A taste of set theory for philosophers

Article · January 2010

CITATIONS
0

READS
112

1 author:

Jouko Väänänen
University of Helsinki
106 PUBLICATIONS **901** CITATIONS

SEE PROFILE

JOURNAL OF INDIAN COUNCIL OF PHILOSOPHICAL RESEARCH

Issue on the Theme of
"Logic and Philosophy Today"

Guest Editors:
Amitabha Gupta and Johan van Benthem

Volume XXVII
Number 1
January-March 2010

Editor: Mrinal Miri
Executive Editor: Godabarisha Mishra

Indian Council of Philosophical Research
Darshan Bhawan
36, Tughlakabad Institutional Area, Mehrauli-Badarpur Road
New Delhi 110062

Articles published in this Journal are indexed in the
Philosophers' Index, USA

ISSN - 0970-7794

Typeset & Printed in India
at Datagraph Creations Pvt. Ltd., Delhi 11052
(D. K. Fine Art Press)
and Published by Member-Secretary
for Indian Council of Philosophical Research
Darshan Bhawan
36, Tughlakabad Institutional Area
Mehrauli-Badarpur Road, New Delhi 110062

Contents

Logic and Philosophy Today: Editorial Introduction

AMITABHA GUPTA AND JOHAN VAN BENTHEM

The Initiative This special issue of the *Journal of the Indian Council of Philosophical Research* (JICPR; http://www.icpr.in/journal.html), is the result of a recent initiative aimed at improving the interactions between contemporary logic and philosophy at universities and colleges in India. This initiative arose out of a chance meeting between Professors Mrinal Miri, the Editor of the JICPR, and Amitabha Gupta. During that meeting, Professor Miri expressed his desire to bring out a Special Issue of the JICPR on the interface of recent developments in Logic and Philosophy. The Journal has maximum reach throughout the country. It was thought that it would be the best instrument to disseminate knowledge of modern logic and its relationship to philosophy in order to enhance the levels of research and education of logic in India. There are already eminent and outstanding Indian logicians residing outside India. What we need now is a strong group inside India involved in advanced research and in training brilliant Indian minds, unleashing local energies in the field - as in ancient times with the Nyāya-Vaiśheṣhika, Jaina and Buddhist schools.

Efforts have already started in India to rejuvenate advanced research and education in logic and its applications, with successful outreach into mathematics and computer science, by organizing Conferences and Winter Schools and forming a new *Association for Logic in India* (ALI; http://ali.cmi.ac.in/), overseeing a wide range of initiatives, including scientific events and various publications. The initiative to publish a Special Issue of the JICPR is in line with this, complementing these efforts by specifically targeting the field of philosophy in India and its activities and programmes relating to research, teaching and learning, by highlighting recent developments in logic and their relevance to philosophy.

The urgent need to come up with a publication that would impact a broad philosophy community in India by making modern logic accessible to it struck a sympathetic chord with Professor Johan van Benthem, a logician based at Stanford University and the University of Amsterdam, who has initiated and supported the cause of propagating logic the world over, in-

cluding recently in China, and who has been associated with the recent Indian efforts from their very inception. Thus, Gupta and van Benthem were invited as Guest Editors entrusted with the ambitious task of bringing out an innovative and distinctive volume on "Logic and Philosophy Today" of the JICPR, soliciting articles from among first-rate logicians in all continents. The volume that you are holding in your hand right now is the result of this editorial collaboration between two Dutch and Indian colleagues. But at the same time, it is much more than that, being the concrete outcome of a truly international effort. It is a pleasure to note the overwhelming response of top-ranking logicians to help enliven the interface of logic and philosophy in India by contributing a paper to this Special Issue of the JICPR. Likewise, the support of the Indian Council of Philosophical Research (ICPR) in Delhi, http://www.icpr.in/, for this enterprise has been generous and gracious all the way.

After this brief history and acknowledgment, let us now turn to matters of content. What you see here before you is a lively panorama of logic research today in a broader setting, written by a large group of distinguished authors who each open a window to their field of expertise for a general philosophical audience. Our aim in all this is to give our readers an impression of what is going on, as well as a path into the literature. Let us first say a bit more about the intellectual background as we see it.

Logic and philosophy over time The juxtaposition of two fields in our title needs no justification. There is a millennia-old history of fruitful interactions between logic and philosophy, in both Western and Eastern traditions. But paths have diverged in recent years. During the last half-century, modern logic has been undergoing a fast expansion of themes and new interdisciplinary alliances, a rich new reality that has hardly registered in the consciousness of philosophers, even those well-disposed toward logic indeed, even those who teach it. What we have tried to do with this issue is provide the reader with a map of major thematic developments in modern logic and its current interfaces.

Logic today Broadly speaking, modern logic was forged in the study of the foundations of mathematics, its rigour and consistency. In effect, this concern with truth and proof in mathematics was a contraction of the traditional agenda of reasoning in general domains, still found with great 19th century logicians like Bolzano or Peirce. But it led to the Golden Age of Mathematical Logic with Frege, Russell, Hilbert and Gödel, whose results

are still central to the discipline as we know it today. At the same time, these new technical insights turned out to be relevant to philosophy, illuminating old issues and creating new directions, witness the work of Wittgenstein, Carnap, or Quine. What has happened after the Second World War is both a continuation of these streams, with many new eminent names joining the pioneers, and also the rise of a wealth of new interfaces of logic with other disciplines. These include linguistics, computer science, and in recent years, also economics and psychology. Logical structures and methods have turned out to be crucial in studying natural language, computation, information flow, interaction, and above all, our cognitive abilities in general. Thus, in a sense, logic is returning to its old broad agenda once more, but with new mathematical tools.

Migrations This broad contemporary role of logic also presents philosophy with new interfaces. It would be hard to write the intellectual history of major themes in logic and philosophy in the last century without tracing their striking further intellectual migrations back and forth across academia. Here is one such saga out of many. It was philosophers who started the study of counterfactual conditionals in their analysis of natural laws; logicians then developed these ideas into conditional logics beyond what mathematical logic provides, and this topic then turned out to be crucial to understanding non-monotonic consequence relations for practical default reasoning in artificial intelligence, while finally, the later logic systems are now being applied in areas as far apart (to the superficial observer) as legal argumentation, the linguistic semantics of normality, brain research with neural nets, and recently, even the study of traditional Indian logic. Van Benthems paper 'Logic in Philosophy' [H. B. Jacquette, ed., 2007, *Handbook of the Philosophy of Logic*, Elsevier, Amsterdam, pp. 65–99] discusses many further examples of this interplay between logic, philosophy and other disciplines, with key logical themes such as knowledge and information coming to reach from practical philosophy to game theory and the social sciences, or dynamic theories of meaning that bridge philosophy, linguistics and computer science.

Logic in India While the above trends make sense for logic and philosophy generally, there is a special interest in bringing these developments to attention in India. It may not be evident *a priori* why people in diverse cultures, with distinct pursuits, disparate convictions, divergent customs and a veritable feast of viewpoints would develop what Amartya Sen called ar-

gumentative traditions and ingeniously nurture them. But they have. And while there are scholarly debates about just what characterized the old Indian study of logic, it is clear that inspired by a robust and vibrant tradition of *naturalism*, India made its mark in the world history of logic, with famous names such as Akapda Gautama, Vasubandhu, Nagarjuna, and Siddhasena Divkara, representing a wealth of schools, in particular, Nyaya, Buddhist Logic, Navya Nyaya, and Jainist logic.

When modern Western logic came to India, scholars first took the Frege-Russell stance, interpreting and reformulating traditional Indian logic to fit that mould, even when the linguistic realities of Sanskrit needed to be twisted occasionally. Whether biased or not, these studies did provide the first significant links, and thereby started a potential conversation across traditions. A later generation of distinguished scholars, influenced more by Quine, then produced much more sensitive analyses of Indian logical thought, widening the contacts. This volume contains a paper by Prabal Sen and Amita Chatterjee, illustrating this by reviewing Navya-Nyya Logic and explaining its difficult ideas and terminology in an accessible fashion, using first order language in the tradition of Sibajiban Bhattacharyya, Daniel Ingalls, Bimal Krishna Matilal, Frits Staal, and in particular, Jonardon Ganeri. In recent years, we see a third wave of studies, many of them bringing the broader logic perspectives outlined in the above to bear on understanding Indian logic. This makes sense, because now that the agenda of Western Logic itself is in flux, its openness to ideas from other traditions tends to increase. These newer perspectives on interpreting Indian texts in logic include case-based reasoning developed by Jonardon Ganeri, paraconsistent logic by Graham Priest, non-monotonic logic by Claus Oetke, dialogical logic by Shahid Rahman, or modern situational logics of information flow, games, and social software by Sarah Uckelman. Our collection includes a paper adding yet one more perspective; Fabien Schang surveys two Indian dialectical traditions and shows how the ancient Indian logicians successfully buttressed the dialectic tradition.

We see in all these phases of contacts historically important stages in increasing mutual understanding between traditions, and we hope that this issue will encourage such studies even further.

Contents of this issue In designing this issue, we have chosen a number of broad areas that allowed us to sample major developments, some extending proven classical lines, others opening new ones. Even so, this publication is not a textbook, but an invitation. Each chapter consists of a

description of an area, with some special highlights, and pointers to further literature. If an author has succeeded in getting you interested, you will then know where to look further.

In Part 1, **History of Logic**, Wilfrid Hodges and Stephen Read give a masterly survey of Western logic, including its subsequent ramifications in Arabic logic. Fabien Schang then samples the Indian tradition through the theme of dialectical logics, while Prabal Sen and Amita Chatterjee introduce its major flowering in Navya-Nyaya Logic. Fenrong Liu and Wujing Yang then conclude with a brief history of a perhaps less-known tradition, that of Chinese logic since Antiquity.

Part 2, **Mathematical Logic and Foundations**, gives some essential technical pillars of the field, with chapters on model theory by Anand Pillay, set theory by Jouko Väänänen, proof theory and the philosophy of mathematics by Jeremy Avigad, computability theory by Barry Cooper, and algebraic logic by Hiroakira Ono.

Part 3, **Logics of Processes and Computation**, charts the thriving interface of logic and computer science (arguably the locus of the bulk of logic research today), with chapters on temporal and dynamic logic by Frank Wolter and Michael Wooldridge, logic and categories by Samson Abramsky, and logic and automata theory by Ramaswamy Ramanujam.

Part 4, **Logics of Information and Agency**, broadens the theme of computation to communication, agency, and logical structures in social organization. Eric Pacuit describes logics of informational attitudes and informative actions, Richard Booth and Tommie Meyer survey modern logics of belief change (the engine of learning and adaptation), and Rohit Parikh, the originator of the well-known program of Social Software employing logic to understand (and improve) social procedures, ends with a key piece on knowledge, games and society.

While many of the earlier pieces are of great relevance to philosophers interested in logical analyis, Part 5, **Logic and Its Interfaces with Philosophy**, tells a more explicit story of contacts between logic and philosophy today. Out of a large set of possible topics, we have selected a representative sample from philosophy of language (Isidora Stojanovic), formal epistemology (Jeffrey Helzner and Vincent Hendricks), logic and philosophy

of science (Bas van Fraassen), logic and ethics (Sven Ove Hansson), quantified modal logic (Horacio Arlo Costa), logic and philosophy of mathematics (Hannes Leitgeb), and logic and metaphysics (Edward Zalta).

We continue this exploration, in line with what we said about migrations earlier, with a number of congenial further interfaces in Part 6, **Logic and Other Disciplines**. Its chapters cover logic and quantum physics (Sonja Smets), logic and probability (Kenny Easwaran), logic and argumentation theory (Dov Gabbay), logic and cognitive science (Alistair Isaac and Jakub Szymanik), decision and game theory (Olivier Roy), and many-valued and fuzzy logics (Petr Hajek).

Taken together, the articles in our issue paint a very broad picture of our field. But pictures arise as much from omitting as applying brush strokes. We could have included many more topics, and we may, in later extensions of this issues. But for now, the material presented here should be enough to open anyone's eyes to the power, sweep and beauty of logic today.

Conclusion This volume does not stand in a vacuum. Indian logicians today are active in university departments of mathematics, computer science, and philosophy and they have been remarkably active in recent years in joining the international community. Organizational efforts began with a series of successful Conferences (2005 and 2007) and Winter Schools (2006) held at IIT Bombay on Logic and its Relationship with other Disciplines that are documented in two forthcoming books: *Proof, Computation, and Agency: Logic at the Crossroads, Vol. 1*, Amitabha Gupta, Rohit Parikh and Johan van Benthem, eds., and *Games, Norms, and Reasons: Logic at the Crossroads Vol. 2*, Johan van Benthem, Eric Pacuit and Amitabha Gupta, eds., both published by Springer Verlag.

Our present initiative hopes to strengthen this process by drawing in more of the Indian philosophical community than was done so far, both through the papers in our volume and an associated meeting in a Conference *Week on Logic* to be held at the University of Delhi from January 5 11, 2011. We plan to bring together our authors with teachers, research scholars and students from Departments of Philosophy in the country as well as participants of ALI Winter School.

But let content have the final say. The various contributions in this issue paint a rich picture of logic today, in a way that we hope will be of interest to philosophers. It has amazed us to see how easy it was to collect a distinguished galaxy of both senior and junior logicians from all over the world, willing to share their ideas and insights with a broader audience. The articles collected here may not all be 'easy reads', but if you make the effort, they will show you something that is rare: both the broader vision of todays researchers on their broader areas, and their enthusiasm about specific themes. Indeed, the editors themselves have learnt a lot of new things about logic today, beyond what they imagined. Of course, not all our authors will agree on what modern logic is exactly, or where it is heading. We stated our own view in the above, but that was just an 'editorial license': taken together, it is the papers in this volume that tell the real story of the field today. But no matter how one construes the march of history, we are certain that, once these contacts have been made, Indian logicians will come to be noticed more and more at the world-wide stage, adding original insights in philosophy, mathematics, language, computation, and even the social sciences. And we would not be surprised at all if some of this innovation would come about by drawing upon India's own rich logical tradition.

Amitabha Gupta and Johan van Benthem
October 2, 2010

Acknowledgement

We would like to acknowledge the contributions of many people who helped produce this volume. We have already mentioned our indebtedness to the Editor of JICPR and the leaders of the ICPR who initiated and supported this initiative right from the start.

Our Introduction also put the spotlight on our authors, the people who provided the real content and sparkle for these volumes. But in addition, we are grateful for various other forms of essential support.

The following colleagues reviewed papers and, through their comments, helped improve overall quality and coverage:

S.D. Agashe, Alexandru Baltag, Dietmar Berwanger, Giacomo Bonanno, Marisa dalla Chiara, Paul Dekker, Igor Douven, Jan van Eijck, Peter van Emde Boas, Jonardon Ganeri, Valentin Goranko, Siegfried Gottwald, Nirmalya Guha, Wilfrid Hodges, Wesley Holliday, Thomas Icard, Ulrich Kohlenbach, David Makinson, Eric Pacuit, Gordon Plotkin, Henri Prakken, K. Ramasubramanian, Manuel Rebuschi, Jan-Willem Romeijn, Hans Rott, Jeremy Seligman, Keith Stenning, Raymond Turner, Albert Visser, as well as Jan Wolenski.

We also thank the type-setting team (Hari Priyadarshan and Mr. Nirmesh Mehta) for its herculean and prodigious help in the preparation of the final manuscript and producing the camera ready copy of this massive document while completing the work in the shortest possible time.

Finally, we thank Sunil Simon of The Institute of Mathematical Sciences (IMSc.), Chennai and CWI, Amsterdam for his quiet and efficient logistical assistance in coordinating this complex international process.

PART I

History of Logic

Western Logic

WILFRID HODGES AND STEPHEN READ

The editors invited us to write a short paper that draws together the main themes of logic in the Western tradition from the Classical Greeks to the modern period. To make it short we had to make it personal. We set out the themes that seemed to us either the deepest, or the most likely to be helpful for an Indian reader.

Western logic falls into seven periods:

(1) Classical Greece (Parmenides, Plato, Aristotle, Chrysippus; 5th to 1st centuries BC)

(2) The Roman Empire (Galen, Alexander, Porphyry, John Philoponus, Boethius; 1st to 7th centuries AD)

(3) The Arabs (Al-Fārābī, Ibn Sīnā, Khūnajī, Qazwīnī; 8th century – present)

(4) The Scholastics (Peter Abelard, Peter of Spain, William of Ockham, John Buridan; 12th - 15th centuries)

(5) Renaissance to Enlightenment (Ramus, Port-Royal Logic, Leibniz; 15th to 18th centuries)

(6) Transitional (Boole, Peirce, Frege, Peano, Russell, Gödel, Tarski, Gentzen; 19th century – mid 20th century)

(7) The modern period (mid 20th century – present)

The division is rather neat; each period built on the one before it. The chief exception to this is Arabic logic; its high point partly overlapped the beginning of Scholastic logic, and after the 13th century its development was independent of European logic. Of course all the dates are approximate, and there were many important logicians besides those named above.

We finish this paper at the end of period (6), in the mid 20th century. That period saw some major changes of paradigm in the study of logic. By the time of the Second World War those changes had worked their way

through the system, and post-war logicians set their minds to exploiting the new paradigms. The rest of this volume tells you how they did it.

Our thanks to Khaled El-Rouayheb, Robert Gleave, Graham Priest, Karen Thomson, Johan van Benthem and an anonymous referee for various corrections and suggestions. All remaining errors are our own.

1 Classical Greece

1.1 Aristotle's predecessors

Early in the 5th century BC a Greek philosopher named Parmenides, who lived in the Greek colony of Elea in South Italy, published a poem called the *Way of Truth*. In the Introduction he promised his readers that they would learn about the 'well-rounded truth' as well as the 'utterly untrustworthy common opinions (*doxai*) of humans'. Like the Advaita Vedānta, he believed that there is only one real entity. He claimed to prove this by assuming the opposite (the 'untrustworthy common opinion' that there is more than one thing) and deducing a contradiction.

His arguments were embarrassingly bad. But he established several of the key traditions of Greek logic. First, he showed (or claimed to show) that we can learn new and surprising things by using methods of pure thought. The chief method that he used is known today as Proof by Contradiction, or Reductio Ad Absurdum (Indian *prasaṅga*, traditionally ascribed to Nāgārjuna in around AD 200). But although Parmenides used this method, he didn't describe it. That was left to Aristotle around 150 years later, and is one of the reasons why Aristotle is reckoned the inventor of logic.

Second, Parmenides invented the Greek tradition of devising paradoxes; in fact 'paradox' means 'contrary to common opinion', as in Parmenides' use of the word *doxai* above. But again it was later Greeks who first devised paradoxes that really challenge our thinking. The first of these later Greeks was Parmenides' follower Zeno of Elea, who invented several well-known mathematical paradoxes, including 'Achilles and the Tortoise'. Around 350 BC, the Megarian logician Eubulides discovered some of the best logical paradoxes, including the Liar. (Am I telling the truth or lying when I say 'I am now telling a lie'?)

Third, he would have been horrified to know it, but Parmenides was probably one of the origins of the Greek tradition of eristic, which is the art of winning arguments regardless of whether you have a good case. (The lawyers must have had something to do with it too.) In the early 4th century the Athenian philosopher Plato wrote a number of fictional dialogues,

mostly involving his philosophical hero Socrates. One of them, the *Euthydemus*, is an entertaining account of a performance by two itinerant experts in eristics. Their arguments rest mostly on obvious ambiguities in words (just as in Parmenides' poem, but their ambiguities are generally funnier). Aristotle, who was a student of Plato's, wrote a book *Sophistical Refutations* which analysed the methods of eristics. This book had a huge influence in late 12th century Europe after the Latin translation became available around 1140 (and was arguably the main stimulus to the creation of terminism and the theory of properties of terms — see §4 below). In later Western logic, eristics survived as a kind of undercurrent; Schopenhauer wrote a textbook of it in 1831.

There is an obvious parallel between eristic argument and the *jalpa* debates described in the *Nyāyasūtra* a few centuries later, where the aim is to win by fair means or foul. Aristotle in several places (for example *Sophistical Refutations* 2) gave classifications of arguments according to their purpose, and several of the kinds that he mentions are really kinds of debate. For example he mentions 'didactic', 'dialectical', 'examinational', 'contentious' and 'rhetorical' arguments. It seems that the *Nyāyasūtra* classification is completely independent of Aristotle's; a comparison would be interesting.

Plato made important contributions of his own to logic. He had learned from Socrates that one essential ingredient of correct reasoning is to have sound and well-defined concepts. In his dialogues he developed a technique of definition which is called Division. To define a class X which interests us, we take a class A that includes X, and we divide it into two clearly defined parts $A1$ and $A2$, so that one of the parts, say $A1$, contains all of X. Then we split $A1$ into two parts $A11$ and $A12$, so that one of the parts contains all of X. We carry on subdividing until we have narrowed down to a class that contains all of X and nothing else. Then we can define X as the class of things that are in A and in $A1$ and .... The fullest account of this method is in Plato's dialogue *Sophist*.

1.2 Aristotle

But the main breakthrough in Classical Greek logic was certainly Aristotle's work *Prior Analytics*. Its contents were probably written in the third quarter of the 4th century BC. Aristotle's works are a strange mixture of books, lectures and notes, and we are often unsure that he intended to write treatises in their present form. Nevertheless the *Prior Analytics* contains one of the world's first tightly integrated formal systems, comparable in a

way with Pāḍini's description of Sanskrit. In this work Aristotle described rules of argument, and showed how all his rules could be derived from a small starting set. The rules were called 'syllogisms'.

Ignoring the modal syllogisms (which are still controversial — see §4.1 below), Aristotle described what were later enumerated as nineteen syllogisms. In the Middle Ages they were given mock-Latin names for easy memorising. (See §4.1 below.) The first and most famous syllogism was the one that the medievals called Barbara. As Aristotle himself presents it, it takes the form

> If *C* belongs to all *B*, and *B* belongs to all *A*, then *C* belongs to all *A*.

The letters mark places where one can put terms, i.e. (in general) nouns or noun phrases; the same noun should be put for '*A*' at both occurrences, and likewise with '*B*' and '*C*'. Probably he intended that different terms should be put for different letters too. It's virtually certain that Aristotle took the idea of using letters from the Greek geometers.

For example Aristotle might write

(1) If every fisher is a hunter, and every angler is a fisher, then every angler is a hunter.

This is our example and not his; the few explicit examples that he did give are mostly tricky cases that needed special analysis. We took the idea of this example from Plato's definition of 'angler' in *Sophist*; many people believe that Aristotle first devised his argument rules through developing Plato's definitions in this kind of way. However that may be, Aristotle's next move was to see that the validity of the argument in (1) doesn't depend on the terms that are put for the letters. We could use any terms, provided that the resulting sentences make sense and we always use the same term for the same letter. So he had discovered not just valid arguments but *valid argument forms*; every argument of that form is guaranteed to be valid. He could have written this form as

(2) Every *B* is a *C*. Every *A* is a *B*. So every *A* is a *C*.

just as most later logicians did. Perhaps he used the roundabout phrasing '*C* belongs to all *B*' because he realised that he had invented a completely new discipline, and he wanted to mark this with some new technical terminology.

What was most distinctive of Aristotle's contribution to logic, however, was that he gave general form to two methods: the method of showing

syllogisms to be valid, and the method of showing invalid argument forms to be invalid. The former method was to reduce all valid syllogisms to what he called the 'perfect' or 'first figure' syllogisms, and ultimately to two of these, Barbara (as above), and Celarent:

No *B* is a *C*. Every *A* is a *B*. So no *A* is a *C*.

The other method, of showing arguments invalid, was to find replacements for the constituent descriptive terms, or the symbolic letters, such that the premises are true and the conclusion false. E.g., take the argument form:

Every *A* is a *B*. Some *B* is a *C*. So some *A* is a *C*.

If we replace '*A*' by 'horse', '*B*' by 'animal' and '*C*' by donkey', we can see that the conclusion cannot follow from the premises, since it is false and they are true.

Although Aristotle began his career as a follower of Plato, he later asserted his independence, and for some centuries his followers (the Peripatetics) and the Platonists formed competing schools. This rivalry generated a number of myths that still survive today; you can find some of them on the internet. For example it was claimed that Pythagoras and Parmenides both had systems of logic, and that Plato had inherited them. But in fact there is not the slightest evidence that Pythagoras ever had anything to do with logic, and certainly Parmenides had nothing like a system.

1.3 Stoic Logic

Attempts by Platonists to establish a platonist logic to rival Aristotle's logic never succeeded: Aristotle had cornered all the logically worthwhile ideas in Plato's work. But the later classical Greeks were fortunate in having a second substantial theory of logic besides Aristotle's, namely that of the Stoics (who inherited logical insights from some earlier logicians, notably the Megarians). The leading figure of the Stoic school was Chrysippus, who lived in the second half of the 3rd century BC. Unfortunately no complete logical works from this school survive — though we are told that Chrysippus himself wrote over a hundred logical treatises, including seven on the Liar Paradox. But we know enough to point to some important innovations by this school.

First, they invented propositional logic. Second, their notion of modality was formally different from Aristotle's. For Aristotle (at least on one reading of his rather obscure explanations), humans are 'necessarily animals' but 'possibly writers'; the modality goes with the description. For

the Stoics, necessity and possibility are properties of whole assertions: 'It is night' is possibly the case but not necessarily the case. In the later terminology, Stoic modalities were *de dicto*, 'about something said'. (And as in some Indian traditions, Stoic logicians used modal notions as properties of propositions rather than as parts of propositions.) Third, Stoics had the notion of 'incomplete' meanings, which (to use modern terminology) have an argument place that needs to be filled. For example 'writes' is incomplete because it needs a subject argument, as in 'The moving finger writes'. Fourth, they had at least the beginnings of a sophisticated philosophical theory of meanings, intended to answer questions like 'What entities are most properly described as having a truth value?' The Stoics also had a reputation for being formalistic, but at this distance in time and with the scanty records that we have, it would be unsafe for us to ascribe to them any particular formalistic doctrine.

The first three of these Stoic contributions eventually passed into the general practice of logic. But by the time of Arabic logic the Stoics as a distinct school of logic had faded from the record.

1.4 Acquisition of knowledge

Writers on Indian logic have often remarked that Indian logic, unlike most modern Western logic, is about how an individual comes to know something that he or she didn't know before. Inference is a process that happens in the mind of the reasoner. It is not always realised that, with only marginal exceptions, exactly the same was true for all proofs in Western logic before the beginning of the twentieth century. For example the syllogisms that Aristotle counted as not 'perfect' were those where the conclusion doesn't *obviously* follow from the premises. His reductions of these syllogisms to perfect syllogisms were not just abstract validity proofs; they were chains of reasoning that a reasoner could use in order to be convinced of the truth of the conclusion of a non-perfect syllogism.

One of the main purposes of logic in the West has been to validate arguments by bringing them to some appropriate kind of 'logical form'. But this meant something different in traditional Western logic from what it came to mean in the twentieth century. The traditional logicians reckoned that a piece of informal reasoning could be reduced to steps, and each step introduces its own piece of knowledge. The steps could be formalised separately; for example there was no requirement even to use the same terms in one step as in the next. So a complicated argument would be reduced to a mixture of logical steps — each simple in itself — and linguistic rearrange-

ments or paraphrases. But in the late nineteenth century logicians started to take a very different approach: the terms in an argument are symbolised, and the same assignment of symbols applies to the whole argument from start to finish, even if the argument consists of many pages of mathematics. As a result, a modern Western student of logic learns how to operate formal proofs of much greater complexity than in the traditional format. In this modern style the separate steps of a formalised argument are not each intended to convey a separate piece of knowledge — at least, not in any straightforward way. One reason for the current interest in the Scholastic obligational disputations (see §4.3 below) is that unlike syllogisms, they do generate arguments with some significant complexity, though these arguments are not really proofs that give us new knowledge.

2 The Roman Empire

During the first century BC, Aristotle's logical writings — which had previously been kept in the private hands of Peripatetics — were edited and published as a group of books called the *Organon*. The editor (said to be Andronicus of Rhodes) put first the book *Categories*, which is about the meanings of single words. Book 2 was *On Interpretation*, which discussed the ways in which words are arranged in sentences. Then he put book 3, the *Prior Analytics*, which explained how to arrange sentences into valid arguments. Book 4, the *Posterior Analytics*, was about how to use syllogisms in order to increase our knowledge. Book 5, the *Topics*, was about debate. Book 6 was the *Sophistical Refutations*; we mentioned it in §1.1 above. In one tradition, two more of Aristotle's books were included in the *Organon*, namely the *Rhetoric* and the *Poetics*; the first of these was about persuasive public argument and the second was about the expressive force of poetry and drama.

The *Organon* and other works of Aristotle contained an immense amount of learning, but they were hard to read. Around AD 200 the Peripatetic philosopher Alexander of Aphrodisias wrote commentaries on the main works, including the *Prior Analytics*. His is the first commentary to survive of a tradition which lasted for a thousand years. The commentary format was so successful that throughout the first millennium AD and for some while after, the main research in logic appeared in the form of commentaries on books of the *Organon*.

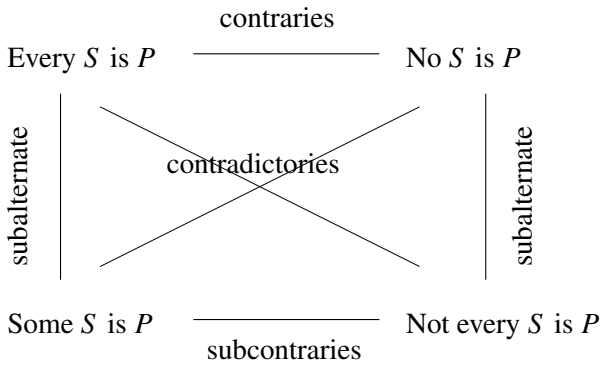
Students working in this tradition were shown how to break a text down into separate inferences, and to check each inference by logic. Each infer-

ence needed to be tickled into the shape of a syllogism by suitable paraphrasing. The result was that any substantial piece of logical analysis used partly logic and partly paraphrase. The paraphrase was done by intuition based on studying many cases, and not by rule. Leibniz later described such steps of paraphrase as 'valid non-syllogistic inferences'.

We can illustrate this with an important example from the Roman period. Aristotle had been interested in the nature of mathematical knowledge, and his views about this may well have influenced later Greek mathematical writing, for example Euclid's *Elements*. But it seems unlikely that the reasoning procedures of Greek mathematics had any influence on Aristotle's syllogisms — the mismatch is too great. For example most statements in geometry use relations: 'lines L and M are parallel', 'point p lies on line L' and so on. Syllogisms had no machinery that handles relations naturally. Nor had the propositional logic of the Stoics. The logicians of the 2nd century AD made the first attempts to reconcile logical methods and mathematical ones. It was apparently Alexander of Aphrodisias who took the crucial step of representing relations by allowing the 'Every' and 'Some' in syllogisms to range over pairs or triples as well as individuals.

In fact, in the 1880s C. S. Peirce took up this idea of using pairs, triples etc. (which he credited to his own student Oscar Mitchell who had introduced 'propositions of two dimensions'). On the basis of it Peirce invented what we now recognise as the earliest form of first-order predicate logic. But there is an important difference between Alexander's idea and Peirce's. Alexander never introduced any method for passing from statements about individuals to statements about pairs, or from statements about pairs to statements about triples, etc. For him and the traditional logicians who followed his lead, no such method was needed, because one could take care of the switch by using paraphrase between the logical steps of an argument. But Peirce's predicate logic allows us to use facts about pairs to deduce facts about individuals, and so on, all within the same formalism. Today no logician would dream of stepping outside a formal proof in mid stream in order to cover a step by paraphrasing.

The Roman Empire commentators tidied up several other aspects of Aristotle's logic. One important contribution from this period was the Square of Opposition, a diagram which records the logical relations between the four propositions in the corners of the square:



A proposition entails its subalternates; contraries cannot both be true, but can be false together; subcontraries cannot be false together, but could be true together; contradictories cannot both be true and cannot both be false.

Most of the Roman Empire commentators on Aristotle after Alexander of Aphrodisias were in fact Platonists or Christians, not Peripatetics. How could they justify teaching the views of the founder of a rival philosophy? They found a tactful solution to this problem. Logic was so obviously valuable that all students should learn it. But the commentators found that they could detach the logic from Aristotle's philosophy and metaphysics. A philosophy-free logic was taught as a first step, and when the students had it under their belt, they would move onto the higher truths of Platonism (or later, Christianity or Islam). But the commentators didn't want to teach logic by pure rote, so they found a kind of justification for it in semantics — the study of the meanings of words and sentences. Thus the students would learn semantics from the first two books of the *Organon* and then move on to syllogisms in the third book.

An example may help comparison with Indian traditions. The point comes up in various Indian treatises that when we make a deduction from a general rule, e.g. 'Whenever there is smoke there is fire', we need to point to an instance that confirms the rule (a *sādharmya-dṛṣṭānta*). The Roman Empire commentator tradition wouldn't have put it like that. If the reason for giving the instance is that a general rule doesn't count as true unless it has an instance, then that should have been said in the explanation of the meaning of general rules. It should be made a point of semantics, not a step in arguments. And in fact some of the commentators of this period did count an affirmative universal statement 'Every *A* is a *B*' as false unless there is at least one *A*. (But they allowed the negative statement 'No *A* is a *B*' to be true when there are no *A*s.)

This meaning-based logic must have been the brainchild of many different scholars, but the Palestinian Platonist philosopher Porphyry of Tyre in the late 3rd century is believed to have played a key role. Porphyry also wrote an elementary introduction to logic; he called it the *Introduction* (*Eisagōgḗ* in Greek), and for several hundred years it was read by every student of logic. In it he mentions some philosophical problems about genera and species (like 'animal' and 'human'); these problems later became known as parts of the 'problem of universals'. For example do genera and species really exist as entities in the world? Porphyry adds that he is deliberately not discussing these problems. The Scholastics couldn't resist the challenge of tackling the problem of universals, and the result was that in the West the ideal of a philosophy-free logic went down the drain. It was recovered in a more scientific form through the work of Carnap, Tarski and other logicians in the period between the two world wars of the twentieth century. (See §6 on Tarski.)

3 The Arabs

Logic has had a good reputation through most of Islamic history. There are many statements in the Qur'an along the lines 'Thus do We explain the signs in detail for those who reflect' (10.24), and these are commonly understood as calls to Muslims to develop their rational thinking. In the early days of the Islamic empire there were a number of well-to-do Arabic speakers, spread across the world from Spain to Afghanistan, who regarded skills of debate as a mark of culture. So they bought logic texts and took lessons in logic. Tamerlane had two distinguished Arabic logicians at his court in Samarqand. Ibn Sīnā (known in Europe as Avicenna) reported that in the late 990s the library of the Sultan of Bukhara (in present-day Uzbekhistan) had a room full of logical texts. Probably it contained translations of most of the Roman Empire commentaries. Most of this material is lost today, or at least uncatalogued. We know there are important Arabic logical manuscripts that have never been edited; for example some are in Turkey and some are in the Indian National Library.

Logic did sometimes have to fight its corner. There were demarcation disputes between logicians and linguists about which aspects of language should be studied in which discipline. A more serious problem developed later: some of the main experts in logic had unorthodox religious views. In around 1300 Ibn Taymiyya — whose religious and political writings have inspired Osama bin Laden — argued strongly against Aristotle's logic. But

about 200 years earlier Al-Ghazālī had mounted a largely successful campaign to convince Muslims that Aristotle's logic was theologically innocent and a great help for reaching the truth. It's largely thanks to Al-Ghazālī that logic has been a major part of the madrasa syllabus ever since.

The first Arabic logician of distinction was Al-Fārābī in the early 10th century. Today philosophers cite him for his views on the relation of logic to determinacy, among other things. A century later came Ibn Sīnā, a logical giant comparable in various ways to Leibniz. It almost passes belief that the medieval scholars who translated classical Arabic philosophy into Latin thought his logic was not worth translating, so that it was unknown in Europe. (But they did translate the more conservative modal logic of Ibn Rushd — Averroes — which influenced the English logicians Kilwardby and Ockham.) One of Ibn Sīnā's books (*Easterners*, which unfortunately is available only in an unreliable Arabic version) has a long section on how he thought logic should be done; in comparison with Aristotle's logic, this section had much less about rules of proof, much more about how to interpret statements and texts, and a long section on definition. Here and elsewhere Ibn Sīnā emphasised that what we mean is nearly always a good deal more complex than what we say — we mentally add 'conditions' to the public meanings of our public words.

Here is a typical example of Ibn Sīnā's semantic analysis. What does a statement 'Every *A* is (or does) *B*' mean? If we look at examples we can see that there are various patterns. When we say that every horse is a non-sedentary animal, we don't mean just now or sometimes, we mean always. But again we don't imply that every horse is eternal; every horse is a non-sedentary animal for as long as it lives. But when we say God is merciful, we mean it for all time. Next take the statement that everyone who travels from Ray in Iran to Baghdad in Iraq passes through Kermanshah near the border. A person who says this certainly doesn't mean that every such person passes through Kermanshah for as long as he lives; she means sometime during the journey. On the other hand a biologist who says 'Everything that breathes in breathes out' doesn't mean that it breathes out at some time while it was breathing in! And so on. Ibn Sīnā did some preliminary cataloguing of these and other cases. But his general position on these examples seems to have been that we should be alert to the possibilities, and we should aim to reason with them in ways that we find intuitively natural. He believed that a training in Aristotle's syllogisms would help us to do that.

Note the form of the statement about the traveller from Ray to Baghdad. It can be written as follows:

> Every traveller who makes a journey from Ray to Baghdad reaches Kermanshah during the journey.

Sentences with a similar form appeared in Scholastic logic in the early 14th century. One due to Walter Burley reads:

> Every person who owns a donkey looks at it.

Thanks to Burley's example, sentences of this kind have become known as 'donkey sentences'. The mark of a donkey sentence is that it contains two parts, and the second part refers back to something introduced by an implied existential quantifier inside the first part. For first-order logic these sentences are nonsensical: the reference in the second part is outside the scope of the quantifier. In the English-speaking world the question what we can infer from a donkey sentence has been seminal for research into natural language semantics. About the same time as this research began, Islamic jurists independently realised that they had a donkey sentence in a verse of the Qur'an (49.6):

> If a person of bad character brings you a report, you should scrutinize it carefully.

(Note the quantifier 'a report' in the first part, and the back reference 'it' in the second — strictly the 'it' is missing in the Arabic, but it is clearly understood.) A number of jurists have published analyses of this verse and its implications. They make no direct reference to logic, but it's plausible to see in their analyses an indirect influence of Ibn Sīnā, through the logic of the madrasas. The most famous of these jurists is well known for other reasons: Ayatollah Khomeini.

A second feature of the traveller example is that there are quantifiers both over the traveller and over time. This makes it a 'proposition of two dimensions' in Oscar Mitchell's sense (see §2 above). Sadly Ibn Sīnā had no Peirce to transmute his ideas into a radically new logic. In fact later Arabic logicians recognised the originality of Ibn Sīnā's examples, but often their tendency was to introduce new moods of 'syllogism' for each new kind of example. Later Arabic logicians studied further examples and duly added further 'syllogisms'. This style of logical research made strides in the Ottoman empire during the 18th century and led to a relatively sophisticated logic of relations, at a time when European logic was largely moribund.

It seems to have been the Arabic logicians who began the study of reasoning in conditions of uncertainty. Both Al-Fārābī and Ibn Sīnā reprimanded the doctor and logician Galen (2nd century AD) for missing the

fact that most medical statements are probabilistic. Much later, in Paris in 1660, the *Port-Royal Logic* of Arnauld and Nicole discussed how to think rationally about the danger of being struck by lightning. In the 19th century Augustus De Morgan and George Boole tried to incorporate quantitative probability reasoning into logic. But the trend was against them, and in the 20th century probability theory came to be recognised as a separate discipline from logic. (See also the paper 'Logic and probability' by K. Easwaran in this collection.)

One would expect some mutual influence between Arabic and Indian logic because of the geographical closeness. But no direct influences have been discovered. For example one of the leading Arabic scientists, Al-Bīrūnī, being compelled to visit North India in the early 11th century as part of the entourage of a warlord, used the opportunity to collect information on Indian science and culture. He wrote a long report with a mass of information about Indian achievements, including philosophy and astronomy. But his book makes no mention of Indian logic. He does refer to one logical text, the *Nyāyābhāsa*, but he describes it as a book on Vedic interpretation. If he came across Indian logic at all, he simply didn't recognise it as logic.

4 The Scholastics

Although there are important discussions of logical issues in such eleventh and early twelfth century thinkers as Anselm of Canterbury, Peter Abelard and Adam Balsham, the distinctive contribution of medieval logic as a body of doctrine began in the late twelfth century in the study of consequence and fallacies. This began with the rediscovery in the Latin West of Aristotle's doctrine of fallacy in his *Sophistical Refutations* (known in Latin as *De Sophisticis Elenchis*) and of the syllogism in his *Prior Analytics*. However, although Boethius (480-525) had translated all of Aristotle's *Organon* except the *Posterior Analytics* (as part of a grand project, never completed, of translating all of Aristotle's works into Latin with commentaries on them), only Boethius' translations of the *Categories* and *On Interpretation* (*De Interpretatione* in Latin, *Peri Hermeneias* in Greek) were known and in circulation at the start of the twelfth century — these two were termed, along with Porphyry's *Introduction*, the *logica vetus*. During the rest of the century, Boethius' translations of the other works emerged (from where is unknown) and in addition translations of both *Analytics*, the *Topics* and the

Sophistical Refutations were made by James of Venice (who had studied in Constantinople) around mid-century. These became known as the *logica nova*. The theory of the syllogism became the basis of the medieval theory of consequence. What is important to realise is that the assertoric syllogism only takes up a relatively small part of the *Prior Analytics*. Aristotle there also developed a theory of the modal syllogism. But whereas his theory of the assertoric syllogism was clear and convincing, his theory of the modal syllogism was highly problematic.

4.1 Consequence

In fact, the syllogism is not the whole of Aristotle's logic. For as we noted, Aristotle's method of validating syllogisms was to reduce all syllogisms to the perfect syllogisms of the first figure — and ultimately to Barbara and Celarent. The method of reduction depended on a number of one-premise inferences elaborated in *On Interpretation*, in particular, simple conversion, conversion *per accidens*, subalternation, and *reductio per impossibile*. The assertoric syllogism is concerned with so-called categorical propositions (a better translation is "predicative proposition", or "subject-predicate proposition", since the Latin *categorica* is simply a transliteration of the Greek *katēgorikē*, 'predicative'), in particular, the four forms 'Every *S* is *P*' (so-called A-propositions), 'No *S* is *P*' (E-propositions), 'Some *S* is *P*' (I-propositions) and 'Some *S* is *P*' (or better, 'Not every *S* is *P*', O-propositions).

One of the main sources of our knowledge of late twelfth and early thirteenth century logic is Peter of Spain. For many centuries, he was thought to be the same Peter of Spain as Pope John XXI, who was killed in 1276 when the roof of his new library fell on him. Recently, however, it has been established that this was a misidentification, and that the logician Peter was a Dominican from Estella (Lizarra) in the Basque country. His *Tractatus* ('Treatises') record the state of the art, and contain the famous mnemonic by which students learned the theory of the assertoric syllogism:

> Barbara Celarent Darii Ferio Baralipton
> Celantes Dabitis Fapesmo Frisesomorum;
> Cesare Camestres Festino Baroco; Darapti
> Felapton Disamis Datisi Bocardo Ferison.

There are here three figures. Aristotle conceived of syllogisms as pairs of premises, asking from which such pairs a conclusion could be drawn. Those pairs of categorical propositions containing between them three terms

could share their middle terms as subject of one and predicate of the other (figure I), as predicate of both (figure II — Cesare - Baroco) and as subject of both (figure III — Darapti - Ferison). In the first figure, the predicate of the conclusion could be the predicate in its premise (concluding directly — Barbara - Ferio) or the subject (concluding indirectly — Baralipton - Frisesomorum). Only when the syllogism was thought of as an arrangement of three propositions, two premises and a conclusion, did it seem better to call the indirect first figure a fourth figure, as some Stoics (e.g., Galen) and some medievals (e.g., Buridan) did.

The mnemonic lists 19 valid syllogisms. Five more result from weakening a universal conclusion by subalternation. The first three vowels give the type of the constituent propositions; certain consonants record the reduction steps needed to reduce the mood to a perfect syllogism, that is, one in the direct first figure. E.g., Baralipton (aai in the indirect first figure) is reduced to Barbara by converting the conclusion of Barbara *per accidens* (from 'Every S is P' to 'Some P is S'), as indicated by the 'p' following the 'i'. The initial consonant indicates the perfect syllogism to which it reduces.

The modal syllogism results from adding one of three modalities to one or more of the premises and the conclusion. The modalities Aristotle considers are 'necessary', 'possible' and 'contingent' (or 'two-way possible'). In its full articulation, the theory was very complex. But there was something puzzling right at its heart, sometimes known as the problem of the two Barbaras. In ch. 3 of the *Prior Analytics*, Aristotle says that E- and I-propositions of necessity convert simply, that is, 'No A is necessarily B' converts to 'No B is necessarily A' and 'Some A is necessarily B' converts to 'Some B is necessarily A', and necessary A-propositioins convert *per accidens*, that is, 'Every A is necessarily B' converts to 'Some B is necessarily A'. But in ch. 9 of that work, he says that adding 'necessarily' only to the premise of Barbara containing the predicate of the conclusion validly yields a necessary conclusion (i.e., 'Every B is necessarily C, every A is B, so every A is necessarily C' is valid), but not if 'necessarily' is only added to the other premise (i.e., 'Every B is C, every A is necessarily B, so every A is necessarily C' is invalid). The challenge is to find a common interpretation of 'Every S is necessarily P' which verifies these two claims. A very natural interpretation of the remark in ch. 3 is that he takes necessity *de dicto*, or as the medievals would say, in the composite (or "compounded") sense, or as modern logicians would say, with wide scope, so that it predicates necessity of the *dictum*, the contained assertoric proposition. But on this reading, the modal Barbara of ch. 9 would not be valid. For necessar-

ily every bachelor is unmarried (taken *de dicto*), but supposing everyone in the room is a bachelor, it does not follow that necessarily everyone in the room is a bachelor. One way to make the syllogism valid is to take the necessity *de re*, or as the medievals would say, in the divided sense, or in modern terms, with narrow scope: every *B* is of necessity *C*. For then the syllogism reduces to a non-modal case of Barbara with a modal predicate, 'of necessity *C*', and so is valid. But the conversions of ch. 3 fail when taken *de re*.

Forcing a choice between *de re* and *de dicto* interpretations of the modal premise may be anachronistic and out of sympathy with Aristotle's metaphysical projects. Nonetheless, this and other problems with the modal syllogism led to much discussion of modal propositions and a variety of logics of modality in the scholastic period and among the Arabs. We will return to a further problem of interpretation of modal propositions shortly.

The main development of medieval logic (the *logica modernorum*, the logic of the "moderns", as it came to be known), however, was to develop a general theory of consequence. In the twelfth century, one focus of concern was a claim of Aristotle's, endorsed by Boethius, that no proposition entailed its contradictory, since they could not both be true, nor did any single proposition entail contradictories, so if it entailed one of a contradictory pair, it couldn't entail the other. But there is, at least with hindsight, an obvious counterexample, namely, an explicitly contradictory proposition, which entails (by the rule known as Simplification, from a conjunction to each of its conjuncts) each of its contradictory conjuncts. Moreover, a contradiction entails not just its conjuncts, but any proposition whatever. For we can disjoin one of the contradictory conjuncts with any other proposition, and since the other conjunct contradicts the first disjunct, that other arbitrary proposition immediately follows. Such surprising results showed that what was needed was a general theory, and it developed along two fronts. The primary line of development was a theory of inference, framing inference rules in terms of the structure of the propositions in question. At the same time, the theory of fallacies developed, building on Aristotle's theory of fallacy in his *Sophistical Refutations* and on his method of counterexamples from the *Prior Analytics*. In time this led to a second and supplementary account of consequence in terms of truth-preservation.

4.2 Properties of Terms

Aristotle had had relatively little to say about propositional consequence in *On Interpretation* apart from the rules that the later commentators incor-

porated in the Square of Opposition (§2 above). But what, for example, explains subalternation, from 'Every *S* is *P*' to 'Some *S* is *P*'? To explain such inferences, the medievals developed their distinctive theory of properties of terms. As the twelfth century proceeded, many properties were mooted: signification, supposition, appellation, copulation, ampliation, restriction and relation were some of them. Part of the spur to this was metaphysical: if, as Aristotle had said, everything is individual, and the only universals were names, one needed a theory of signification, or meaning, to explain the functioning of names. Supposition then explained how terms functioned in propositions, and in particular picked out that class of things the term stood for, and how it did so. Thus the theory of supposition has two aspects, the first concerning what the term stands for, the other the mode of supposition. Sometimes, for example, a term supposits for itself, or some other term (one which it doesn't signify), as in 'Man is a noun', or 'The spoken sounds "pair" and "pear" sound the same' — we nowadays mark such uses with inverted commas. The medievals said the term had material supposition. Other cases where a term does not supposit for the things it signifies are when it stands for the universal (if there is one) or the concept, e.g., in 'Man is a species'. This was said to be a case of simple supposition. Some authors, especially realists, thought supposition should be restricted to subjects, and predicates had copulation (i.e., coupled to the subject). Others thought predicates had simple supposition, for the universal. The hard-line nominalists, however, like William of Ockham and John Buridan, in the fourteenth century, thought both subject and predicate stood for individuals. For example, in 'A man is running', 'man' and 'running' stand for men and runners, respectively, and have so-called personal supposition. The proposition is true if subject and predicate supposit for something in common — if the class of men overlaps the class of runners. Thus subject and predicate in personal supposition stand for everything of which the term is presently true. 'Man' supposits for all (present) men and 'running' for all those presently running.

What, however, of, e.g., 'Some young man was running'? Suppose Socrates is now old, but in his youth he ran from time to time. 'Young' restricts 'man' to supposit only for young men; but 'was running' ampliates the subject 'young man' to supposit for what is now, or was at some time, a young man. So the proposition is true of Socrates, since he was at some time a young man and ran. Indeed, it is true if he never ran in his youth but ran yesterday, say. So ampliation and restriction analyse 'Some young man was running' to say 'Something which is or was at some time a young man was at some time (not necessarily the same time) running'. Not

only tense, but also predicates such as 'dead' ampliate the subject. 'Some man is dead' is true because something which was a man is now dead.

Modal verbs also ampliate their subjects. But there was disagreement how they did so, and what the truth-conditions of modal propositions were. For example, 'Cars can run on hydrogen' is true even if no existing cars can run on hydrogen provided something which could be a car could run on hydrogen. So the modal verb ampliates the subject to supposit for possible cars. What of 'A chimera is conceivable' ('chimera' is ambiguous, but in one sense means an impossible combination of the head of a lion, the body of a goat and the tail of a serpent)? — Buridan claimed the modal '-ble' here ampliates only for possibles (so the proposition is false); others, e.g., Marsilius of Inghen in the next generation, thought such verbs ampliate for the imaginable, even the impossible (so the proposition is true). More problematic is the supposition of the subject in a proposition of the form 'Every S is necessarily P'. Buridan claimed that 'necessarily' again ampliates the subject to what is possible, so that 'Some S might not be P' is its contradictory. William of Ockham disagreed. He eschews the language of ampliation, and thinks that 'Some S might not be P' is ambiguous between 'Something which is S might not be P' and 'Something which might be S might not be P', but 'Every S is necessarily P' is not ambiguous, and can only mean 'Everything which is S is necessarily P'. Thus one reading of 'Some S might not be P' contradicts 'Every S is necessarily P', the other does not. Ockham is arguably truer to the everyday understanding of modal propositions than Buridan, who has a tendency to regiment language to his theory, and in the face of opposition responds that language is a matter of convention and he intends to use words the way he wants.

However, none of this explains subalternation. That comes from the theory of modes of common personal supposition, that is, of the supposition of general terms for the things they signify. There are two divisions, into determinate and confused supposition, and of confused supposition into confused and distributive and merely confused. Broadly, the divisions were characterised syntactically in the thirteenth century and semantically in the fourteenth, though accompanied by syntactic rules. Determinate supposition is that of a general term suppositing "for many as for one", as do both terms in 'Some S is P'; confused and distributive that of a term suppositing "for many as for any", as do both terms in 'No S is P'; merely confused that of a term like 'P' in 'Every S is P' or in 'Only Ps are Ss'. In confused and distributive supposition, one can descend (as they termed it) to every singular, indeed, to a conjunction of singulars, replacing the term in question by singular terms: 'Every S is P' entails 'This S is P and that S is P

and so on for all *S*s', so '*S*' in the original has confused and distributive supposition. This kind of descent is invalid for '*P*' in 'Every *S* is *P*'. One can ascend from any singular (so 'Every *S* is this *P*' entails 'Every *S* is *P*') but one can only descend through what was called a "disjunct term": 'Every *S* is *P*' entails 'Every *S* is this *P* or that *P* and so on'. In determinate supposition, one can descend to a disjunction of singulars, and ascend from any singular. Thus is subalternation explained: 'Every *S* is *P*' entails 'This *S* is *P* and that *S* is *P* and so on', which in turn entails 'Some *S* is *P*'. From confused and distributive supposition to determinate supposition is valid, but not conversely.

Buridan used the doctrine of supposition, and in particular, the notion of distribution in confused and distributive supposition, to provide an alternative to Aristotle's explanation of the validity of syllogisms.

> It should be noted that by these three conclusions, that is, the sixth, seventh and eighth, and by the second, the number of all the modes useful for syllogizing in any of the three figures both direct and indirect is made manifest.

The second conclusion showed that nothing follows from two negative premises, the sixth and seventh that the middle term must be distributed, and the eighth that any term distributed in the conclusion must be distributed in its premise.

4.3 Obligations

Logic lay at the heart of the medieval curriculum, and a further distinctive medieval doctrine was the mainstay of the education in logic, that of obligational disputations. This was a disputation between an Opponent and a Respondent, where the Opponent poses various propositions, as he chooses, and the Respondent is obliged to grant them, deny them or express doubt about them according to closely circumscribed rules — hence the description "(logical) obligations". There were several types of obligation: let us concentrate on just one, *positio*. In *positio*, the Opponent starts by describing a hypothetical situation and posing (or "positing") a certain proposition, the *positum*. The Respondent must accept it, unless it is explicitly contradictory; in "possible *positio*", provided it could be true. E.g., suppose as hypothesis that Socrates is not running, and take as *positum*, 'Every man is running'. The Respondent accepts this, and the disputation now starts. The Opponent proposes a succession of propositions; each proposition is "relevant" if it follows from (*sequens*) or is inconsistent

with (*repugnans*) the *positum* or any proposition previously granted, or the contradictory of one previously denied; otherwise it is "irrelevant". If it is relevant, the Respondent must grant it if it is *sequens* and deny it if it is *repugnans*; if irrelevant, he must grant it if it is known by the participants to be true, deny it if known to be false, and express doubt if its truth or falsity is unknown — a classic example is 'The king is sitting', which is standardly doubted if irrelevant. Here is a typical sequence of challenge and response:

Opponent	*Respondent*
Suppose Socrates is not running	
Positum: Every man is running	Accepted (possible)
Socrates is running	Denied (irrelevant and false)
Socrates is a man	Denied (relevant and *repugnans*)

If the Respondent makes a mistake (that is, grants contradictories, or grants and denies the same proposition) or after a certain agreed time, the disputation ends and an analysis of the disputation ensues.

Not every obligation is as simple as this. Walter Burley, who wrote a treatise on *Obligations* in 1302 which is usually credited as representing the standard doctrine, noted that there were certain tricks an Opponent could use to force the Respondent to grant any other falsehood compatible with the *positum*. For example:

Opponent	*Respondent*
Positum: Every man is running	Accepted (possible)
Socrates is not running or you are a bishop	Granted (irrelevant and true, since by hypothesis Socrates is not running)
Socrates is a man	Granted (irrelevant and true)
Socrates is running	Granted (relevant and *sequens*)
You are a bishop	Granted (relevant and *sequens*)

Once one has understood how the Respondent was forced to concede the falsehood 'You are a bishop' (assuming it is false), one can see that the Respondent can be forced to concede any falsehood whatever.

Like noughts-and-crosses (*aka* tic-tac-toe), the rules mean that there is always a winning strategy for the Respondent — keeping a clear head, the responses can be kept consistent. But mistakes are easy, because of the way relevant and irrelevant proposition must be so differently dealt with. If a *positum* really is inconsistent, it should not have been accepted to start

with. Or the disputation may exploit a paradox. Consider this example:

Opponent	*Respondent*
Positum: A man is an ass or nothing posited is true	Accepted (the second disjunct could be true)
A man is an ass	Denied (irrelevant and false)
Nothing posited is true	Granted (relevant and *sequens*)
The *positum* is true	Granted (relevant and *sequens*?)
Something posited is true	Granted (relevant and *sequens*)
Time's up.	

Contradictories have been granted. So has the Respondent made a mistake? Burley points out that the *positum* is an insoluble. Insolubles were, in Ockham's famous phrase, so called not because they could not be solved but because they were "difficult to solve". They constitute various kinds of logical paradox, including the Liar paradox itself: 'What I am saying is false'. It seems that this cannot be true, since if it were, it would, as it says, be false; and it cannot be false, for if it were, things would be as it says, so it would be true.

4.4 Insolubles

A variety of solutions to the Liar paradox were explored during the medieval period. Nine solutions were listed in Thomas Bradwardine's treatise on *Insolubles* in the early 1320s; fifteen are listed in Paul of Venice's *Logica Magna* ('The Great Logic') composed during the 1390s. The majority fall into three classes: the cassationists (*cassantes*), who claim that nothing has been said; the restrictionists (*restringentes*), who claim that no term can refer to a proposition of which it is part; and those, like Bradwardine, who diagnose a fallacy *secundum quid et simpliciter* (of relative and absolute), following Aristotle's comments in *Sophistical Refutations* ch. 25. The cassationist solution is known almost entirely by report by logicians who reject the suggestion, only one surviving text, from the early thirteenth century, advocating it. The idea is that any attempt to construct a proposition containing a term referring to this very proposition, fails on grounds of circularity to express any sense. More popular, at least before Bradwardine's devastating criticisms, was the restrictionist solution, sometimes in a naive version, similar to the cassationist story but inferring not that nothing had been said, but that the term trying to refer to the proposition of which it is part, in fact must refer to some other proposition of which it is not part — its scope for referring is thus restricted. A more sophisticated

version, put forward by Burley and Ockham, among others, proposed that the restriction only applied to insolubles, and prevented a term supposing for propositions of which it is part, or their contradictories. For example, Burley's diagnosis of the error in the Respondent's responses in the obligation above was that the proposition 'The *positum* is true' should not be granted, for 'the *positum*' cannot refer to this *positum*, for part of the *positum* contradicts the proposition of which 'the *positum*' is part. 'The *positum*' must refer to some other *positum*, so the proposition is irrelevant and should be responded to according to what holds of that *positum*. In any case, contradiction is avoided.

Bradwardine attacked the restrictionist view mercilessly, pointing out how implausible it was. His own view was taken up directly by very few (Ralph Strode, writing a generation later in the 1360s, was one of his champions), but he seems to have indirectly influenced most of the later proposals. The central idea to all these subsequent solutions is that an insoluble says more than appears on the surface. For whatever reason (and the reasons were multifarious), an insoluble like 'What I am saying is false' says not only that it is false but also that it is true — all insolubles, or perhaps all propositions, say implicitly of themselves that they are true. Hence no insoluble can be true, since it is self-contradictory. All insolubles are false.

5 Renaissance to Enlightenment

5.1 The Renaissance

During the fifteenth century a major change came over European logic. Some people have tied this change closely to the French logician Petrus Ramus (Pierre de la Ramée, 1515–72), who for his Master's degree in 1536 defended the thesis that 'Everything said by Aristotle is a pack of lies'; logic texts that are seen as influenced by Ramus are often referred to as Ramist Logic. But there may be a misunderstanding here. As a masters' student Ramus may well have been given his thesis title by his teachers — so he was being required to defend an obvious falsehood rather in the spirit of the obligational disputations that we described in §4 above. In fact his logic was not at all anti-aristotelian, but it does illustrate a general trend to relate logic to humanism.

This trend can be traced back earlier than Ramus. In fact some of its main features are already visible in the colourful Majorcan eccentric Ramon Llull (c. 1300), who proposed to use logic as a tool for converting North African Muslims to Christianity. Llull seems to have had little influence in

his lifetime, but many later logicians have seen his ideas as prophetic. Four features of his work are worth recording.

First, Llull addressed his logic to the general public, not just to university students and colleagues. (He was deported from Tunis three times for attempting public debates with Muslims there.) During the seventeenth and eighteenth centuries, most publications in logic were for general readers, particularly those with an interest in raising their level of culture. In Britain the authors were often literary figures rather than university teachers; we have logic texts from the poets John Milton (17th century), Isaac Watts (18th century) and Samuel Coleridge (early 19th century). Inevitably these works avoided all the subtler points of Scholastic logic and said more about general improvement of the mind.

Second, Llull wanted to use logic as an instrument of persuasion. In the early 15th century Lorenzo Valla argued that the central notions of logic should be not deduction but evidence and testimony; the best logician is one who can present a sound case persuasively. This whole period saw debates about how to speak both to the 'heart' and to the 'mind' (as Blaise Pascal put it in the 17th century). For example one way of catching the interest of the listener or reader is visual display. Llull himself had some strange display consisting of rotating disks with Latin words written on them — we will say more on these below. Several writers devised ways of making logic itself more appealing by presenting it as a 'game'; for example in the 16th century Agostino Nifo wrote a 'Dialectica Ludicra', which one might translate as 'Logic by games'. This trend towards associating logic with games has become a permanent feature of Western logic. Lewis Carroll joined it in 1887 with his book *The Game of Logic*. Today Katalin Havas in Hungary uses games to teach logic to schoolchildren, and Johan van Benthem in the Netherlands does something similar at a more advanced level, using some elementary mathematical game theory to advertise epistemic logic. It's worth noting here that in the late 20th century game theory was used partly to restore links between logic and probability, which (as we remarked in §3) were broken when probability theory became an independent discipline.

Third, this period saw logic drawing closer to mathematics, in the sense that logical deductions came to be seen more as calculations. Exactly what Llull contributed here is unclear, but many later logicians were inspired by his use of a mechanical device for making logical points. Some people even honour him as a forerunner of computer science. Leibniz named Llull as someone who had anticipated Leibniz's own project (on which more below) for building a logical calculus based on mathematics. We should

also mention the mathematicians Leonhard Euler (18th century) and John Venn (19th century) who gave us respectively Euler Diagrams and Venn Diagrams as ways of using visual display to help logical calculations.

The fourth prophetic feature of Llull's approach to logic was his use of classifications. His rotating disks were meant to illustrate different combinations of properties from a given set. Leibniz saw this as an anticipation of his own view of logic as a 'combinatorial art'. But it was also an anticipation of the enormous interest that some logicians of this period took in classification and cataloguing. Ramus was famous for his binary classifications; in Christopher Marlowe's play *The Massacre at Paris* (c. 1592) Ramus is murdered for being 'a flat dichotomist'. It was during this period that notions from Aristotle's theory of definition, such as 'genus', 'species' and 'differentia', were adapted to provide a structure for biological taxonomy. Some of the least appealing logic texts of the period are long catalogues of logical definitions, for example the 236-page 'Compendium' of Christian Wolff's *Logica* published in the mid 18th century by Frobesius.

5.2 Leibniz

The most powerful logician of this period was Gottfried Leibniz (1646–1716). He was also a mathematician, in fact one of the founders of the differential and integral calculus. Some of his most lasting contributions to logic are about combining logic and mathematics. He wrote several papers developing a logical calculus of 'coincidence', i.e. identity. He devised a way of translating definitions into numbers, so that logical properties of the definitions could be checked by arithmetical calculation. Above all he is remembered for his project of designing a 'universal characteristic', by which he seems to have meant an ideal language in which all human reasoning can be expressed in a form that can be checked by calculation. He imagined a day when scholars or lawyers would resolve their differences by writing down their arguments in his language and saying to each other '*calculemus*' ('let us calculate'). The project never came anywhere near completion, but Leibniz's calculi of identity and definitions were certainly intended to be contributions to it.

It might seem a short step from claiming that all logical proofs can be checked by calculation, to claiming that all logical problems can be solved by calculation. Leibniz himself seems never to have taken this step. It was left to the 1930s to sort out these claims. By that date, higher-order logic had replaced syllogisms, and a much wider range of logical problems could

be formulated. Thanks to work of Kurt Gödel and Alan Turing above all (for which see Barry Cooper's chapter 'Computability Theory' in this volume), we now know that Leibniz's instincts were sound: for higher-order logic we can check by elementary calculation whether a supposed proof is correct, but by contrast there is no mechanical method of calculation that will tell us whether any given sentence of the language of higher-order logic is a logical truth. The same holds for first-order logic.

Since the mid 20th century, Western modal logicians have often used the notion of possible worlds: a sentence is necessarily true if and only if it is 'true in' every possible world. The notion is often credited to Leibniz, who certainly did talk about alternative worlds that are possible but not actual. But he himself didn't use this notion for logical purposes. In any case one might argue that the 'possible worlds' of modern modal logicians are not alternative worlds but reference points or viewpoints, as when we say that something will be true at midday tomorrow, or that something is true in Smith's belief system. (The study of things being true or false at different times goes back to Aristotle, though Ibn Sīnā may have been the first to build a logic around it. The study of the logic of belief systems is much more recent in the West; the Jaina logicians got there first with their notion of perspectives, *anekāntavāda*.)

5.3 The philosophical turn

During the late 18th and early 19th centuries several metaphysicians made attempts to base logic on a theory of rational thinking. The results of these attempts were strictly not a part of logic at all, but comments on logic from the outside. But we need to mention them, both because they had an influence in logic, and because their importance in Western logic has been exaggerated in a number of recent comparisons between Western and Indian logic.

Thus Immanuel Kant (1724–1804) believed he had identified a central core of logic, which he called 'pure general logic' or 'formal logic'. The defining feature of pure general logic was that it studies the absolutely necessary laws of thought without regard to subject matter. All other logic was dependent on this central core. Some later logicians agreed with Kant that there is a central 'genuine logic'; in a few cases their definition of it (which was nearly always different from Kant's) influenced the direction of their research, and in this way Kant's notion indirectly affected the history of logic.

Among these later logicians, pride of place goes to Gottlob Frege (1848–

1925), who aimed to show that arithmetic and mathematical analysis are parts of pure general logic. Frege achieved a combination of depth and precision that had certainly never been seen in logic before him, and has rarely been equalled since. But his actual historical influence is another matter; it is quite subtle to trace and has often been overblown. For example one reads that Frege founded mathematical logic; but as we will see in section §6, both the name 'mathematical logic' and its initial programme were proposed by Giuseppe Peano, quite independently of Frege.

The period around 1800 also saw the formulation of some 'fundamental laws of thought', such as the Law of Non-Contradiction and the Law of Excluded Middle. These two laws were popularised in the 1830s by the Scottish metaphysician William Hamilton in his lectures on logic. Formulations of the laws have changed over the years, and today few people would recognise Hamilton's versions. The broad sense of the Law of Non-Contradiction is that it can never be correct both to assert and to deny the same proposition at the same time. The broad sense of the Law of Excluded Middle is that every proposition either can be correctly asserted or can be correctly denied (though we might not know which).

The claim that these are fundamental laws bears little relation to traditional practice in Western logic. True, many logicians from Aristotle onwards said things that look like the laws; but one has to allow for simplifications and idealisations. In fact many traditional Western logicians accepted that a proposition could fail to be straightforwardly true or false in several circumstances: for example if it was ambiguous, or a borderline case, or paradoxical, or a category mistake. Likewise many traditional logicians were happy to say that a sentence or proposition (the two were often confused) could be true from one point of view and false from another. Perhaps no Western logician pursued this last point to the same extent as the Jaina logicians, though Ibn Sīnā came close at times. In any case, to treat the Laws as a basic difference between Western and Indian logic is certainly a distortion.

In the twentieth century it became common to use purpose-built formal systems of logic. The Laws then served as ways of classifying formal systems. For example, it is crucial to distinguish Excluded Middle, the claim that every proposition or its contradictory is true, from Bivalence, that every proposition is either true or false.

6 Transitional

Aristotelian logicians through the ages claimed that logic can free us from errors in reasoning. The claim was fraudulent. From Aristotle onwards, logicians made catalogues of types of fallacious argument, in terms of kinds of ambiguity. But until the early 19th century no aristotelian logician made any serious attempt to discover whether the ambiguities are causes or symptoms of the breakdown of reasoning, or what are the best ways of protecting ourselves against falling into fallacies. (By contrast the Buddhist logicians appreciated early on that making errors is a failure of our cognitive apparatus, and at least part of the reason for errors must lie in facts about that apparatus and its powers of 'constructive thinking', *vikalpa*.)

In the West the first step into intellectual honesty seems to have been taken by the philosopher Jeremy Bentham, who never published his views on logic — they were reported later by his nephew George Bentham. Jeremy Bentham argued that a good strategy for avoiding errors is to translate arguments into what we now call set theory. One should identify the classes of objects that one is talking about, and express the argument in terms of relations between these classes.

These remarks were ahead of their time. Bentham couldn't have foreseen another result of translating logic into set theory. Namely, set theory provides a universe of abstract objects and a set of sharply defined rules for operating with them. So it gives us building materials and a space for developing logic in a way that had never been possible before.

The three logicians most responsible for moving logic into set theory were George Boole (1815–64), Giuseppe Peano (1858–1932) and Alfred Tarski (1901–83, whom we postpone for a moment). Boole is well known for boolean algebra; in his hands it was an algebra of classes. Peano was interested in avoiding errors in mathematics. He believed that the best protection against errors was to translate mathematical arguments wholesale into a symbolic language of his own invention, and he wrote down rules for operating with this language. With this he had invented a new discipline. Peano named this discipline 'mathematical logic' and did vigorous propaganda for it. His most important convert was Bertrand Russell, whose book *Principia Mathematica* (written with Alfred North Whitehead) became the showpiece for the Peano programme. Peano had little interest in traditional logic and he largely made things up as he went along; this certainly helped him to break free from some unhelpful traditional views. With hindsight, perhaps his biggest breakthrough was that he formalised entire complex

arguments, not single inference steps as in the aristotelian tradition. (Recall §1.4 above.) This forced him, and the later logicians who plugged the holes in his work, to rethink logic from the ground up. Logicians would no longer state more and more complicated inference rules. Instead they would isolate fundamental ideas and principles, that you could use to derive whatever arguments you were interested in. For example logicians had been using existential quantification since Aristotle, but Peano was one of the first logicians to isolate the existential quantifier (the symbol $\exists$ is his) so that it could be used in any context. One consequence of Peano's work was that for half a century, mainstream logicians more or less abandoned any interest in natural language arguments.

Peano worked in a kind of mishmash of logic and set theory. Russell brought some order into the system with his 'theory of types'. At the bottom level one has expressions for talking about individuals; then there are expressions for talking about relations between individuals, then expressions for talking about relations between relations between individuals, and so on. David Hilbert in 1928 published what became accepted as the standard version of this logical system; it was known as the 'logic of finite types', or 'higher-order logic'. One could separate out parts of it. For example if one stopped at the relations between individuals, one got 'first-order logic'. Cutting out even more, one got 'propositional logic'. This was probably the main source of the idea that logicians study formal systems called 'logics'. In the 1920s and 1930s logicians came to realise that they could bring their subject to levels of precision and accuracy undreamed of before, by defining formal systems. A definition of a formal system would say precisely what symbols were used (and even list the variables exactly — people never did that before about 1930). The rules for manipulating the symbols were exact mechanical rules that could be used by a machine (and today often are). It was largely thanks to this that the new subject of computability theory, founded in the 1930s by Alan Turing, came to be regarded as a branch of logic, although it had no real antecedents in earlier Western logic.

Now we can explain Tarski's role. Tarski studied various fundamental notions of logic such as truth and logical consequence, which are meta-level notions, i.e. they are used for *talking about* formal systems rather than being *expressed in* them. He showed that these notions have set-theoretical translations that for practical purposes we can use instead of the original ones. This had an effect that he didn't foresee at first. We can in principle throw the whole of logic — both the formal systems themselves and the meta-level study of them — into the formal system of first-order set theory.

Tarski's work neatly complemented ideas of Hilbert. Hilbert had argued in the 1920s that we can discover useful mathematical facts by studying the ways in which mathematicians move symbols around on the page, without bothering with the meanings of the symbols; this would be a way of studying mathematics from the outside, and Hilbert coined the name 'metamathematics' for it, to distinguish it from mathematics proper. (Hilbert was a leading mathematician of the early 20th century, and his own contributions ranged far beyond metamathematics.) Hilbert's metamathematics was an important step towards the invention of digital computers, which do move symbols around without any understanding of what they mean. It also led to a new mathematical theory of formal proofs; Gerhard Gentzen proved deep results in this theory, and most of today's proof calculi trace back to him in one way or another.

Within the framework of first-order set theory we can do logic without any philosophical assumptions; we need not even agree on why the principles of set theory are true or usable, so long as we agree to use them. Old philosophical questions about logic haven't gone away, but today we can separate them off as questions *about* logic, not questions that logicians themselves need to think about. (It's like the difference between doing history and doing philosophy of history.) One can argue that some Indian logic compares better with Western *philosophy* of logic than it does with Western logic proper. Recall here the remarks about Frege and others in subsection §5.3 above.

Frege's writings contain by far the most penetrating account of what needed to be changed and corrected in aristotelian logic. So it was ironic that a contradiction in his own work seemed for a time to threaten the coherence of set theory. Frege had a principle, known today as the Unrestricted Comprehension Axiom, which said that for every property P there is a class whose members are exactly those things that have the property P. We get a contradiction by applying this to the property 'class which is not a member of itself'. If C is the class of all classes that are not members of themselves, then we easily show that C is a member of itself, and that it isn't a member of itself. This contradiction was pointed out by Russell in 1901, and is known as Russell's Paradox. Ernst Zermelo (who independently noticed the paradox) developed a way of doing set theory that — as far as we know, one century later — doesn't lead to paradoxes. We do know that there is a price to pay: if Zermelo's set theory is consistent, then it leaves some important questions unanswered, such as the size of the continuum.

Could we perhaps *prove*, by mathematical or logical methods, that Zer-

melo's system of set theory will never lead to a contradiction? Hilbert hoped that his metamathematics would give an answer. The aim would be to prove, by studying the ways in which a mathematician using Zermelo's system moves symbols around, that this mathematician could never reach the point of writing '0 = 1'. The proof should use only what Hilbert called 'finitist' reasoning about the symbols, so as to avoid circularity. In one of the strongest *tours de force* of logic — perhaps of mathematics too — Gödel showed in 1931 that this is impossible. To state his result briefly but a little imprecisely, Gödel showed that there is no argument that can be expressed in elementary arithmetic and proves that elementary arithmetic itself is consistent, unless in fact elementary arithmetic is inconsistent.

In the first decades of the 20th century Russell's Paradox made many mathematicians and philosophers suspicious of set theory, and it certainly encouraged some logicians to develop logical systems that didn't depend on set theory. One important example was the intuitionist system which L. E. J. Brouwer devised as a basis for mathematics. His central notion was not truth but mental construction. A meaningful sentence is either true or false; but we may not be able to make a mental construction that justifies the sentence, or one that justifies its negation. So intuitionists do not accept the Law of Excluded Middle in the form of the law 'Either p or not-p'.

Some more radical followers of Brouwer constructed systems of logic which avoid the notion of negation altogether. This makes a curious contrast with the frequent appearances of double negation in Buddhist logic under the influence of the doctrine of *apoha*. The contrast is not hard to explain when we remember that these Western logicians were studying mental constructions while the Buddhist logicians were thinking about classification by universals. A general point to draw from this is that the history of logic makes little sense if one doesn't appreciate what the various logicians were aiming to do with their logic. We hope our short survey has illustrated this point in a range of ways.

7 References

Classical Greece

- Plato, *The Collected Dialogues*, Eng.tr. Edith Hamilton and Huntington Cairns, Princeton University Press, Princeton 1963.

- Aristotle, *The Complete Works (2 vols.)*, Eng.tr. Jonathan Barnes, Princeton University Press, Princeton 1984.

- Aristotle, *Prior Analytics Book 1*, Eng.tr. Gisela Striker, Clarendon Press, Oxford 2009.
- Paul Thom, *The Syllogism*, Philosophia Verlag, Munich 1981.
- Jonathan Barnes et al., 'Logic', in *The Cambridge History of Hellenistic Philosophy*, ed. K. Algra et al., Cambridge University Press, Cambridge 1999, pp. 77–176.

The Roman Empire

- Richard Sorabji (ed.), *Aristotle Transformed: The Ancient Commentators and Their Influence*, Duckworth, London 1990.
- The series *Ancient Commentators on Aristotle*, ed. Richard Sorabji, Duckworth, London makes available several of the main texts in English translation with commentary. Published volumes include Alexander of Aphrodisias on several parts of the *Prior Analytics*, Porphyry on the *Categories* and Ammonius on part of *On Interpretation*.
- Porphyry, *Introduction*, Eng.tr. and ed. Jonathan Barnes, Clarendon Press, Oxford 2003.

Arabic Logic

- Ibn Sīnā, *The Deliverance: Logic*, Eng.tr. Asad Q. Ahmed, Oxford University Press, Oxford 2010.
- Khaled El-Rouayheb, *Relational Syllogisms and the History of Arabic Logic, 900–1900*, Brill, Leiden 2010.
- Tony Street, 'Arabic and Islamic Philosophy of Language and Logic'. The Stanford Encyclopedia of Philosophy, ed. E. Zalta, URL = `http://plato.stanford.edu/entries/arabic-islamic-language`.
- The website `http://wilfridhodges.co.uk` contains a number of English translations of Arabic logic, mostly Ibn Sīnā but also the later textbook *Shamsiyya*.

The Scholastics

- Mediaeval and Renaissance logic (Handbook of the history of logic vol.2), edited by Dov M. Gabbay and John Woods (Amsterdam: North-Holland 2008).

- The Cambridge History of Medieval Philosophy, vol.1 Part II (chs.10-15), edited by Robert Pasnau (CUP 2010).
- Tuomo Aho and Mikko Yrjönsuuri, 'Late medieval logic', in Haaparanta (2009) below, pp. 11–78.
- Stephen Read, 'Medieval Theories: Properties of Terms, The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), ed. E. Zalta, URL = http://plato.stanford.edu/archives/fall2008/entries/medieval-terms
- Paul Vincent Spade and Stephen Read, "Insolubles", The Stanford Encyclopedia of Philosophy (Winter 2009 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2009/entries/insolubles.
- John Buridan, *Summulae de Dialectica*, Eng. trans. Gyula Klima (New Haven: Yale UP 2001).

Renaissance logic

- E. J. Ashworth et al., 'The defeat, neglect, and revival of scholasticism', in *The Cambridge History of Later Medieval Philosophy*, Cambridge University Press, Cambridge 1982, pp. 785–852.
- Walter J. Ong, *Ramus, Method and the Decay of Dialogue: From the Art of Discourse to the Art of Reason*, Harvard University Press, Cambridge Mass. 1984.
- Gottfried Wilhelm Leibniz, *Logical Papers*, trans. and ed. G. H. R. Parkinson, Clarendon Press, Oxford 1966.
- Massimo Mugnai, *Leibniz' Theory of Relations*, Steiner, Leinen 1992.

Transitional

- George Boole, *An Investigation of the Laws of Thought, on which are Founded the Mathematical Theories of Logic and Probabilities*, Dover, New York 1958. (Original 1854.)
- D. Hilbert and W. Ackermann, *Principles of Mathematical Logic*, Chelsea, New York 1950. (German original 1928, edited from Hilbert's lectures in Göttingen 1917–22.)

- Alfred Tarski, *Introduction to Logic and to the Methodology of Deductive Sciences*, 4th edition ed. Jan Tarski, Oxford University Press, New York 1994. (Original 1936.)
- Jean van Heijenoort, *From Frege to Gödel, A Source Book in Mathematical Logic, 1879–1931*, Harvard University Press, Cambridge Mass. 1967.
- Leila Haaparanta (ed.), *The Development of Modern Logic*, Oxford University Press, Oxford 2009.

Two Indian Dialectical Logics: *saptabhaṅgī* and *catuṣkoṭi*[*]

FABIEN SCHANG

Abstract. A rational interpretation is proposed for two ancient Indian logics: the Jaina *saptabhaṅgī*, and the Mādhyamika *catuṣkoṭi*. It is argued that the irrationality currently imputed to these logics relies upon some philosophical preconceptions inherited from Aristotelian metaphysics. This misunderstanding can be corrected in two steps: by recalling their assumptions about truth; by reconstructing their ensuing theory of judgment within a common conceptual framework.

Contents

[*]I want to thank Jonardon Ganeri for a fruitful informal correspondence with him, as well as Shrikant Joshi for his educated support about sanskrit and pali expressions and the anyonymous referee for his/her helpful requirements.

1 Two logics?

A note on Indian "logics" is in order, to begin with. By a logic, it is ordinarily meant a specific set of consequence relations between a set of premises Γ and a conclusion B such that, for every formula $A \in \Gamma$, if A is true then so is B. Formally: if $v(A) = T$ then $v(B) = T$, where v is a valuation function from a set of formulas to a set of truth-values. But such a modern definition of logic as a set of rules for truth preservation cannot be properly applied to ancient logics, including those from India. Rather, ancient and medieval logics include epistemology in the scope of the formal discipline: how to assess the content of a judgment isn't separable in Aristotle's *Organon* or the Port-Royal Logic, for instance, and Indian logics are not an exception.

The epistemological import of Indian logics largely accounts for their peculiar content; the metaphysical assumptions that underlie these Indian schools of philosophy also results in specific theories of truth, and the main aim of the present paper will be to give a formal presentation of the ways to produce a judgment or predication such as "S is P" or "S is not P" (where S is the subject-term and P the predicate-term). As a matter of rule, Indian logics are about judgments and not about the sentences expressing them; we will restrict our attention to two such cases: the Jaina *saptabhaṅgī*; and the *catuṣkoṭi* from the Buddhist school of Mādhyamaka (literally, "Middle Way").

As a general rule, the logics emerging from the Jaina and Mādhyamika schools include both a theory of knowledge (about how to come to know something) and a complementary theory of judgment (about how to express this something known). Concerning the theory of knowledge, the *nayavāda* is a Jaina theory (*vāda*) of standpoints (*nayas*) that includes seven kinds of justification for the truth of a sentence.[1] Furthermore, a set of seven (*sapta*) distinct judgments (*bhaṅgī*) can be made about a given

[1]The seven kinds of justification (*nayas*) include metaphysical, physical and grammatical features. These are the following: *naigama-naya* (non-distinguished standpoint); *saṃgraha-naya* (collective standpoint); *vyavahāra-naya* (particular standpoint); *ṛju-sūtra-naya* (momentary viewpoint); *śabda-naya* (synonym viewpoint); *samabhirūḍha-naya* (etymological viewpoint); and, finally, *evaṃbhūta-naya* (momentary etymological viewpoint). For instance, "the existence of an entity such as a pot, depends upon its being a particular substance (an earth-substance), upon its being located in a particular space, upon its being in a particular time, and also upon its having some particular (say, dark) feature. With respect to a water-substance, it would be non-existent, and the same with respect of another spatial location, another time (when and where it was non-existent), and another (say, red) feature. It seems to me that the indexicality of the determinants of existence is being emphasized here." ([12], p. 132).

topic. There is no causal relation between the number of standpoints and judgments, however. After all, the Greek skeptic Agrippa proposed five kinds of justification while sticking to an Aristotelian or bivalent view of judgments: either S is P or S is not P, period. Rather, the number of the Jaina judgments is due to their endorsement of a metaphysical pluralism according to which reality is many-faceted and cannot be restricted to a unique predication. As to the Mādhyamika school and its founder Nāgārjuna ($\simeq$ 100 C.E.), they did not present a competing theory of knowledge but advanced four (*catuṣ*) main sorts of stances (*koṭi*) for any subject-matter.

As noted in [15], "logic is not metaphysically neutral", and the difference between the Jaina seven and Nāgārjuna's four judgments is due to their rival views of truth. Ganeri advances (in [6], p. 268) a relevant distinction between three semantic views of truth-assignment, namely: doctrinalism, skepticism, and pluralism. According to the doctrinalist view, "it is always possible, in principle, to discover which of two inconsistent sentences is true, and which is false." This doctrine is related to Aristotle's two-valued logic, where only two judgments can be made about any subject-matter (S is P, S is not P) and only one of which comes to be accepted as true while the other is to be false. Bivalence is the logical cornerstone of such a doctrine and entails that every judgment is either a truth- or a falsity-claim, i.e. a statement. Skepticism and relativism challenge this binary view in opposite directions. According to skepticism, "the existence both of a reason to assert and a reason to reject a sentence itself constitutes a reason to deny that we can justifiably either assert or deny the sentence", so that some sentences can be taken to be neither true nor false. Conversely, the pluralistic watchword is "to find some way conditionally to assent to each of the sentences, by recognizing that the justification of a sentence is internal to a standpoint"; in this sense, one and the same sentence can be taken to be both true and false depending upon the condition under which its content is assessed.

We take these three doctrines of truth-assignment to be the crucial path for a better understanding of Indian logics. While these have been dismissed by Western thinkers, as having "irrational" or "unintelligible" outlook[2], we suspect this uncharitable preconception to stem from a narrow reading of bivalence that takes Frege's modern logic as a standard for any

[2] "Manifoldness in this context is understood to include mutually contradictory properties. Hence on the face of it, it seems to be a direct challenge to the law of contradiction. However, this seeming challenge should not be construed as an invitation to jump into the ocean of irrationality and unintelligibility" ([12], pp. 129-30).

meaningful judgment. If so, the next sections insist upon the discursive and non-standard form of judgments in Jaina and Mādhyamika logics: it is still possible to preserve bivalence within these Indian theories and, thus, to preserve their intelligibility, but only if such a bivalence is not defined in Fregean terms and reformulated as a question-answer game between speakers.

2 Two opposite logics?

An intriguing feature of Jaina and Mādhyamika logics concerns their attitude towards inference: the relativist doctrine of truth seems to entail a fully inconsistent logic, whereas the skeptic doctrine of truth would entail a fully incomplete logic. This means that, for any sentences A and B, B seems to be inferred from every premise A in Jaina logic (say, J): A $\models_J$ B (for every B); whereas no sentence B would be inferred from A in Nāgārjuna's logic (say, N): A $\not\models_N$ B. Parsons described in [14] these cases in terms of ultimate eclecticism and complete nihilism, respectively[3].

Is Jaina logic a formal system of eclecticism, and Nāgārjuna's logic a system for nihilism? This is not so, at least for one simple reason: nihilism assumes that the premise A is accepted as true, while the coming exposition of Nāgārjuna's Principle of Four-Cornered Negation amounts to a denial of every sentence including A. As to the Jaina logic, the role of standpoints means that not every conclusion B can be inferred from A irrespective of the context in which A and B are assessed. This entails that not everything can be inferred from every given context, and Priest recalls this fact in [15] to make his own dialetheist reading of Jaina logic immune from triviality. We will return to this modern translation in Section 5.

Two Sanskrit notions will be introduced now, in order to throw some light upon the Jaina and Mādhyamika ways of doing logic. The first concept is *anekāntavāda*: this term means non one-sidedness and characterizes the Jaina conditional view of truth, according to which the truth of a sentence is never one-sided (*ekānta*) but always depends upon the context in which it is assessed. The second concept is *prasajya pratisedha* (see

[3] See [14], p. 141. Roughly speaking, eclecticism refers to the view that sentences of two different theories can be accepted consistently within a third embracing theory: $T_1 \models p$, $T_2 \models q$, $T_3 \models p$ and $T_3 \models q$. This is not the point of Jainism. As to nihilism, it refers to the belief that nothing is true. This is not the point of Mādhyamaka, either. The difference between such nihilists and the latter could be made clearer by the difference between atheism (negative assertion about the existence of God) and agnosticism (mere denial about the existence of God).

[5],[11],[13]); Mohanta mentions this concept in [13] as a non-relational negation which somehow corresponds to the contemporary denegation or illocutionary negation[4]. In contrast to the Jaina conditions for truth - assignment, the Mādhyamikas defended the view that being dependent upon anything else is a sufficient ground for denying a corresponding predication: S cannot be said to be P or not to be P whenever S is not self-originated and is caused by another substance than itself. This refers to the two-truths doctrine and its distinction between *absolute* truth (*paramārtha-satya*) and *conventional* truth (*saṃvṛti-satya*) in the Mādhyamika's *sūnyavāda* (doctrine of emptiness); we will see how this doctrine leads to Ganeri's previous distinction between the pluralist and skeptic conditions for truth-assignment. While the Jains favor a contextual theory of affirmation, Nāgārjuna endorses a peculiar use of denial which is to be rigorously distinguished from negative assertion and departs from falsity-assignment. Thus, saying that S is not P results in an ambiguous judgment between affirming that the sentence "S is P" is false and denying that "S is not-P" is true. From an Aristotelian or doctrinalist approach, affirming S not to be P and denying S to be P are synonymous with each other; from a Mādhyamika or skeptic approach, however, P may be denied to be true of S without being affirmed to be false of S. Such a confusion amounts to a harmful confusion between two sorts of Indian negations (*pratisedha*), namely: the previous *prasajya pratisedha* and *paryudāsa pratiṣedha*, which is a relational (see [13]) or locutionary negation used by the later Navya school.

To sum up, Jaina and Mādhyamika logicians do oppose each other with respect to their underlying criterion for truth-assignment. Given two opposite sentences "S is P" and "S is not P", how to decide on the truth of either? The main difference between Jainas and Mādhyamikas lies in their answer to this question. Thus, Matilal claims (in [12], p. 129) that "the difference between Buddhism and Jainism in this respect lies in the fact that the former avoids by *rejecting* the extremes altogether, while the latter does it by *accepting* both with qualifications and also by reconcil-

[4]Illocutionary negation (denial, or denegation) has been defined by John Searle in [19]. Let the speech act F(p) = "I promise that I will come", where F is the act of promise and p the sentential content "I will come"; then its locutionary negation F($\sim p$) is "I promise that I will not come", while its illocutionary negation $\sim$(Fp) is "I do not promise that I will come". Denial has been ordinarily rendered as a reversed turnstile $\dashv$, in reference to Frege's turnstile of assertion, while Keiff views it in [11] as a merely failed assertion $\nvdash$. In both cases, denial occurs as an operator; in **QAS**, however, denial is an operand (a logical value: the no-answer $\mathbf{a}_i = 0$).

ing them." It is worthwhile to note that these opposite modes of truth-assignment also foreshadow the contemporary opposition between semantic realism and anti-realism: [22] and [23] notice that the Jains countenance a correspondence theory of truth, whereas Siderits' comparison (in [21]) between Nāgārjuna's denials and Dummett's anti-realist semantics entails that Nāgārjuna's conception of truth doesn't transcend recognitional capacity by a given agent.

Before approaching this last problem about the relations between judgments, let us consider the way to describe their various admitted judgments within a clear and uniform formal semantics.

3 Two many-valued logics?

One of the primary aims of the paper is to insist upon the dialectical nature of Indian logics, i.e. their presentation in terms of speech-acts within an argumentative framework of questions and answers. To put it in other words, each truth- or falsity-assignment proceeds by means of an intermediary act of affirmation and denial. Importantly, we take the asymmetry between the pairs true-false and affirmation-denial to be the key for a better understanding of Indian logics. A number of logical techniques have been proposed in the literature to catch the dialectical or discursive feature of Indian logics: relational or possible-world semantics ([15]), dialogics ([8],[11]), and algebraic or many-valued semantics ([6],[15],[18],[20])[5].

In order to give a more fine-grained description of Jaina and Mādhyamika logics, we resort here to many-valuedness. Roughly speaking, the various ways of making a judgment require the introduction of alternative logical values beyond the doctrinalist values of truth and falsity. In the case of Jaina philosophy, no judgment uniquely claims plain truth or falsity because of its underlying one-many correspondence theory of truth: a given sentence partly describes a fact following the perspective from which its

[5]Gokhale rejected the many-valued interpretation of Jain logic because, according to him, a difference is to be made between epistemological and logical values. Thus: "The middle value designated by the term *avaktavyam* is therefore better understood as the epistemic middle rather as the logical middle. It is closer to the middle truth-value called 'undeterminable' of Kleene's three-valued system than to the Łukasiewiczian third truth-value called 'indeterminate'. (...) As a result we can say that *avaktavya* is not the third truth-value in the logical sense of the term, because it does not arise out of the violation of the laws of logic such as non-contradiction and excluded middle" ([7], p. 75). This objection assumes that every logical value should have an ontological import, but our purely algebraic viewpoint of logic does not require this and Belnap's four-valued system is an instance where all the logical values have an epistemological import.

content may be described.[6] In the case of Nāgārjuna's Principle of Four-Cornered Negation, it will be shown that the assumption of bivalence cannot make sense of the four negative stances together (see section 5). At the same time, the metaphysical pluralism of the Jains does not entail that new truth-values should be devised in addition to the Aristotelian framework of bivalence. Rather, these alternative logical values are various *combinations* of truth and falsity inside the initial set of values T (for true) and F (for false).

In particular, the Jaina theory of sevenfold predication (*saptabhaṅgī*) reminds one of Belnap's system of generalized truth-values and Shramko & Wansing's extension from 2 to n truth-values (see [3],[20]). Taking 2 = {T,F} as a basic set and its two elements of truth and falsity, an extension from 2 to 4 results from its powerset $\wp(2)$, that is the set of the subsets of 2. Thus **4** = {{T},{F},{T,F},∅}, and Belnap symbolized the new combinations of truth-values as {T,F} = B (for "both true and false") and ∅ = N (for "neither true nor false") in its four-valued logic FDE (First Degree Entailment). The same process can be applied indefinitely, leading to a set of $\wp(n)$ elements for any n-valued logic (where $n \geq 1$). Another such generalized set is $\wp(3)$, with $n = 3$ basic elements T, F and {T,F}. One of these generalized sets is

8 = {{T},{F},{B},{{T},{F}},{{T},{B}},{{F},{B}},{{T},{F},{B}},∅}.

We will see that the latter set can be made very similar to the Jaina semantics, even though the odd number of the seven Jaina judgments may surprise at a first blush. Moreover, Bahm rightly noted in [2] that Indian logics are not just formal combinations of truth-values but require a more comprehensive reading of their original texts.

For this purpose, we propose now a conceptual framework to grasp the rationale of Indian logics: a Question-Answer Semantics (**QAS**) that encompasses Belnap's generalizations and helps to account for the Mādhyamika's dialectical logic of Four-Cornered Negation.

[6] Sylvan noted that "Jainism apparently entailed a correspondence theory of truth" (p. 62), so that the Jain values have an ontological import that differs from Belnap's four values in FDE: a sentence *is* true and false (in some respects), rather than *told* true and *told* false. The difference between Jain and Aristotelian logic relies upon their underlying ontology: the latter takes a true sentence to correspond to a fact, while the former reject such a one-one correspondence between sentences of a language and states of affairs of the world. Thus Tripathi argued in [23] that "Jainism is a realistic system. It not only holds that reality is pluralistic, but also that reality is many-faced (*anantadharmātmakam vastu*)." ([21], p. 187) The Wittgensteinian *Bildtheorie* should be strictly kept apart from the Jain view of reality, consequently.

DEFINITION 1. A question-answer semantics is a model $\mathbf{QAS} = \langle \mathfrak{M}, \mathbf{A} \rangle$ upon a sentential language $\mathscr{L}$ and its set of logical connectives ©. It includes a logical matrix $\mathfrak{M} = \langle \mathbf{Q}, V, D \rangle$, with:
- a function $\mathbf{Q}(\alpha) = \langle \mathbf{q}_1(\alpha), ..., \mathbf{q}_n(\alpha) \rangle$ that turns any sentence α of $\mathscr{L}$ into a specific speech-act (the sense of which is given by appropriate questions about it);
- a set V of logical values (where $Card(V) = m^n$);
- a subset of designated values $D \subseteq V$.

It also includes a valuation function $\mathbf{A}$, such that the logical value $\mathbf{A}(\alpha) = \langle \mathbf{a}_1(\alpha), \ldots, \mathbf{a}_n(\alpha) \rangle$ of V that characterizes a statement by giving an ordered set of m sorts of answers to each question $\mathbf{q}_i$ in $\mathbf{Q}(\alpha) = \langle \mathbf{q}_1(\alpha), \ldots, \mathbf{q}_n(\alpha) \rangle$. This semantic framework results in a variety of logics $\mathcal{L} = \langle \mathscr{L}, \models_{\mathcal{M}} \rangle$ that include an entailment relation in a model $\models_{\mathcal{M}}$ such that, for every set of premises Γ and every conclusion α in $\mathscr{L}$, if $\mathbf{A}(\mathbf{\Gamma}) \subseteq D$ then $\mathbf{A}(\alpha) \subseteq D$: $\Gamma \models_{\mathcal{M}} \alpha$.

A crucial difference with the more familiar logics is the meaning of the semantics values in **QAS**: each element $\{\mathbf{a}_1(\alpha), \ldots, \mathbf{a}_n(\alpha)\}$ of $\mathbf{A}(\alpha)$ is a basic answer $\mathbf{a}_i(\alpha)$ (where $1 \geq i \geq n$) with the symbol 1 for *affirmations* (yes-answers) and the symbol 0 for *denials* (no-answers). Let us call by the general heading of "logical value" every such ordered set of answers, rather than the customary "truth-values": these values are a combination of yes-no answers to corresponding questions, whereas not every question is to be asked about the truth-value of a sentence in **QAS**.

Once the formal structure is set out for any question-answer game, let us have a closer look at our two Indian logics at hand while attempting to reconstruct their argumentative games.

4 Jaina's theory of seven-fold predication

It has been previously claimed that not everything can be derived from every premise from a Jaina perspective: meaningfulness presupposes that a restricted set of sentences can be accepted on the basis of certain premises in a given language, while the remaining sentences of the language should not be accepted. But the question is how the Jaina predications do make sense in a consistent set of statements. In particular, the Jaina theory of seven-fold predication (*saptabhaṅgī*) has been viewed as a challenge to Aristotle's logic.

According to Aristotle, the Principle of Non-Contradiction (PNC) is a universal law of thought that cannot be violated without committing its

opponent into plain nonsense. It is stated in [1] as follows:

"It is impossible for the same thing to belong and not to belong at the same time to the same thing and in the same respect." (Book IV, 1005b19-20)

An instant reflection suffices to see that the Jains did not oppose to this principle as it stands: their semantic pluralism relies upon a doctrine of conditioned, relative or partial truth (*syādvāda*). The Jaina philosopher Vādiveda Sūri (1086-1169 C.E.) displayed the following set of seven predications and witnessed the crucial role of *syād* ("arguably", or "in some respect") in every corresponding statement, where every predication expresses a conditioned judgment about a sentence[7]:

(1) *syād asty eva*: arguably, it (some object) exists.
(2) *syān nāsty eva*: arguably, it does not exist.
(3) *syād asty eva syān nāsty eva*: arguably, it exists; arguably, it does not exist.
(4) *syād asty eva syād avaktavyam eva*: arguably, it exists; arguably, it is non-assertible.
(5) *syād asty eva syād avaktavyam eva*: arguably, it exists; arguably, it is non-assertible.
(6) *syān nāsty eva syād avaktavyam eva*: arguably, it does not exist; arguably, it is non-assertible.
(7) *syād asty eva syān nāsty eva syād avaktavyam eva*: arguably, it exists; arguably, it does not exist; arguably, it is non-assertible.

Each of these predications is a combination of three basic semantic predicates (*mūlabhaṅgas*)[8], namely: assertion, or *truth*-claim; denial, or *falsity*-

[7]The *saptabhaṇgī* clearly departs from the Fregean logic of propositions, where a sentence expresses a thought and refers to a unique truth-value. To the contrary, the seven arguments of *nayavāda* assume that the meaning of a sentence is context-dependent and doesn't refer to some eternal entity as the True. Thus Matilal: "Realists or believers in bivalence (as Michael Dummett has put it) would rather have the proposition free from ambiguities due to the indexical elements - an eternal sentence (of the kind W. V. Quine talked about) or a Thought or *Gedanke* (of the Fregean kind) - such that it would have a value, truth or falsity - eternally fixed (...) We may assume that a proposition has an eternally fixed truth-value, but it is not absolutely clear to us what kind of a proposition that would be. For it remains open to us to discover some hidden, unsuspected determinants that would force us to withdraw our assent to it." ([12], p. 136)

[8]A judgment proceeds as a statement in which a semantic value is predicated of the sentence. Gokhale claims for this higher-order level of discourse: "A *syāt*-statement, in so far as it is a statement about a sense of a sentence, is a metalinguistic statement and not an object-linguistic one." ([7], p. 80).

claim[9]; and a third sort of judgment that Jains called by non-assertibility (*avaktavya*). Before discussing the meaning of this third predicate #, it follows from their combinations that the three basic statements are very similar to the set **3** = {T,F,#} and its eight combined subsets in $\wp(3)$ = **8** = {{T},{F},{#},{T,F},{T,#},{F,#},{T,F,#},∅}. The logical structure of **QAS** brings out the two main features of this sevenfold predication, where each component is to be rendered in terms of corresponding questions and answers.

DEFINITION 2. A Jaina predication expresses an ordered answer $\mathbf{A}(\alpha) = \langle \mathbf{a}_1(\alpha), \mathbf{a}_2(\alpha), \mathbf{a}_3(\alpha) \rangle$ to $n = 3$ basic questions $\mathbf{Q}(\alpha) = \langle \mathbf{q}_1(\alpha), \mathbf{q}_2(\alpha), \mathbf{q}_3(\alpha) \rangle$, such that $\mathbf{q}_1$: "Is α asserted?", $\mathbf{q}_2$: "Is α negated?", and $\mathbf{q}_3$: "Is α non-assertible?". There are $m = 2$ kinds of exclusive answers $\mathbf{a}_i(\alpha) \mapsto \{0, 1\}$ to each ordered question $\mathbf{q}_i$, where 0 is a denial "no" and 1 is an affirmation "yes". This yields the following list of $m^n = 2^3 = 8$ predications and their counterparts in a Belnap-typed set **8**:

(1) = ⟨1, 0, 0⟩ for {T} (2) = ⟨0, 1, 0⟩ for {F}
(3) = ⟨1, 1, 0⟩ for {{T},{F}} (4) = ⟨0, 0, 1⟩ for {#}
(5) = ⟨1, 0, 1⟩ for {{T},{#}} (6) = ⟨0, 1, 1⟩ for {{F},{#}}
(7) = ⟨1, 1, 1⟩ for {{T},{F},{#}} (8) = ⟨0, 0, 0⟩ for ∅

Each of the seven Jaina statements is an expression of single yes-answers ($\mathbf{a}_i = 1$) among three possible ones, while the remaining no-answers ($\mathbf{a}_i = 0$) are left silent by the affirmative nature of Jaina philosophy. The first two statements (1) and (2) mean that *every* standpoint is such that it makes a given sentence true or false, respectively. (3) means that there are standpoints for asserting the truth and the falsity of the sentence, while noting that a standpoint does not make this sentence both true and false at once. The *internal* consistency of the standpoints is stated in terms of *successive* assertion and denial. (4) is the troublesome statement that the sentence is non-assertible: although this semantic predicate seems to entail merely that a given sentence cannot be asserted (made true), this should leave place for strong denial (falsity-claim); but such a translation would collapse (4) into (2), all the more that this third *mūlabhaṅgi* is translated as a case of *simultaneous* assertion and denial. How can one and the same sentence be non-assertible and asserted at once? We return to this point in the next

[9]Jain "denial" corresponds to the relational negation of the realists (*paryudāsa pratiṣedha*), by contrast to the Mādhyamika non-relational negation (*prasajya pratisedha*). Accordingly, the "denial" of the second *mūlabhaṅgi* (2) amounts to an act of negative assertion or falsity-claim and stands for a commitment of the speaker about how the world is not, whereas every disciple of Mādhyamaka typically endorses an attitude of non-committment.

paragraph. The three remaining predications are combinations of the four preceding ones: (5) and (6) mean that there are standpoints that make the sentence true and non-assertible, or false and non-assertible. (7) is a combination of the three basic predications such that the available standpoints make the sentence true, false, and non-assertible.

The ultimate subset (8) doesn't appear in the list of the Jaina predications, however; hence the odd number of $8-1 = 7$ elements. A combinatorial account for this odd number of predications can be given as follows: there is an infinite number of particular arguments for any predication, and all of these are classified among a set of seven general standpoints in the *nayavāda*. Now since any two different kinds of standpoints may result in one and the same statement of the *syādvāda*, it follows from it that every sentence is made (or claimed to be) either true, false or non-assertible by a variable set of related standpoints. Therefore, there is always at least one standpoint $\mathbf{a}_i(\alpha) = 1$ for any sentence α. This entails that no sentence a can be an exception to these three basic judgments $\langle \mathbf{a}_1(\alpha), \mathbf{a}_2(\alpha), \mathbf{a}_3(\alpha)\rangle$, and the answer $\mathbf{A}(\alpha) = \langle 0, 0, 0\rangle$ is made an impossible case.

As rightly noted by Priest[10], no contemporary counterpart has been devised for the so-called "Jaina logic": the Jains have not defined any closed formal language with a set of constants (connectives) and a closed set of consequences. However, we can develop a plausible Jain logic within **QAS**.

DEFINITION 3. Jain logic is a model $\mathbf{J}_7 = \langle \mathfrak{M}, A\rangle$ upon a sentential language $\mathscr{L}$ and its set of logical connectives $\copyright = \{\sim, \wedge, \vee, \rightarrow\}$. It includes a logical matrix $\mathfrak{M} = \langle \mathbf{Q}; \mathbf{7}; D\rangle$, with:
- a function $\mathbf{Q}(\alpha) = \langle \mathbf{q}_1(\alpha), \mathbf{q}_2(\alpha), \mathbf{q}_3(\alpha)\rangle$;
- a set $\mathbf{7}$ of logical values;
- a subset of designated values $D \subseteq \mathbf{7}$.

The cardinality of D and the different matrices for $\copyright$ cannot be uniquely determined without solving an intermediary problem: the meaning of the "non-assertible" *avaktavya* in $\mathbf{q}_3$, by contrast to the two "assertible" *vaktavya* (*asti, nasti*) that constitute expressible predications in $\mathbf{q}_1$ and $\mathbf{q}_2$. Each ordered answer is a logical value from our many-valued perspective, and the meaning of the semantic predicate "non-assertible" is crucial to determine whether a positive answer to $\mathbf{q}_3(\alpha)$ results in a designated or non-

[10] "What are the semantic values of such compound sentences? Such a question is not one that Jaina logicians thought to ask themselves, as far as I know. So we are on our own here." ([15], p. 268).

designated value[11]. For if $\mathbf{A}(\alpha) = (4) = \langle 0, 0, 1\rangle$, then $\mathbf{a}_1(\alpha) = \mathbf{a}_2(\alpha) = 0$ and $\mathbf{a}_3(\alpha) = 1$. Assuming with Priest that a semantic value is designated if it expresses truth, then a non-assertible sentence should be asserted to be at least true in order to be designated. Is it so?

There are three main interpretations of *avaktavya*: (4.1) neither true nor false, (4.2) both true and false, (4.3) none (taking to be granted that not two of these can be accepted without extending the set of semantic predicates from $8-1$ to $16-1 = 15$ elements)[12]. Given the crucial role of the number 7, only one of these three possibilities is to be accepted as *the* third *mūlabhaṅgi*. The first interpretation is defended by [6], [7], and [9]; the second is urged by [4], [12], and [15]. [15] and [18] admit both interpretations, while the third interpretation is supported by [2] and [23].

Those who advocate (4.1) usually claim that the Jains always sustained internal consistency or non contradiction as an unquestionable meta-principle (*paribhāsā*); this amounts to reject any case of simultaneous assertion and denial from the same standpoint. Ganeri advanced in [6] a *reductio* argument against the inconsistent interpretation, to the effect that admitting a simultaneous assertion and denial would reduce the logical values (5) and (6) to (4). This collapsing argument is rejected in [18]], insofar as it omits to take the difference between the standpoints $\mathbf{a}_1$ and $\mathbf{a_2}$ into account[13].

[11] An alternative way consists in characterizing logical consequence in terms of an ordering relation between the elements of V, such that $p \models_{\mathbf{J}_7} q$ if and only if $\mathbf{A}(p) \leq \mathbf{A}(q)$. See [3],[22] about this process. An algebraic presentation for Jain logic is also given in [20],[22] and results in a bi-and- a-half-lattice (a product of two Belnap's bi-lattices) with no lower bound $\varnothing$ ($\langle 0, 0, 0\rangle$, in $\mathbf{J}_7$). But given that nothing seems to justify a specific hierarchy between the seven logical values, we stick to the view of logical consequence as preserving the designated value.

[12] Priest mentions the possibility of four-valued facets or *mūlabhaṅgi* and a subsequent 15-valued logic in [15], in such a way that a sentence could be said to be either asserted or denied, or both, or neither. Some other extensions of the basic predications have been entertained in [2] for Jain logic, assuming it to be a positive counterpart of the *catuṣkoṭi*; these yield an extension from 4- to 8- and 12-valued logics, where a given standpoint is "more asserted (or not)" than another. But such a probabilistic extension misleadingly takes the doctrine of relative truth for a logic of partial truth-values. Gokhale argues against this reading, because "*nayavāda*, as has generally been held, gives us a class of 'partial truths', whereas *syādvāda* gives us a class of whole truths (or the whole truth)." ([7], p. 74). In other words, each sentence is plainly true (or not) from each given standpoint.

[13] Ganeri's argument (see [6], p. 272) proceeds as follows: if *avaktavyam* means (4.2): {T,F}, then the fifth and sixth predicates yield (5.2): {T,{T,F}} and (6.2): {F,{T,F}}, respectively; now (5.2) and (6.2) are "logically equivalent" with {T,F}, given the trivially twofold occurrence of T and F. Hence the adoption of (4.2) entails that (5) and (6) conflate into (4), and the *seven*fold predication is done. Ganeri's mistake is due to his set-theoretical equation between sets and subsets of elements in V: "this argument seems to rely upon a

As a further argument for (4.3), Tripathi claimed that the incomplete interpretation (4.1) cannot square with the *affirmative basis* of the Jaina predications[14]. The latter means that any sentence can be made true from at least one standpoint, so that no sentence can be said to be neither true nor false. Assuming that "affirmative basis" essentially refers to an act of assertion (the second predication is a negative assertion), this implies that every Jaina predication asserts something about a sentence and cannot amount to a pure denial without assertive counterpart[15].

Conversely, Priest quotes some sources in support of (4.2) and takes them to mean a plausible admission of internal inconsistency[16]. The present paper does not purport to have the final word, but to note two main properties of $\mathbf{J}_7$ that are established in [18][17]. On the one hand, the essential occur-

conflation of two distinct standpoints: to state that p is asserted from one standpoint and both asserted and denied from another standpoint doesn't entail that p is merely asserted and denied, unless the crucial *syād* is suddenly removed from the meaning of a statement. But it could not be so, and Ganeri unduly commits the following simplification: $p \wedge (p \wedge \sim p) = (p \wedge \sim p)$." ([18], pp. 63-4)

[14]"To say that a thing neither exists (*asti*) nor does not exist (*nāsti*) is sheer skepticism, and the Jaina would never accept it as a *bhaṅga* (predicate), and as one of the *mūlabhaṅgas* (primary predicates) at that. (...) What is worse, the interpretation of the *avaktavya* as "neither" would make it indistinguishable from the fourth *koṭi* (alternative viewpoint) of the Mādhyamika *catuṣkoṭi*, as also from the *anirvanacanīya* (indescribable as either being or not-being) of the Vedānta." ([21], pp. 187-8). The argument is unconvincing, however, given that the Mādhyamikas deny the "neither ... nor"- position and don't affirm it (see Section 5); no confusion should arise from (4.1), accordingly.

[15]It could be objected to the view of a pure denial that any first-order denial implicitly contains a second-order assertion. Such an objection suggests that (4.3) includes a second-order affirmative basis (something like "arguably, I assert that I don't assert anything about p"); see Section 6 about this.

[16]Priest adduces his usual argument for dialetheism, according to which some (but not every) contradictions are true: "What should seem to be meant by two things being contradictory here is that they cannot obtain together. If [(4)] is *both true and false*, then [p] and [$\sim p$] are precisely not contradictories in *this* sense." ([14], pp. 271-2). Does this mean that a difference should be made between possibly true and impossibly true contradictions? A plea for possibly true contradictions has been made in [16], arguing that (4.1) could mean that some standpoint affords an evidence both for and against the truth of p. But the latter explanation does not seem to match with the definite value of a sentence in each standpoint, according to Gokhale (see note 12 above). This is why the third interpretation (4.3) will be favored in the following.

[17]A quantified epistemic interpretation of the standpoints has been suggested in [17]: each standpoint stands for a single belief within a community of agents, so that each Jain statement about α is translated as $\exists x B_x(\alpha)$ and reminds us of Jaśkowski's discussive logic $\mathbf{D}_2$. Such a translation helps to explain the *paraconsistent* behavior of the Jains: a set of inconsistent standpoints does not entail the truth of everything. Nevertheless, it doesn't account for Jain *realism* (see note 6 above): a standpoint is not the mere epistemic expression

rence of standpoints gives rise to a *quasi-value-functional* set of logical matrices for $\mathbf{J}_7$ where the logical value of a complex sentence is partly determined by the value of its components[18]. On the other hand, the incomplete or inconsistent interpretation of *avaktavya* makes $\mathbf{J}_7$ quasi-equivalent to two famous many-valued systems: Kleene's 3-valued logic $\mathbf{K}_3$ or Priest's 3-valued Logic of Paradox **LP**, respectively. This can be stated by the two following theorems:

THEOREM 1. $\mathbf{J}_7$ is a *paranormal* logic that is either paraconsistent or paracomplete. That is: for some sentences α, β of $\mathscr{L}$, either $\alpha, \sim\alpha \nvDash \beta$ or $\nvDash \alpha$ does not entail $\vDash \sim\alpha$. $\mathbf{J}_7$ is paracomplete and quasi-equivalent with $\mathbf{K}_3$ if and only if (4) is interpreted incompletely, and $\mathbf{J}_7$ is paraconsistent is quasi-equivalent with Priest's 3-valued logic **LP** if and only if (4) is interpreted inconsistently.

THEOREM 2. The matrices for the connectives © of $\mathbf{J}_7$ are invariant, irrespective of the interpretation of (4). For every connective $\bullet \in$ ©, $\mathbf{A}(\alpha \bullet \beta)_{icm} = \mathbf{A}(\alpha \bullet \beta)_{ics}$ for every value of α and β including the incomplete (icm) or inconsistent (ics) reading of #.

Apart from these technical results, it remains that no definite interpretation of *avaktavya* occurs in the literature and thus leaves the Jaina set of logical consequences indeterminate. The next point is to see whether a meaningful interpretation can be given to the third interpretation (4.3): what can be meant by *avaktavya*, if it is neither "both asserted and denied" nor "neither asserted nor denied"? For even though such an alternative reading prevents Jaina logic from reducing to what Matilal called a mere "facile relativism"[19], a formal approach hardly makes obvious any statement beyond being either true, or false, or both true and false, or neither true nor false.

For one thing, Bahm takes it (in [2]) to mean something like an *incomplete* thought: a sentence is non-assertible whenever no property P can be

of a belief or opinion, but the genuinely ontological expression of a facet of reality.

[18]Quasi-truth-functionality is due to the relative truth of standpoints. Two any sentences α and ψ can be true from two different standpoints; but there may be no standpoint from which α and ψ should obtain at once, according to the existential translation of a standpoint in [18]: $v(\exists x \mathrm{B}_x(\alpha)) = \mathrm{T}$ and $v(\exists x \mathrm{B}_x(\psi)) = \mathrm{T}$ don't entail $v(\exists x \mathrm{B}_x(\alpha \wedge \psi)) = \mathrm{T}$, but $v(\exists x \mathrm{B}_x(\alpha \wedge \psi)) = \mathrm{T}$ or F. On the origins of quasi-truth-functionality, see [17].

[19]"It also amounts to a view which announces that all predicates are *relative* to a point of view; no predicates can be *absolutely* true of a thing of a thing or an object in the sense that it can be applied unconditionally at all times under any circumstances. Jainas in this way becomes identified with a sort of facile relativism." ([12], p. 133). Again, the crucial role of standpoints clearly points out that the Jain logic is not a real challenge to PNC.

completely predicated of S. But this is the essential feature of *anekāntavāda*, the partial truth for every standpoint of the Jaina *nayavāda*: the cornerstone of their pluralist metaphysics is that reality is an indefinite collection of incomplete perspectives. Assertion and denial are not categorical or one-sided speech-acts, therefore, and the essential incompleteness of any *syād* is likely to undermine Bahm's explanation.

A more insightful reading seems to emerge in [23], where non-assertibility is synonymous with *non-distinction*: a sentence is non-assertible whenever its object S cannot be said to be properly P or not P. The difference is thus made with the interpretation (4.2), in the sense that S is said to be both P and not-P by including both opposite properties from one contradictory standpoint. But again, Tripathi claims in [23] that the Jains fully subscribed to the law of non-contradiction and would have refused any *self*-contradictory statement[20]. A plausible account of being indistinguishable refers to the Hegelian view of an internal or inclusive contradiction without exclusive opposition between its terms. In support of this awkward view of contradiction, it is worthwhile to note that most of the Jaina or Mādhyamika sentences are about such metaphysical subjects as ātman, Brahman and their being existent. One may be hesitant about the logical form of an expression like "ātman is self-existent", where existence occurs as a predicate; but a more charitable reading would be to the effect that the subject-term S is elliptically said to exist or to be as falling under a certain property P. Consequently, *avaktavya* might mean that S is not any more P that non-P. But which sort of S could be so indistinguishable as not only to cover both P and all its complementary properties, but also to cancel any distinction between these properties? Tripathi mentions as a "non-expressible" sentence that which can be thought but cannot be expressed (for want of a distinguishable set of properties)[21]. Such a subject should be

[20]"No system of philosophy can afford to accept self-contradiction as valid, because if self-contradiction is accepted as valid without any qualifications, then there remains no weapon for criticism, anything which is said will have to be accepted, because even self-contradictories is valid. It is certain that the Jaina does not take leave of logic and consistency; he does criticize others by pointing out self-contradiction. Every system of philosophy has its contradictory which is regarded as false. This is why when a system has to accept a synthesis of contradictories as valid, it has to invent one device or another which at least seems to take off the edge from the contradictories." ([21], p. 188).

[21]Bahm's account must be distinguished from Meinong's famous example of a "round square", which has frequently been mentioned as a case of impossible object and a challenge to PNC. A round square is an object that can be expressed (described) but cannot be thought (imagined, or conceived mentally). To the contrary, the third interpretation of *avaktavya* refers to something that can be thought but cannot be expressed. Is there such a subject S that can fulfill this requirement? A Wittgensteinian reader would answer neg-

kept silent, according to the Wittgensteinian stance that the limits of language are the limits of thought. (But our former reference to Hegel should give rise to a non-Wittgensteinian relationship between language and the world.) While noting that Hegel's philosophy supported a transcendental idealism and clearly differs from the Jaina realism, a common point between Jainism and the Buddhist trend of Mādhyamikas seems to be their common rejection of *logical atomism*: reality is not a whole whose parts would be objects and their properties, or at least not for some extra-natural entities that transcend the empirical level of illusory data (*prātibhāsika*). This plausible account of (4.3) will be pursued in the next section, because it might make sense of Nāgārjuna's radical skepticism.

To conclude our discussion of Jaina logic, Priest uses in [15] an analogy with the cube to make sense of complete truth: every facet of reality is a side of a cube, and reality is the collection of every such facet. But Jaina cubism is such that the indefinite number of facets turns the cube into a polygon even more complex than Descartes' chiliagon. Just as Picasso wanted to catch a conceptual reality by pooling different perspectives of a character together in one and the same profile, the Jaina philosophy relies upon a plurality of standpoints to grasp the essence of reality. A logical translation of this view is given in [4]): plain truth amounts to a complete knowledge (*pramaña*) whose expression in a complete judgment consists in the addition of the seven sorts of predication. Is this a right way to describe the transition from partial to complete truth[22]?

An alternative account would be to state that a subject is completely described when absolutely every particular standpoint is listed, rather than just the seven kinds of argument from the *nayavāda*. Such an exhaustive completion is impossible, given the infinite sort of standpoints that consti-

atively to this question, assuming that "whereof one cannot speak, thereof one must be silent".

[22]The following definition of plain truth is given in [4]: "An object X can be viewed from any one of the seven standpoints. However, since the totality of all these seven possibilities comprises the *pramāṇa-saptabhaṅgī* (complete judgment of the phenomenal world in terms of seven possibilities), the disjunction, denoted by $\wedge$, of these seven predications should lead to a tautology." ([4], p. 186). In algebraic terms, the Jains would thus assimilate one-sided truth with logical tautology and define the latter as the union of the seven elements of V. That is: $\top = ((1) \cup (2) \cup (3) \cup (4) \cup (5) \cup (6) \cup (7))$. This definition of tautology clearly differs from that of Priest's in [15] or $\mathbf{J}_7$ in [18]: a sentence is a tautology if it is designated from every standpoint. But this is a definition of tautology in the *conventional* sense of truth, by contrast to the aforementioned absolute sense of truth that uniquely leads to a *pramāṇa*. One could wonder another thing, with respect to this definition of one-sided tautology: does it correspond to the union of the seven kinds of standpoints or, rather, should it collect the indefinitely many particular standpoints that are included in each of these seven kinds?

tute the proper description of any object.

A natural translation of (4) within $\mathbf{J}_7$ might be taken to be the twofold answer "yes and no" to the third basic question: $\mathbf{a}_3(\alpha) = \{1, 0\}$. But it is not so, given that this third question is positively answered if the corresponding sentence is inexpressible. No yes-no answer occurs in the Jaina question-answer game, consequently: two different questions can result in the same answer or not, but no single question can be answered oppositely by "yes" and "no" at once[23]. This is the gist of self-contradiction, and even the third basic predicate of inexpressibility does not state it because non-distinction does not mean an internal coexistence of opposite properties. These cannot coexist, by definition.

Whatever the final word may be about (4), we argue two things about complete truth: it does not mean for a given sentence either to be assigned a designated value (this is partial truth) or to be uniquely asserted and, therefore, be given the logical value (1) in $\mathbf{J}_7$[24]; partial truth is a sufficient condition of truth-assignment for the Jains, while the skeptic Mādhyamikas take complete truth to be a necessary condition for truth-assignment. Let us now consider this skeptic logic within a question-answer game of **QAS**.

5 Nāgārjuna's Principle of Four-Fold Negation

Nāgārjuna's radical skepticism is summarized in his *Mūlamadhyamaka-kārikā*, where the first verse includes four sentences (or lemmas) that are

[23]Three levels of inconsistency can be graded within the framework of **QAS**: *light* inconsistency, or inconsistency from two different standpoints: {{T},{F}}, i.e. $\mathbf{a}_i(\alpha) = \mathbf{a}_j(\sim\alpha) = 1$ (where $i \neq j$); *mild* inconsistency, or inconsistency from one and the same standpoint: {{T,F}}, i.e. $\mathbf{a}_i(\alpha) = \mathbf{a}_i(\sim\alpha) = 1$; and *strong* inconsistency, or inconsistency in one and the same answer: {{T,~T}}, i.e. $\mathbf{a}_i(\alpha) = \mathbf{a}_i(\alpha) = \{1,0\}$. The Jain *anekāntavāda* embodies a logic of light inconsistency; Priest's Logic of Paradox **LP** argues for a mild inconsistency that corresponds to the inconsistent interpretation (4.1) of *avaktavyam*; but no counterpart seems to occur for the strong inconsistency of self-contradiction, going beyond the so-called "impossible" values of [20]. Indeed, strong inconsistency consists of non-empty subsets including an element and its complement. Such a case is impossible even in a combinatorial approach of semantic values, insofar as Priest's value {T,F} assumes that T and F are not complementary to each other.

[24]Returning to the comparison with Jaśkowski's Discussive logic $\mathbf{D}_2$, the Polish logician rendered each standpoint by the modality of possibility, $\Diamond$. Accordingly, any sentence α that is uniquely asserted (such that $v(\alpha) = (1)$) is logically necessary because it is cannot be but asserted, and it is not possible for it to be denied or taken to be non-assertible. Thus $v(\alpha) = (1)$ means the same as α. This modal interpretation squares with the idea of one-sidedness; however, the Jain view of *pramāṇa* still goes beyond such a logical necessity (see note 22 above).

equally denied by means of stances (*dṛṣṭis*, or *koṭi*) and result in the the so-called Principle of Four-Cornered Negation (thereafter: 4CN) or Tetralemma (*catuṣkoṭi*). Thus:

(a) Does a thing or being come out itself? No.
(b) Does a thing or being come out the other? No.
(c) Does it come out of both itself and the other? No.
(d) Does it come out of neither? No.

How can Nāgārjuna consistently deny all the four questions at once? While noting that their content refers to the Mādhyamika's doctrine of emptiness (*sūnyavāda*), a problem arises about the meaning of negation in the four aforementioned answers. A tentative formalization of (a)-(d) yields the following, where a is a predication of the form "S is P" (with S for "thing" and P for "coming out iself") and $\sim$ is classical negation:

(a′) Not (S is P) = $\sim(\alpha)$
(b′) Not (S is not P) = $\sim(\sim\alpha)$
(c′) Not (S is P and S is not P) = $\sim(\alpha \wedge \sim\alpha)$
(d′) Not (neither S is P nor S is not P) = $\sim(\sim(\alpha \vee \sim\alpha))$

Assuming that negation is the relational *paryudāsa pratiṣedha*, the set of four negative statements is clearly inconsistent: (b′) is equivalent with the affirmation α (by double negation), and this is patently contradictory with its negation in (a′). Even more than for the Jains, it is commonly acknowledged that the Mādhyamikas unexceptionably subscribed to PNC and cannot then accept both (a′) and (b′). Furthermore, (d′) occurs as a denial of the denial of the Principle of Excluded Middle (PEM), according to which every sentence or its negation is true. But it clearly appears that the double denial arising in (d′) does not amount to an affirmation of PEM, since (a′) and (b′) already reject the affirmation of both α and $\sim\alpha$.

A way to avoid the contradiction (a′)-(b′) has been urged by Horn (in [10]), who claimed that the negation of every sentential content should be rendered as a *predicate-term* negation rather than a *predicate* negation[25].

[25]Horn claims that "crucially, no distinction between contradictory and contrary negation was regularly made within classical Indian logic." ([10], p. 80) However, the contrary or contradictory feature of a negation crucially depends upon the nature of the subject in a predication: are the subjects of a Jain predication sometimes universal, sometimes particular? No definite answer seems to be available to disentangle the meaning of 4CN; it is only the later school of Navya-Nyāya that will deal with such equivocation cases. See in this respect J. Ganeri: "Towards a formal regimentation of the Navya-Nyāya technical language" (parts I,II), in *Logic, Navya-Nyāya and Applications* (Homage to Bimal Krishna Matilal), M.K. Chakraborti and Löwe, B. and Mitra M.N. and Sarukkai S (eds.), College Publications, London, 2008, pp. 105-121.

The distinction between predicate-term and predicate negation cannot be expressed in a modern or Fregean logic, where predicate-terms and predicates are collapsed into a unique function. By using term logic, (b′) should be read as "S is not-P", the contrary opposite of (a′). The conjunction (a′)-(b′) results in a stronger relation of incompatibles, and Horn is right to say that two contraries can be consistently negated without entailing any self-contradiction. In this respect, an application of *intuitionistic* negation ($\neg\alpha$ for "S is not-P") should fill the bill and be preferred to the classical negation ($\sim\alpha$ for "S is not P"): $\sim(\sim\alpha)$ becomes $\sim(\neg\alpha)$, and the latter cannot be reduced to a by the law of double negation.

Does this mean that intuitionistic logic should be seen as a proper logic for 4CN? It is not, given that the last statement (d′) leads to another contradiction. For since one of de Morgan's laws states that $(\sim\alpha \wedge \sim(\neg\alpha))$ is equivalent to $\sim(\alpha \vee \neg\alpha))$, how to claim with (a′)-(b′) that S is neither P nor not-P: $\sim(\alpha \vee \neg\alpha))$ while denying it at the same time with (d′): $\sim(\sim(\alpha \vee \neg\alpha)) \leftrightarrow (\alpha \vee \neg\alpha)$?

The whole result turns 4CN into a case for radical skepticism: not only does the speaker Nāgārjuna ignore whether S is P or not, but he goes on denying that he does ignore it. This troublesome stance has been noted by Raju[26] and accounts for the difference between Buddhism and nihilism, as currently urged by a number of commentators: nihilism is the affirmation that nothing is real or can be known to be so; whereas Buddhism argues for a mere denial without any positive counterpart. The positive basis of each Jaina statement included a case of negative assertion, as witnessed by the predication (2); but no such assertion arises in 4CN, where negation is pure denial. Before answering to whether there can be a negation without any positive counterpart, we suspect the core difficulty with 4CN to lie in the meaning of its wide scope negation (the answer "No"): it is used to produce a denial, and this no-answer should find a proper treatment within the formal framework of **QAS**.

Unlike the Jaina statements, and following the connection established between Mādhyamika skepticism and anti-realism, we assume that each *koṭi* deals with the impossibility of knowledge: the human failure to catch any absolute truth (*paramārthasatya*) about reality is a sufficient reason

[26]The alleged founder of 4CN, Sañjaya (≃ 6th century B.C.), would have influenced the Greek philosopher Pyrrho in his radical skepticism; Raju states this point by claiming that Pyrrho "maintained that 'I am not only not certain of the knowledge of any object, but also not certain that I am not certain of such a knowledge' " ([16], p. 695). It is worthwhile to note that the Greek principle of indifference *ou mallon* (not any more than) strikingly parallels 4CN.

to deny any justifiable belief and thus any truth-assignment, according to Nāgārjuna's *sūnyavāda*. If so, we introduce a four-valued logic of acceptance and rejection for 4CN.

DEFINITION 4. A logic of acceptance and rejection is a model $\mathbf{AR}_4 = \langle \mathfrak{M}, \mathbf{A} \rangle$ upon a sentential language $\mathscr{L}$ and its set of logical connectives $\copyright = \{\sim, \wedge, \vee, \rightarrow\}$. It includes a logical matrix $\mathfrak{M} = \langle \mathbf{Q}; 4; D \rangle$, with :
- a function $\mathbf{Q}(\alpha) = \langle \mathbf{q}_1(\alpha), \mathbf{q}_2(\alpha) \rangle$;
- a set **4** of logical values;
- a subset of designated values $D \subseteq \mathbf{4}$, where $D = \{\langle 1, 0 \rangle, \langle 1, 1 \rangle\}$.

$\mathbf{Q}_(\alpha)$ is an ordered set of $n = 2$ questions about the sentence α, with $\mathbf{q}_1$: "is a justifiably be true?" and $\mathbf{q}_2$: "is a justifiably false?"[27], and $n = 2$ sorts of answers such that $\mathbf{a}(\alpha) \mapsto \{0, 1\}$. It results in a set V of $m^n = 2^2 = 4$ logical values, each standing for an explicit belief-attitude in $\mathbf{4} = \{\langle 1, 0 \rangle, \langle 1, 1 \rangle, \langle 0, 0 \rangle, \langle 0, 1 \rangle\}$. The difference with $\mathbf{J}_7$ is that no third question $\mathbf{q}_3$ occurs here: *avaktavya* is not a Mādhyamika concept, so that only two basic semantic predicates or *muladṛṣṭis* are required in 4CN. At the same time, $\mathbf{AR}_4$ is a general logic of statements that could include the Jaina stances as well: the Jaina value $\langle \mathbf{a}_1(\alpha), \mathbf{a}_2(\alpha), \mathbf{a}_3(\alpha) \rangle$ can be equated with the value $\mathbf{Q}(\alpha) = \langle \mathbf{q_1}(\alpha), \mathbf{q_2}(\alpha)$ of $\mathbf{AR}_4$ by canceling the third *bhañga* $\mathbf{a}_3(\alpha)$. Then $\langle 1, 0 \rangle = \{\langle 1, 0, 1 \rangle, \langle 1, 0, 0 \rangle\}$, $\langle 1, 1 \rangle = \{\langle 1, 1, 1 \rangle, \langle 1, 1, 0 \rangle\}$, and $\langle 0, 1 \rangle = \{\langle 0, 1, 1 \rangle, \langle 0, 1, 0 \rangle\}$. A relevant exception concerns the third value $\langle 0, 0 \rangle = \{\langle 0, 0, 1 \rangle, \langle 0, 0, 0 \rangle\}$, which includes the eighth forbidden value $\langle 0, 0, 0 \rangle$ in $\mathbf{J}_7$. This forbidden value is our key to a better understanding of Nāgārjuna's four stances, with the following definition of negation and its distinction with the speech-act of denial.

DEFINITION 5. For every sentence α such that $\mathbf{A}(\alpha) = \langle \mathbf{a}_1(\alpha), \mathbf{a}_2(\alpha) \rangle$:
$\mathbf{A}(\sim\alpha) = \langle \mathbf{a}_2(\alpha), \mathbf{a}_1(\alpha) \rangle$.

The import of **QAS** is to bring an algebraic distinction between logical negation and denial: contrary to the usual perplexing presentation of 4CN, denial should not be rendered as a connective that is part of the sentential content α; rather, a denial is a no-answer that does not stand for a function but its resulting value. Correspondingly, a proper formalization of 4CN is suggested in the following style:

(a″) $\mathbf{a}_1(\alpha) = 0$
(b″) $\mathbf{a}_1(\sim\alpha) = 0$

[27]The second question "Is α justifiably false?" is equivalent with "Is $\sim\alpha$ justifiably true?". This results in the following equation for negation in $\mathbf{AR}_4$: $\mathbf{a}_1(\sim\alpha) = \mathbf{a}_2(\alpha)$, and conversely.

(c″) $\mathbf{a}_1(\alpha \wedge {\sim}\alpha) = 0$
(d″) $\mathbf{a}_1({\sim}(\alpha \vee {\sim}\alpha)) = 0$

Only one valuation of $\mathbf{AR}_4$ accounts for the consistency of (a″)-(d″), namely: $\mathbf{A}(\alpha) = \langle 0, 0\rangle$, the forbidden value of Jaina logic. Following the definition of conjunction and disjunction in $\mathbf{AR}_4$[28], (a″) and (b″) entail that $\mathbf{a}_2(\alpha) = \mathbf{a}_1(\alpha\wedge {\sim}\alpha) = \mathbf{a}_1({\sim}(\alpha\vee {\sim}\alpha)) = \mathbf{a}_2(\alpha \vee {\sim}\alpha) = 0$.

Once again, the usual perplexity caused by Nāgārjuna's stance is due to a confusion between the relational and non-relational reading of negation. The former negation (*paryudāsa pratiṣedha*) is not an answer about whether the sentence α is true or false, given that it occurs within its sentential content in the whole expression ${\sim}\alpha$; most importantly, it assumes bivalence and entails that ${\sim}\alpha$ is false whenever α is true (and conversely). Therefore, no sentence can be given a "gappy" value (neither true nor false) with such a relational use of negation. Furthermore, introducing the intuitionistic negation $\neg$ for this purpose is not the solution either: that α is said to be neither true nor false cannot explain again why this gappy solution is insufficient to account for the fourth stance (d'). This leads to the conclusion that Nāgārjuna's denial should be strictly distinguished from assertive negation and be equated with the "absolutely no"-answer $\langle 0, 0\rangle$.

Our point about logical values actually holds for every negation, in the sense that there is no functional difference between classical and intuitionistic negation $\mathbf{AR}_4$. For the difference between the two negations does not lie in the definition of their mapping from $\mathscr{L}$ to V but, rather, in the domain of values they range over. Given that classical negation assumes a one-one correspondence theory of truth, this entails that a sentence cannot be said to be either both true and false or neither true nor false; hence a restriction of the range from $V = \mathbf{4}$ to $V = \mathbf{2} = \{\langle 1, 0\rangle, \langle 0, 1\rangle\}$. As to the intuitionistic theory of truth as justifiable truth, no sentence can be said to be true unless the justification is definite and this stringent view of justification implies another restriction from $V = 4$ to $V = 3 = \{\langle 1, 0\rangle, \langle 0, 0\rangle, \langle 0, 1\rangle\}$. The Jaina case embodies a paraconsistent variant, where a sentence can be said to be both true and false but excludes the possibility that it be none; hence a corresponding restriction from $V = 4$ to $V = 3 = \{\langle 1, 0\rangle, \langle 1, 1\rangle, \langle 0, 1\rangle\}$. The relative truth of *nayavāda* also accounted for the combination of such basic

[28] A complete description of the semantics for $\mathbf{AR}_4$ is not required in the context of 4CN, but it includes maximal and minimal functions (*max*,*min*) upon the values of V, given a total ordering function $<$ between these elements proceeds as follows: $\langle 0, 1\rangle < \langle 0, 0\rangle < \langle 1, 1\rangle < \langle 1, 0\rangle$. Hence the following definition of the connectives of conjunction and disjunction: $v(\alpha \wedge \psi) = min(\alpha, \psi)$, and $v(\alpha \vee \psi) = max(\alpha, \psi)$.

answers into new logical values in $\mathbf{J}_7$, unlike the non-relative, absolute or one-sided view of truth in the Mādhyamika school.

But that is not the whole story of 4CN. Recalling a former quotation by Raju, two problems remain to be solved. Firstly: does Nāgārjuna deny absolutely everything, including his own denials? And secondly: is the *catuṣkoṭi* a mere reversal of the *saptabhaṅgī*, i.e. the transformation of a common set of positive statements into negative statements?

6 Two contrary logics?

Let us note about the first question that a distinction can be made between two generic forms of skepticism, a moderate and a radical one. The former is closer to what the Buddhists meant by nihilism and wanted to be strictly distinguished from; it means that nothing can be known about reality, but one least thing to be known is precisely that nothing mundane can be known. In contrast to this, the radical version goes on denying any denial about our knowledge about reality: ignorance is not asserted but doubted itself. Whether or not such a distinction relates to the Greek schools of the New Academy (Arcesilas, Carneades) and Pyrrhonism (Pyrrho, Timon of Phlius) does not really matter in what follows. Rather, the point is whether Nāgārjuna endorsed radical skepticism and what his rejection consisted in. In the light of **QAS**, the complete denial of 4CN means that only no-answers are given to preceding questions.

As to the second question, Bahm replies in [2] that the two Indian logics cannot merely seen as mutual contraries: Jaina logic cannot be reduced to a Principle of Four-Cornered Affirmation. **QAS** already brought this point out by the cardinality of the sets of logical values, given the essential occurrence of a third question (about *avaktavya*) in $\mathbf{J}_7$. Nevertheless, there is a reason to claim that these philosophical schools are really opposite to each other in some respect. The *catuṣkoṭi* can be taken to be a reversal of *saptabhaṅgī* only if the sentential content of a denial or an affirmation is of the first order, i.e. stands for a declarative sentence about reality; but the same cannot be safely said for higher-order questions about the answerer's attitudes[29].

[29]The order of attitudes and their statements can be reformulated in terms of iterated modalities: the statement "α" is an affirmation and correlated belief about α, $\mathrm{B}(\alpha)$; the statement "I affirm that α" is an affirmation and correlated belief about the affirmation and correlated belief about α, $\mathrm{B}(\mathrm{B}\alpha)$; and so on for any n-ordered statement as a sequence of n beliefs: $\mathrm{B}^n(\alpha)$. The difference between $\mathbf{AR}_4$ and modal logic is that iterated attitudes are not rendered as modal operators but as logical values in the former semantics. See note 31

Let us exemplify this symmetrical behavior by means of two Socratic dialogues, where an initial question about the atomic sentence p is accompanied with a sequence of oratory questions (the questioner expects to have a given answer) and answers. The answerer to a common questioner (the doctrinalist Aristotle) is a Jaina speaker (Vādiveda Sūri) and a Mādhyamika speaker (Nāgārjuna), respectively. It clearly appears that the resulting dialogues are radically opposed to each other, and we bring this out by formalizing them in terms of **QAS**.

DIALOGUE 1: ARISTOTLE VS. VĀDIVEDA SŪRI

1. **Q**: Do you accept p?
   $[\mathbf{a}_1(p) = 1?]$
2. **A**: Yes, I accept p.
   $[\mathbf{a}_1(p) = 1]$
3. **Q**: Therefore you reject $\sim p$?
   $[\mathbf{a}_2(p) = 0\ ?]$
4. **A**: No, I do not reject $\sim p$.
   $[\mathbf{a}_2(p) \neq 0]$
5. **Q**: Does it mean that you also accept $\sim p$?
   $[\mathbf{a}_2(p) = 1\ ?]$
6. **A**: Yes, I also accept $\sim p$.
   $[\mathbf{a}_2(p) = 1]$
7. **Q**: Therefore you accept p and $\sim p$?
   $[\mathbf{a}_1(p \wedge \sim p) = 1\ ?]$
8. **A**: Yes, I accept both.
   $[\mathbf{a}_1(p \wedge \sim p) = 1]$
9. **Q**: Therefore you reject $\sim(p \wedge \sim p)$?
   $[\mathbf{a}_2(p \wedge \sim p) = 0\ ?]$

---

below.

10. **A**: No, I don't reject $\sim(p \wedge \sim p)$.
    $[\mathbf{a}_2(p \wedge \sim p) \neq 0]$
11. **Q**: Does it mean that you also accept $\sim(p \wedge \sim p)$?
    $[\mathbf{a}_2(p \wedge \sim p) = 1\ ?]$
12. **A**: Yes, I also accept $\sim(p \wedge \sim p)$.
    $[\mathbf{a}_2(p \wedge \sim p) = 1]$
13. **Q**: Therefore you reject $\sim((p \wedge \sim p) \wedge \sim(p \wedge \sim p))$?
    $[\mathbf{a}_2(((p \wedge \sim p) \wedge \sim(p \wedge \sim p))) = 0\ ?]$
14. **A**: No, I don't reject $\sim((p \wedge \sim p) \wedge \sim(p \wedge \sim p))$.
    $[\mathbf{a}_2(((p \wedge \sim p) \wedge \sim(p \wedge \sim p))) \neq 0]$
15. **Q**: Therefore you also accept $\sim((p \wedge \sim p) \wedge \sim(p \wedge \sim p))$?
    $[\mathbf{a}_1(\sim((p \wedge \sim p) \wedge \sim(p \wedge \sim p))) = 1?]$
16. **A**: Yes, I also accept $\sim((p \wedge \sim p) \wedge \sim(p \wedge \sim p))$
    $[\mathbf{a}_1(\sim((p \wedge \sim p) \wedge \sim(p \wedge \sim p))) = 1]$
    ...

It emerges from this abortive maieutic that the doctrinalist questioner fails to make the answerer his own reason: the whole answers are perfectly rational albeit inconsistent, in the light of $\mathbf{AR}_4$ and its non-classical logical values that are exclusively positive or negative[30].

THEOREM 3. For every sentence α (including p, $\sim p$, $p \wedge \sim p$, $\sim(p \wedge \sim p)$, and so on), the answer of the Jaina in $\mathbf{AR}_4$ is $\mathbf{A}(\alpha) = \langle 1, 1 \rangle$.

Proof: Let us assume that $\mathbf{a}_1(p \wedge \sim p) = 1$; then $\mathbf{a}_1(p) = \mathbf{a}_1(\sim p) = \mathbf{a}_2(p) =$

[30]The semantics for $\mathbf{AR}_4$ can be said to be bivalent in this respect: for every answer given to question $\mathbf{q}_i$ about the sentence α, the corresponding answer is either positive ($\mathbf{a}_i(\alpha) = 1$) or negative ($\mathbf{a}_i(\alpha) = 0$). *Tertium non datur*. Concerning any positive and negative answer to one and the same question, it has been argued earlier (see note 23) that it is equally impossible in the pluralist approach of the Jains. Hence the ensuing difference between two grades of inconsistency in $\mathbf{AR}_4$: a given answer $\mathbf{A}(\alpha)$ is externally *inconsistent* if and only if $\mathbf{a}_1(\alpha) \neq \mathbf{a}_2(\alpha)$; it is internally inconsistent or *incoherent* if and only if, for any answer x in {0,1}, $\mathbf{a}_i(\alpha) = x$ and $\mathbf{a}_i(\alpha) \neq x$. Accordingly, there is a crucial difference between sentential inconsistency and non-sentential inconsistency (incoherence): two sentences α and $\sim\alpha$ can be mutually inconsistent while the answers $\mathbf{A}(\alpha)$ and $\mathbf{A}(\sim\alpha)$ about them are internally consistent (coherent).

1. And if $\mathbf{a}_1(\sim(p \wedge \sim p)) = 1$ then $\mathbf{a}_2(p \wedge \sim p) = 1$, i.e. $\mathbf{a}_2(p) = 1$ or $\mathbf{a}_1(\sim p) = 1$. Hence for every α, $\mathbf{a}_1(\alpha) = \mathbf{a}_2(\alpha) = 1$. Hence $\mathbf{A}(\alpha) = \langle \mathbf{a}_1(\alpha), \mathbf{a}_2(\alpha) \rangle = \langle 1, 1 \rangle$.

Let us now apply the same process to a *dual* dialogue between the dogmatist questioner Aristotle and his skeptic answerer. This yields the exact reversal of the preceding dialogue, given that each question about whether a given sentence is *accepted* becomes a question about whether it is *rejected*.

DIALOGUE 2: ARISTOTLE VS. NĀGĀRJUNA

1. **Q**: Do you reject p?
   [$\mathbf{a}_1(p) = 0$?]
2. **A**: Yes, I reject p.
   [$\mathbf{a}_1(p) = 0$]
3. **Q**: Therefore you accept $\sim p$?
   [$\mathbf{a}_2(p) = 1$?]
4. **A**: No, I do not accept $\sim p$.
   [$\mathbf{a}_2(p) \neq 1$]
5. **Q**: Does it mean that you also reject $\sim p$?
   [$\mathbf{a}_2(p) = 0$?]
6. **A**: Yes, I also reject $\sim p$.
   [$\mathbf{a}_2(p) = 0$]
7. **Q**: Does it mean that you reject both p and $\sim p$?
   [$\mathbf{a}_1(p \vee \sim p) = 0$?]
8. Yes, I reject both p and $\sim p$.
   [$\mathbf{a}_1(p \vee \sim p) = 0$]
9. **Q**: Therefore you accept $\sim(p \vee \sim p)$?
   [$\mathbf{a}_2(p \vee \sim p) = 1$?]
10. **A**: No, I do not accept $\sim(p \vee \sim p)$.
    [$\mathbf{a}_2(p \vee \sim p) \neq 1$]

11. Does it mean that you reject both $(p \vee \sim p)$ and $\sim(p \vee \sim p)$?
    $[\mathbf{a}_1(((p \vee \sim p) \vee \sim(p \vee \sim p))) = 0\ ?]$
12. **A**: Yes, I reject both $(p \vee \sim p)$ and $\sim(p \vee \sim p)$.
    $[\mathbf{a}_1(((p \vee \sim p) \vee \sim(p \vee \sim p))) = 0]$
13. **Q**: Therefore you accept $\sim((p \vee \sim p) \vee \sim(p \vee \sim p))$?
    $[\mathbf{a}_2(((p \vee \sim p) \vee \sim(p \vee \sim p))) = 1?]$
14. **A**: No, I don't accept $\sim((p \vee \sim p) \vee \sim(p \vee \sim p))$.
    $[\mathbf{a}_2(((p \vee \sim p) \vee \sim(p \vee \sim p))) \neq 1]$
15. **Q**: Therefore you also reject $\sim((p \vee \sim p) \vee \sim(p \vee \sim p))$?
    $[\mathbf{a}_2(((p \vee \sim p) \vee \sim(p \vee \sim p))) = 0?]$
16. **A**: Yes, I also reject $\sim((p \vee \sim p) \vee \sim(p \vee \sim p))$
    $[\mathbf{a}_2(((p \vee \sim p) \vee \sim(p \vee \sim p))) = 0]$
    ...

Again, the doctrinalist questioner failed to make the answerer his reason: the whole is rational albeit incomplete, so long as the answerer refuses to commit in the truth of any sentence.

THEOREM 4. For every sentence α (including p, $\sim p$, $p \vee \sim p$, $\sim(p \vee \sim p)$, and so on), the answer of the Mādhyamika in $\mathbf{AR}_4$ is $\mathbf{A}(\alpha) = \langle 0, 0\rangle$.
Proof: if $\mathbf{a}_1(p \vee \sim p) = 0$ then $\mathbf{a}_1(p) = \mathbf{a}_1(\sim p) = \mathbf{a}_2(p) = 0$. And if $\mathbf{a}_1(\sim(p \vee \sim p)) = 0$ then $\mathbf{a}_2(p \vee \sim p) = 0$, i.e. $\mathbf{a}_2(p) = 0$ or $\mathbf{a}_1(\sim p) = 0$. Hence for every α, $\mathbf{a}_1(\alpha) = \mathbf{a}_2(\alpha) = 0$. Hence $\mathbf{A}(\alpha) = \langle \mathbf{a}_1(\alpha), \mathbf{a}_2(\alpha)\rangle = \langle 0, 0\rangle$.

Just as the Jains refuse exclusive acts of positive assertion and contend themselves with inconsistent affirmations, the Mādhyamikas refuse exclusive acts of negative assertion and contend themselves with incomplete denials.

A parallel can be made here with da Costa paraconsistent logics C_1-C_n: these are non-truth-functional systems where contradictions are variably affirmed or denied according to the structural complexity of the contradictory sentences (p and $\sim p$, in C_0; $(p \wedge \sim p)$ and $\sim(p \wedge \sim p)$, in C_1; and so on). By the same way, a set of dual paraconsistent logics C_1'-C_n' can be devised for the dialectical process of 4CN and states that alternatives are variably affirmed or denied according to the structural complexity of the alternative sentences: (p or $\sim p$, in C_0'; $(p \vee \sim p)$ or $\sim(p \vee \sim p)$, in C_1'; and so on).

But the parallel stops here, because the preceding dialogues have shown that the structural complexity of a sentence does not change the attitude of the answerer. In this respect, the Jains and Mādhyamikas are likely to be considered as two contrary attitudes or judgments in the common logic of statements $\mathbf{AR}_4$: the former affirm everything whereas the latter deny everything.

Returning to a preceding objection, it remains to consider to what extent such radical speakers can be said to affirm "everything" (doxastic eclecticism) or deny "everything" (doxastic nihilism) in their dialectical games[31]. While the concerned texts mention dialectical games about first-order statements only, it hardly makes sense to contend that Nāgārjuna would have denied his own denials with respect to first-order statements.

Let us make a semantic ascent and consider the second-order statement α': "I don't affirm that α (is true)". A no-answer to the question $\mathbf{q}_1(\alpha')$: "is α' justifiably true?" would mean that the answerer denies to have denied (the truth of) α, while a yes-answer would entail that he affirms to have denied α (as he did). The same objection can be made to a universally affirmative stance in the Jains. Likewise, the Jain would hardly give an affirmative answer to α' without refusing the truth to α and thereby violating his policy of non-one-sidedness[32]. Actually, the preceding dialogues have already made clear that the Jain did deny three times (steps 4, 10 and 14) while the Mādhyamika did affirm five times (steps 2, 6, 8, 12, and 16).

If so, the radically opposed attitudes of the Jainas and Mādhyamikas should find their own limits with the sort of sentences to be questioned: denying and affirming are about the nature of reality, rather than about one's own mental states. Such a limit of dialectic might be what Aristotle had in mind, when he attempted to show the attitude of Heracliteus with respect to the PNC is self-defeating. But he failed to make his point with his elenctic strategy, locating the trouble in the propositions (affirming α and affirming not-α) rather than his opponent's propositional attitudes (af-

[31] Nāgārjuna's following stance is the key to his allegedly radical skepticism: "If I had a thesis, I would be wrong. But I have no thesis. Therefore there is nothing wrong with me." ("To keep one away from the vain discussions", Number 29). What is the content of the thesis at hand? It is likely to be a first-order thesis, i.e. a statement about any given state of affairs. Whether Nāgārjuna would have also claimed to have no thesis about his own attitudes remains unclear, however.

[32] This leads to the reintroduction of the law of double negation in the form of an illocutionary law of *double denial*: the denial of $\sim\alpha$ needn't entail the affirmation of α, given that $\mathbf{a}_1(\alpha) = 0$ needn't entail that $\mathbf{a}_2(\alpha) = 1$ (compare with $\mathbf{A}(\alpha) = \langle 0, 0\rangle$); on the other hand, the denial of the denial of α entails the affirmation of a, given that $\mathbf{a}_1(\alpha) \neq 0$ does entail that $\mathbf{a}_1(\alpha) = 1$.

firming α and not affirming α). Admittedly, these Indian logics were much more concerned with metaphysical topics and soteriological ends than having the final word in every yes-no answer game.

7 Conclusion

We have proposed a reconstruction of the Jaina and Mādhyamika logics by means of a question-answer semantics. The result of such an enterprise is a rational reading of these Indian schools through modern logical glasses, including the logical tool of many-valuedness that presented skepticism and pluralism as radically opposed to each other and separated by a middle view of judgment that is Aristotle's bivalent way of doctrinalism. Many-valuedness accounts for the seven judgments of Jaina *saptabhaṅgī*, while a more general logic of attitudes displays Jaina and Nāgārjuna's stances within a four-valued semantics that characterizes both Mādhyamika skepticism (the value $\langle 0, 0\rangle$) and Jaina pluralism (the value $\langle 1, 1\rangle$).

Above all, the main import of **QAS** is to pay attention to the dialectical role of questions and answers in the Indian approach of logic: just as the Megarics emphasized the dialogical nature of philosophical investigation in contrast to the Aristotelian monological view of truth and falsity as transcendental values, we want to keep in mind that the Indian logicians introduced their statements in the form of answers to speculative questions. Jaina metaphysical pluralism also made sense of their inconsistent judgments, while the skeptic flavor of Nāgārjuna's philosophy explains his systematic denial to any question about the nature of reality.

Last, but not least: one of the most intriguing case studies has concerned the meaning of *avaktavya* (non-assertibility), the third basic judgment of Jaina logic. This predicate should not be confused with common self-contradiction, where a sentence and its negation are said to be both true at once and in the same respect. The commentators frequently claimed that the Jainas subscribed to PNC in their various reasonings: so non-assertibility refers to another, milder view of contradiction than coexistence of incompatible properties in the same subject. Rather, we support Tripathi's interpretation of *avaktavya* in the sense of non-distinction: the Jaina third judgment might mean that some objects (S) cannot be predicated by any property, that is, neither of one of them (P) or any of their complementaries (not-P). Rather than a plea for self-contradiction, avaktavya seems to argue for the impossibility to predicate anything of some such "absolute subjects" as ātman or Brahman because these would stand beyond any set

of definite properties. Such a tentative explanation would match with the Hegelian alternative process of *Aufhebung* (or "sursumption"), in contrast to the predicative process of subsumption that systematically describes a subject S as falling under a given set of properties P[33].

References

[1] Aristotle. *Metaphysics*.

[2] A.J. Bahm. "Does Seven-Fold Predication equal Four-Cornered Negation reversed?". *Philosophy East and West*, 7:127–30, 1958.

[3] N. Belnap. "A useful four-valued logic". In Dunn J.M. and Epstein G., editors, *Modern Uses of Multiple-Valued Logic*, pages 8–37. Dordrecht: D. Reidel Publishing Company, 1977.

[4] F. Bharucha and R. Kamat. "Syādvāda theory of Jainism in terms of deviant logic". *Indian Philosophical Quarterly*, 9:181–7, 1984.

[5] S.S. Chakravarti. "The Mādhyamika *Catuṣkoṭi* or Tetralemma". *Journal of Indian Philosophy*, 8:303–6, 1980.

[6] J. Ganeri. "Jaina logic and the philosophical basis of pluralism". *History and Philosophy of Logic*, 23:267–81, 2002.

[7] P. Gokhale. "The logical structure of Syādvāda". *Journal of Indian Council of Philosophical Research*, 8:73–81, 1991.

[8] M.H. Gorisse. "Non-one-sideness: context-sensitivity in Jain epistemological dialogues". In *A Day of Indian Logic*. ILLC Technical Report X-2009-04, Amsterdam, 2009.

[9] M.H. Gorisse. "The art of non-asserting: dialogue with Nāgārjuna". In R. Ramanujam and Sarukkai S., editors, *Springer Lecture Notes in Artificial Intelligence*, volume 5378, pages 257–68. FoLLI Series, Springer, 2009.

[10] L. Horn. *The Natural History of Negation*. University of Chicago Press, 1989.

[11] L. Keiff. "Ultimately and conventionally : some remarks on Nāgārjuna's logic". In *A Day of Indian Logic*. ILLC Technical Report X-2009-04, 2009.

[33] Here is an alleged description of the Brahman by himself: "This whole universe is filled by me in immaterial form; all beings are in me, but I am in them. Yet those born are not within me. Behold my kingly rule: my self sustains all beings, is not in them but creates them. Just as the mighty wind everlastingly occupies the space above us and moves throughout it, so do all created beings occupy me." (*Bhagavad-Gītā*: Chapter 9, verses 4-7). This seems to match with our description of S as an ultimate class.

[12] B.K. Matilal. "The Jaina contribution to logic". In J. Ganeri and H. Tiwari, editors, *The Character of Logic in India*, pages 127–39. State University of New Press, 1998.

[13] D. Mohanta. "The use of Four-Cornered Negation and the denial of the Law of Excluded Middle in Nāgārjuna's logic". In A. Schumann, editor, *Logic in Religious Discourse*, pages 44–53. Ontos Verlag, Paris & Frankfurt, 2009.

[14] T. Parsons. "Assertion, denial, and the Liar Paradox". *Journal of Philosophical Logic*, 13:137–52, 1984.

[15] G. Priest. "Jaina logic: a contemporary perspective". *History and Philosophy of Logic*, 29:263–79, 2008.

[16] P.T. Raju. "The Principle of Four-Cornered Negation in Indian philosophy". *Review of Metaphysics*, 7:694–713, 1954.

[17] N. Rescher. "Quasi-truth-functional systems of propositional logic". *Journal of Symbolic Logic*, 27:1–10, 1962.

[18] F. Schang. "A plea for epistemic truth: Jaina logic from a many-valued perspective". In A. Schumann, editor, *Logic in religious discourse*. Ontos Verlag, Paris & Frankfurt, 2009.

[19] J. Searle. *Speech Acts*. Cambridge University Press, 1969.

[20] Y. Shramko and H. Wansing. "Hyper-contradiction, generalized truth-values and logics of truth and falsehood". *Journal of Logic, Language and Information*, 15:403–24, 2006.

[21] M. Siderits. "Nāgārjuna as antirealist". *Journal of Indian Philosophy*, 16:311–25, 1988.

[22] R. Sylvan. "A generous Jainist interpretation of core relevant logics". *Bulletin of the Section of Logic*, 16:58–66, 1987.

[23] R.K. Tripathi. "The concept of *avaktavya* in Jainism". *Philosophy East and West*, 18:187–93, 1968.

Navya-Nyāya Logic

PRABAL K. SEN [*] and AMITA CHATTERJEE [†]

> In short, the Nyāya strategy is to appeal to our intuitions about knowledge, in order to learn something about reasoning and not vice versa. Bimal Krishna Matilal [1]

> In its first meaning, a logic is a collection of closely related artificial languages... In its second but older meaning, logic is the study of rules of sound argument. Wilfrid Hodges [2]

The expression 'Navya-Nyāya' literally means 'the recent Nyāya' or 'the new Nyāya', usually employed for indicating the later phase of the Nyāya school of philosophy, as distinguished from its earlier phase, which is commonly known as 'Prācīna Nyāya', i.e., 'the earlier Nyāya' or 'the old Nyāya'. Akṣapāda Gautama (c. 100 CE) is traditionally regarded as the founder of the Nyāya school, and a set of aphorisms known as Nyāya-sūtra-s that are ascribed to him happens to be the oldest available text of this school. Quite a few commentaries and subcommentaries on these aphorisms were written, many of which are now lost, and are known only from references to them in later works. The available texts in this series of commentarial literature are (i) Nyāyabhāṣya of Vātsyāyana (fourth century), (ii) Nyāyavārttika of Uddyotakara (seventh century), (iii) Nyāyavārttika-tātparya-tīkā of Vacaspati Miśra I (ninth century), (iv) in Nyāyavārttika-tātparya-pariśuddhi of Udayana (tenth century). The independent works like (i) Nyāyasāra (with the autocommentary Nyāyabhūṣaṇa) by Bhāsarvajña (tenth century) and Nyāyamañjarīby Jayanta Bhatta (ninth century) are also important texts that belongs to this phase of Nyāya philosophy. According to the tradition, Udayana's works formed the watershed between the Old and the Navya-Nyāya, which in the process of defending and explicating the Nyāya tenets also anticipated many theses and approaches of

[*] Professor of Philosophy, University of Calcutta
[†] Vice Chancellor, Presidency University, Kolkata
[1] Matilal (1986) p. 126
[2] 'Classical logic I: First Order Logic' Lou Goble (2001)

the later Naiyāyikas. Indian theory of inference forms part of Indian epistemology (pramāṇavāda) and is intimately connected with the ontology of a system. We, therefore, begin our explorations in Navya-Nyāya logic with a brief account of the metaphysical basis of the system. The Nyāya school of philosophy upholds direct realism and pluralism; and it shares this outlook in common with the Vaiśeṣika school, which is traditionally maintained to be founded by Kaṇāda (second century CE). The Vaiśeṣika system has been described as 'a synthesis between philosophy of nature, ethics and soteriology' [3], and this is also true of the Nyāya school, though here we find in addition a lot of emphasis on epistemology and the rules that should be observed in philosophical debates.

The doctrines of Nyāya philosophy were severely criticised by a number of opponents, the principal among them being the Buddhists of the Mādhyamika, Yogācāra and Svatantra-Yogācāra sects. For the Naiyāyikas, the world contains innumerable entities that are in principle knowable and nameable. Each such entity, whether external, like a pot, or internal, like a cognitive state, is real, and has an intrinsic nature (svabhāva). Many of these entities are eternal, and even those that are non-eternal, are stable, i.e., non-momentary (akṣaṇika). Many of these entities are mutually related, and these relations, which are as real as their relata, are of various kinds. The relation that links most of these existent objects is the relation between (i) the entities that are located (ādheya), and (ii) the entities where these entities are located (ādhāra). This relation between location and locatee is known as dharma-dharmī-bhāva. This general relation may again obtain through some specific relations. For example, when we cognize a man as characterized by a stick, the relation between the man and the stick is that of contact (saṃyoga). Again when we cognize an animal as a white cow, the relation of the animal with white colour and the universal, viz., cowness is that of inherence (samavāya). None of these claims would be admitted by the Buddhists. For the Mādhyamika Buddhists, the objects of our experience are devoid of nature (niḥsvabhāva); for the Buddhists of the early Yogācāra school, pure consciousness (vijñaptimātra) is the sole reality, there being thus no external objects; and according to the Svatantra Yogācāra school, whatever is real is also momentary, which effectively precludes the possibility of such things being either located in, or related with anything else. Each entity, they maintain, is unique (svalakṣaṇa) and unrelated. The commentaries and subcommentaries that grew around the Nyāyasūtra-s tried to defend the Nyāya doctrines by rejecting the Buddhist

[3] Partha Ghosh (2010) p. 258

views. Navya-Nyāya philosophers did not forget these issues when they developed their language and logic.

One of the favourite strategies of the Buddhists was to show that the entities admitted by the Naiyāyika-s cannot be properly defined, and they tried to establish this by pointing out defects in such definitions proposed in the Nyāya texts. Another strategy was to point out that the Nyāya doctrines were beset with logical difficulties like self-dependence, mutual dependence, infinite regress, etc. The Buddhists also tried to show that in many cases what was regarded as a single or unitary entity by the Naiyāyika-s could not be so, since each of them harboured mutually incompatible properties. The adherents of the Nyāya school were hard-pressed to find out some way for answering such criticisms, and this more or less compelled them to find out some techniques for formulating precise and immaculate definitions; and also for answering the dialectical arguments of the Buddhists. In some cases, minor modifications in the earlier doctrines were also made, though the basic doctrines and the commitment to realism and pluralism were not compromised in any way.

II

By combining the Nyāya epistemology with the Vaiśeṣika ontology, philosophers like Śaśadhara, Manikaṇṭha Miśra, TaraṇīMiśra, Sondaḍa Upādhyāya and others initiated a new trend of philosophizing in Mithila – a region in northeastern India. It is, however, Gaṅgeśa (thirteenth century) who integrated and popularized the technique of subtle argumentation in his magnum opus Tattvacintāmaṇi (TCM) and is regarded as the founder of the Navya-Nyāya tradition. The tradition was carried forward in Mithila by Vardhamāna Upādhyāya (fourteenth century), Yajñapati Upādhyāya (fifteenth century), and Pakṣadhara Miśra (fifteenth century), among others. The novelty and originality of the Navya Nyāya school is found not so much in introducing new topics of philosophical discussion but in the method employed, in devising a precise technical language suitable for expressing all forms of cognition. By the time the Navya-Nyāya language was devised, Buddhism, the principal opponent of Old Nyāya had become almost extinct in India. Navya-Nyāya philosophers had the Mīmāṃsaka-s as their chief adversary, but their language was strong enough to withstand attacks from both Buddhism and Vedānta.

From Mithila, Navya-Nyāya travelled to Navadvīpa, in Bengal. Pragalbha Miśra, Narahari Viśārada and Vāsudeva Sārvabhauma are the notable early exponents of Navya-Nyāya in Navadvīpa. The unorthodox logician, Raghunātha Śiromaṇi (sixteenth century), who was a disciple of

Vāsudeva Sārvabhauma wrote a commentary on TCM entitled Dīdhiti, in which he went far beyond Gaṅgeśa by introducing changes in Navya-Nyāya metasphysics and epistemology. Subsequent prominent proponents of Navya-Nyāya in Bengal – including Bhavānanda Siddhāntavāgīśa, Mathurānātha Tarkavāgīśa, Jagadīśa Tarkālaṃkāra, and Gadādhara Bhaṭṭācāryya – wrote sub-commentaries on Dīdhiti, which contributed to the fullest development of Gaṅgeśa's technique of reasoning. The fame of Navadvīpa Naiyāyikas spread all over India, and scholars from other schools too adopted the Navya-Nyāya language. This highly technical language became the medium for all serious philosophical discussion by the sixteenth century, irrespective of the ontological, epistemological, and moral commitments of the discussants. However, one must remember that though the Navya-Nyāya language can be successfully dissociated from its context, Navya-Nyāya was developed as a complete system of philosophy with its epistemology, logic, ontology and soteriology.

'Navya-Nyāya logic', writes Sibajiban Bhattacharya, 'is mainly a logic of cognitions'. [4] A piece of cognition has at least three elements – viśeṣya (qualificandum), prakāra or viśeṣaṇa (qualifier), and saṃsarga or the qualification relation between them. If, for example, one's cognitive content is a-R-b, i.e., b is located in a by the relation R, then says the Naiyāyika, one is directly aware of a, b, and R where a and b are things in the real world and not mere representations of things and the relation R actually obtains between a and b. So a cognitive content a-R-b is true if and only if b is located in a by the relation R. So, when one cognizes a man with a stick, the man is the qualificandum, the stick is the qualifier and the relation between the man and the stick, in this case, is contact or saṃyoga. This piece of cognition will be true (pramā) if and only if the man being perceived has contact with a stick.

It is, therefore, obvious that the Navya-Naiyāyika-s are in favour of giving a de re reading of a cognitive content. This situation, when viewed in terms of locus-located relation is: b is located in a or a superstratum (ādheya) of a in the relation R in a-R-b, and a is the locus or the substratum (ādhāra) of b in the relation R in a-R-b. Generally speaking, according to Navya-Nyāya, the basic combination which expresses a cognitive content is a locus-locatee combination of the form 'a has f-ness' / '(there is) f-ness in a' ('the lotus has redness'/ '(there is) redness in lotus', which is expressed in ordinary language as 'the lotus is red'.). In a perspicuous account of a cognitive content, the Navya-Naiyāyika would like to make

[4] Haaparanta (2009) p.963

explicit the connection between the lotus and its colour in consonance with their own categorical framework.

It is evident from the above analysis that relations play a crucial role in the Navya-Nyāya concept of a cognitive content. Over and above the two relations of contact and inherence admitted by the Vaiśeṣika-s, Navya-Naiyāyika-s define many new relations for precisifying our cognitive content. A standard definition of relation in terms of subjuncts/superstratum (anuyogī) and adjuncts/substratum (pratiyogī) given by Gadādhara is as follows.

> When xRy is a cognitive content, R is a relation of x to y iff x is the adjunct of R (one which is related) and y is the subjunct (to which x is related) of R.

The Navya-Nyāya way of expressing a relation is always as xRy, where the entity to the left of R is the adjunct and the entity to the right of R is the subjunct. The Navya-Naiyāyikas admit two types of relation, occurrence-exacting (vṛtti-niyāmaka) and non-occurrence-exacting (vṛtti-aniyāmaka). An occurrence-exacting relation always gives the impression that one entity is located in another entity, while a non-occurrence-exacting relation does not do so. The latter only makes us aware that the two terms are related. It is easier to identify the adjunct and subjunct of a relation of the former type; the adjunct is that which is located and the subjunct is that where the adjunct is located but in the second type adjunct and subjunct are identified depending on the fiat of the cogniser. The Navya-Naiyāyikas mainly use four types of direct relation: (1) contact (saṃyoga); (2) inherence (samavāya); (3) svarūpa [5]; and (4) identity (tādātmya). Of these, the first two are occurrence-exacting, svarūpa is sometimes so, and identity is not. They admitted some indirect relations (paramparā sambandha) too, e.g., the colour of a cloth's thread resides in the cloth by an indirect relation composed out of inherence and its inverse, viz., sva-samavāyi-samavetatva. According to the Nyāya school all these relations, direct and indirect, are binary relations.

It is now time to give a minimal account of the Navya-Nyāya language, which is a higher-order technical language but, strictly speaking, is not a formal languguage.

The primitive terms of the language are the nouns or nominal stems like ghaṭa (pot), dhūma (smoke), vṛkṣa (tree), kapi (monkey), etc. By adding

[5] Svarūpa will be left untranslated because any English term is bound to distort its meaning; it is identical with either one or both the relata.

the simple suffix 'tva' or tā', many new abstract terms are generated. For example, by adding 'tva' to dhūma, abstract terms like dhūmatva (smokeness or smokehood), which is a universal (jāti), can be generated. The suffix 'tā' is used to generate relational abstract expressions such as causehood (kāraṇatā), locushood (ādhāratā), and their corresponding inverse relational expressions such as effecthood (kāryatā), located-hood or superstratumhood (ādheyatā/vṛttitā). Navya-Nyāya also uses a possessive suffix 'mat' (or its grammatical variant 'vat') meaning 'possessing' to generate new concrete terms as in 'vahnimat' or fire-possessing.

There is an operator known as the determiner-determined-relation (nirūpya-nirūpaka-bhāva) which obtains between correlatives like locushood and locatedhood, causehood and effecthood, motherhood and sonhood, etc. To explain, when a is the locus of b, the relational abstract locushood (ādhāratā) resides in a and its correlative locatedhood (ādheyatā) resides in b. The property of locushood residing in a determines or is determined by the locatedhood residing in b, depending on the direction of the relation. This determining relation guarantees exact description of the content of cognition. Suppose, one sees that there is a plum in a bowl and a book on the table. In terms of locus-locatee these two facts can be described as follows. The plum has a locatedhood determined by the bowl and the book has the locatedhood determined by the table. Similarly, the bowl has the locushood determined by the plum and the table has the locushood determined by the book. As the locatedhood of one entails the locushood of the other and vice versa, there exists a determiner-determined relation between locatedhood and the locushood. Hence the cognitive content, viz.,'the plum is in the bowl' can be rephrased as the plum possesses a locatedhood that is determined by the correlative locushood residing in the cup (kuṇḍaniṣṭha-ādhāratā-nirūpitā-ādheytāvat-vadaram) and 'the cup has a plum in it' can be explained as the cup possesses a locushood that is determined by the correlative locatedhood residing in the plum (vadaraniṣṭha-ādheyatā-nirūpitā-ādhāratāvat-kuṇḍam).

Another very important operator is avaccehadakatā, or limitorhood. This operator performs multiple functions in a cognitive situation. (1) It states explicitly the mode of presentation of an object, (2) it acts as a quantifier in a content-expressing sentence, and (3) it helps us to determine which pair of sentences is contradictory.

The first operation of a delimitor can be explained with the simple example of 'the floor is with a pot'. When we cognize something, some qualifiers are expressed in the first order language and some are merely understood. The qualifiers which are merely understood are called the 'delim-

itors'. So in the above example, the floor is the qualificandum and the pot is the qualifier, both of which have been mentioned. But there are two other unmentioned qualifiers, viz., potness and floorness qualifying respectively pot and floor and hence are the delimitors. A full account of the content undoubtedly requires these delimitors and if the mode of presentation or the delimitor is properly specified, we can set aside all confusions. Besides the delimitors of the qualificandum and the qualifier, there exists a delimiting relation too, which in this context is contact. So, fully spelt out, the sentence 'the floor is with a pot' (ghaṭavat bhūtalam) turns out to be: the floor delimited by floorness possesses a locushood that is determined by the correlative locatedhood residing in the pot delimited by potness in the delimiting relation of contact (saṃyogasambandha-avacchinna-ghaṭatva-avacchinna-ghaṭaniṣṭha-ādheyatā-nirūpita-bhūtalatva-avacchinna-ādhāratā-vat-bhūtalam). The situation mentioned above is being represented by a diagramme [6].

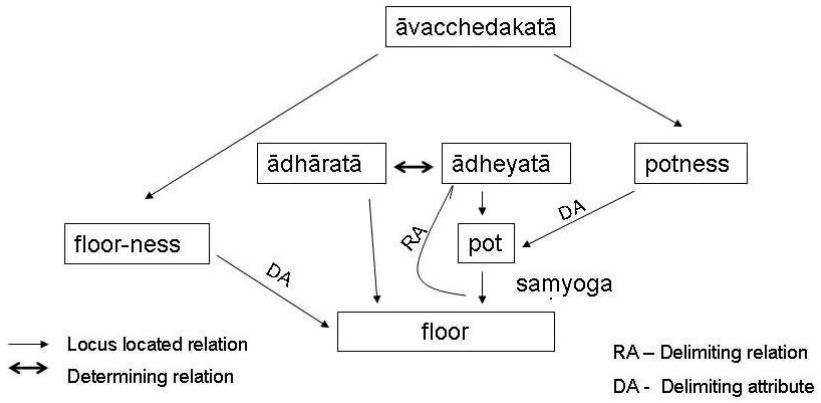


Figure 1: Ghaṭavat bhūtalam (The floor has a pot on it)

To understand how the Naiyāyika -s use 'delimitor' to state the quantity of the cognised structure, we may consider two interesting examples given by Maheśa Chandra Nyāyaratna [7]. When the content-expressing sentence is: 'A person having brahminhood is scholarly', it does not signify that all brahmins are scholarly. Rather this means that the property of being brahmin and scholarship are sometimes found in the same locus, i.e., 'Some brahmins are scholarly'. On the other hand, when the content expressing

[6] All the diagrammes is the text have been adapted from Ramesh Chandra Das (2006)
[7] Mahesa Chandra Nyāyaratna (1973)

sentence is: 'Men are mortal', the qualifier mortality pervades the delimitor of the qualificandum, i.e., humanity. Hence, the sentence should be construed as universally quantified. The general rule is: when the chief qualifier is just co-resident (samānādhikaraṇa) with the property of being the qualificandum, the content-expressing sentence should be taken as particular but when the chief qualifier is delimited by the delimitor of the qualificandum (viśeṣyatāvacchadakāvacchinna), the sentence is to be construed as universal.

To find out how delimitorhood helps us determine contradiction in a pair of cognitive content or sentences, let us consider the following example. When a strong breeze blows over a tree, the leaves and the branches of the tree are seen to tremble. The roots and the trunk of the tree do not, however, tremble. Thus it may be said that the tree is characterized both by trembling (sakampatva) and absence of trembling (akampatva), which are opposed to each other. Using the delimiting operator, the Navya-Naiyāyika would show that though trembling and the absence of trembling are present in the tree, that does not affect the unity of the tree; nor does it amount to the assertion of a contradiction that the same tree is both trembling and non-trembling at the same time. He would rather say that the tree in respect of its branches (śākhāvacchedena) is the locus of trembling, whereas the same tree, in respect of its root (mūlāvacchedena) is the locus of the absence of trembling. In like manner, when a monkey sits on a tree, the tree may very well have contact with that monkey in respect of one of its branches; while the same tree in respect of its roots may simultaneously harbour the absence of that contact. In such cases, the locushood resident in the tree is said to be delimited (avacchinna) by different delimitors (avacchedaka) – the tree, as delimited by its branch is the locus of contact with monkey, and this is in no way opposed to the fact that the same tree, as delimited by its roots, is the locus of the absence of contact. There would be a contradiction if the tree would have been a location of a contact and the absence of that contact with respect to the same delimitor.

In this connection, it must be mentioned that the presence or absence of a certain thing in a certain locus is always through some specific relation. Thus, a pot may be present on the floor of a room through the relation known as contact, and at the same time, present in its own constituent parts through the relation of inherence. The pot, however, is not located in the floor through inherence, or in its own parts through contact. But this does not lead to any contradiction.

A logical language remains incomplete without an account of negation. To understand the Navya-Nyāya concept of negation we need to understand

their ontology of absence. Absence, they point out, is not merely a logical or linguistic operator, it is as objectively real as a positive entity is. Four types of absence are admitted in the system: (1) mutual absence or difference (anyonyābhāva), e.g., a jar is not a pen and vice versa; (2) absence of not-yet type (prāgabhāva), e.g., absence of a bread in flour before it is baked; (3) absence of no-more type (dhvaṃsābhāva),e.g., absence of a vase in its broken pieces and (4) absolute absence (atyantābhāva), e.g., absence of colour in air. So an absence is always of something and that something is called the counterpositive or the negatum (pratiyogī) of that absence. Consider the absence of smoke in a lake. Smoke is the counterpositive (pratiyogī) of the absence of smoke and pratiyogitā or the relation of counterpositiveness is the relation between an absence and its counterpositive. Here, the lake is the locus (anuyogī) of the absence. Hence, anuyogitā connects the absence in question with its locus. Here absence is that of smoke in general (dhūma-sāmānya) and not this or that particular smoke, hence it is called dhūma-sāmānyābhāva. Next, let us explain the notion of a delimitor and the delimiting relation in the context of an absence. When x is in y, x is related to y in a particular relation and that relation is the delimiting relation. Similarly, when there is an absence of x in y, a counterpositiveness must be there in x and there must be a relation to delimit that counterpositiveness. Suppose there is smoke on a mountain. Here the delimiting relation is contact (saṃyoga). There is at the same time absence of smoke on the same mountain by the relation of inherence because smoke never resides in a mountain by the relation of inherence. Again, smoke is absent on the mountain by the relation of identity or tādātmya, since smoke and mountain cannot be identical. So counterpositiveness in the first case is delimited by the relation of inherence whereas in the second case the delimiting relation is identity. At the same time counterpositiveness so related determines (nirūpaka) the said absence. Thus the first absence is determined by the counterpositiveness residing in smoke delimited by the relation of inherence (samavāyasambandhāvacchinna-pratiyogitā-nirūpita-dhūma-sāmānyābhāva) and the second absence is determined by the counter-positive-ness residing in smoke limited by the relation of identity (tādātmyasambandhāvacchinna-pratiyogitā-nirūpita-dhūma-sāmānyābhāva).

III

Navya- Naiyāyika-s, like the Buddhists and the Old Naiyāyika-s divide inference broadly into two types. Svārthānumāna (SA) or inference-for-oneself deals with the psychological conditions, i.e., causally connected

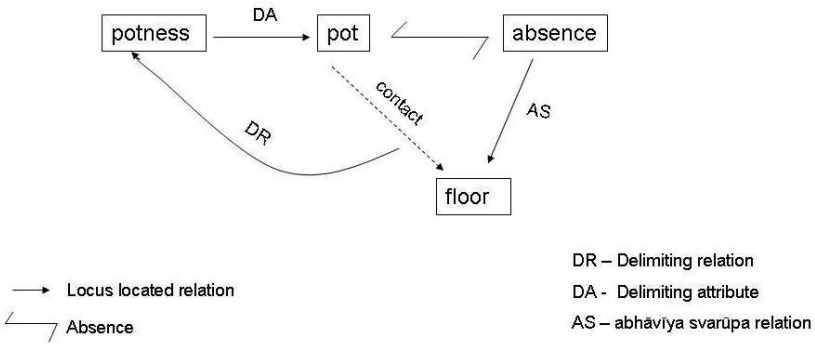


Figure 2: Bhūtale ghaṭābhāva (absence of pot on the floor)

cognitive states leading to one's own inferential cognition, while Parārthānumāna (PA) or inference-for-others essentially deals with the proper linguistic expression of this inference with a view to communicating it to others.

SA which is a process of mental reasoning par excellence consists of four steps, each of which is a state of cognition causally connected with the immediately preceding state. The process can be best explained with their typical example. A person first sees that (a) the hill (pakṣa/ the locus of inference) possesses smoke (hetu/the ground of inference/probans). This is perceptual cognition which reminds him that (b) wherever there is smoke there is fire (sādhya/the provable/probandum) as he has always observed in a kitchen. The first step is technically called pakṣadharmatājñāna, meaning the probans is known to be present in the locus of reasoning. The second step (known as vyāptijñāna) is memory or a recollective cognitive state of the universal concomitance between smoke and fire. Then (a) and (b) are combined to produce a complex form of cognition called 'parāmarśa' or 'consideration' of the form (c) the hill possesses smoke pervaded by fire and then follows the conclusion (d) Therefore, the hill possesses the fire.

PA has five constituents arranged in the order pratijnā or assertion, hetu or reason, udāharaṇa or example, upanaya or application and nigamana or conclusion. The typical example of a fully fledged parārthānumāna is the following.

> Pratijñā: The hill possesses fire. (stating what is to be proved)
>
> Hetu: The reason is smoke (stating the ground of inference)
>
> Udāharaṇa with vyāpti: Wherever there is smoke, there is fire as in a kitchen

Upanaya: The hill is similar (in possessing smoke)

Nigamana: Hence, the hill possesses fire.

Though the conclusion of a PA appears to be the same as the first step, these two perform two different tasks. The first step just asserts the thesis while the conclusion declares that what is to be proved has been proved. According to the tradition, the first step is said to be generated by verbal cognition, the second is established by inference, in the third step, example is acquired through perception and the fourth step is based on cognition of similarity. Since these four steps are established by four sources of true cognition admitted in the Nyāya school, the Naiyāyika considers this five-membered argument as the demonstration par excellence (parama-nyāya).

Gaṅgeśa in the Vyāptivāda of TCM has rejected many definitions of pervasion (vyāpti) given by the opponents of which only the first will be analysed here. Simply stated, the definition runs thus: Pervasion or vyāpti is the absence of occurrence of the hetu in every locus of absence of the sādhya.[8] This definition, however, has been amended quite a number of times to free it from the charges of over-coverage (ativyāpti) and under-coverage (avyāpti). A ramified version of the definition, though it is not the final version, is:

The hetu is pervaded by the sādhya if the hetu is in no way occurrent by the relation of hetutāvacchedaka in any locus of the absence of the sādhya which is characterized by the sādhyatāvacchedakadharma and also by the sādhyatāvacchedakasambandha. [9]

We have said before that pervasion is the relation of invariable concomitance of the ground of an inference (hetu) and the thing to be inferred (sādhya). Without the knowledge of this relation it is not possible to infer. In a valid inference, 'The hill has fire because it has smoke', the sādhya is fire, the hetu is smoke and pakṣa or the locus is the hill. Sādhyatāvacchedaka-sambandha is the relation in which the sādhya resides in the pakṣa. As fire resides in the hill by the relation of contact (saṃyoga), the limiting relation is contact. The property which is the delimitor of the sādhya in this case is fireness (vahnitva) and not the property of producing burns (dāhajanakatva). Similarly by hetutāvacchadkasambandha is meant the relation in which the hetu resides in the pakṣa. In the given instance, that relation is also contact, as smoke too resides in the hill by contact. This

[8] sādhyābhāvavadvṛttitvam

[9] sādhyatāvacchedakasambandhāvacchinna-sadhyatāvacchedakadharmāvacchinna-sādhyatāvacchedakāvacchinna-pratiyogitāka-sādhyābhāvādhikaraṇanirūpita-hetutāvacchedakasambandhāvacchinna-vṛttitāsāmānyābhāvo vyaptiḥ.

absence of occurrence of smoke is again absence of occurrence of smoke in general and not of any particular smoke. So there is the relation of pervasion between the hetu smoke and the sādhya fire as there is general absence of occurrence of the hetu smoke by the delimiting relation of contact, determined by every locus of absence of the sādhya fire, counterpositiveness of which is delimited by the relation of contact and the attributive delimitor firehood. Plainly speaking, fire pervades smoke because no smoke ever resides by way of contact in a lake or anywhere else, which is the locus of absence fire qua fire. While exploring the psychology of reasoning, the Naiyāyika-s have also specified three pre-conditions of the possibility of engaging in a reasoning. The reasoning process cannot even take off if these pre-conditions are not fulfilled. Reasoning process begins

1. if the reasoner is not aware that there is fire on the hill, i.e., that the probandum is present in the locus. Of course, if the reasoner desires to reason to the effect that there is fire on the hill even after being sure of the fact, he may indulge in reasoning;

2. if the reasoner does not believe that there is absence of fire on the hill, i.e., the probandum is absent in the locus; and

3. if the reasoner does not believe or doubt that the hill is characterised by some property which is concomitant with the absence of fire, i.e., the locus is characterised by some probans, which is invariably concomitant with the absence of the probandum.

The second and the third pre-condition require ascription of minimal rationality to the reasoner in the sense that the person naturally avoids the alternatives that lead to contradiction. Next, the Naiyāyika-s discuss in details how a reasoner can be sure that SA will lead to a sound conclusion. According to them, the psycho-cognitive states previously specified ensure the truth of the conclusion provided the probans, which serves as the ground of reasoning is legitimate. A probans is legitimate if and only if it possesses five features, viz.,

a It is present in the locus of reasoning (pakṣa-sattva);

b It is present in a similar location (sapakṣa-sattva);

c It is not present in any dissimilar location (vipakṣa-asattva);

d It is not associated with the contradictory of the probandum in the locus (abādhitatva);

e If another probans tending to prove the contradictory of the probandum is not present in the locus (asatpratipakṣitatva)

These five features provide the truth conditions of the cognitive states involved in SA; a) is the truth condition of pakṣadharmatājñāna, b) and c) are the truth conditions of vyāptijñāna disjunctively and thus become the truth condition of parāmarśajñāna too; d) and e) have a direct relevance to the truth of the conclusion. The violation of these conditions leads to the types of defective probans known as asiddha (unestablished), viruddha (hostile), savybhicāra (deviating), bādhita (contradictory) and satpratipakṣa (counterbalanced) respectively. All these defects of probans can be present in one non-veridical inference, e.g., 'the lake has fire because it has potness'. In this example, the lake is the inference-locus, fire is the probandum and potness is the probans. It violates the first condition because the probans potness is not present in the locus of reasoning, the lake. It goes against the second condition because potness is present only in pots but absent in various loci of fire, hence the probans is opposed or hostile. A more familiar example of this type of fault is: sound is eternal as it is an effect. The inference under discussion is also vitiated by the defect due to a deviating probans. Here the probans potness which is present only in pots can easily reside in a locus which is characterised by the absence of fire. That shows that potness is not invariably concomitant with fire, the probandum. In this example, the probans potness becomes contradictory and hence illegitimate, if the lake does not possess fire. Again, it is easy to show the possibility of the existence of an alternative probans, say, water, capable of proving the absence of fire in the lake, thus counterbalancing the force of the original probans and preventing the conclusion. All these defective probans are faulty because they somehow block the conclusion of the inference. Thus, it is obvious that the psychological conditions of SA are related to the conditions of validity of it in such a way that the fulfilment of the former guarantees the fulfilment of the latter. Having shown this in the context of SA, the Navya-Naiyāyika-s work out what role these conditions play in PA, the full-fledged explicit form of reasoning employed primarily for convincing others. As the theory of PA became more and more developed, many structural and transformation rules of reasoning were abstracted. These truth-preserving rules enabled the reasoners who had access to the same set of premises to arrive at the same conclusion. Thus the theory of reasoning which began as a description of psychology of proof as well as a way of knowing was transformed into a logical theory, not as a formal rule-driven axiomatic theory but as a model-theory.

One area where the adherents of Navya-Nyāya added a novel feature of philosophical discussion was the formulation of anugama (i.e., consecutive or uniform character). It is often found that the same term is applied to indicate a number of entities, even though at first sight, no common feature can be found in them. Normally, one would expect that application of the same word to a number of things depends on the apprehension of some common feature in them; and if such apprehension is to be veridical, then some such common feature should actually be present in those entities. The problem is to find out some such common properties. The relation of pervasion that justifies the inference of sādhya (S) from hetu (H) may be apprehended in two ways:

1. Wherever H is present, S is also present; and

2. Wherever S is absent, H is also absent.

The first of these is known as anvaya-vyāpti, while the second is known as vyatireka-vyāpti. It may be noted here that (i) and (ii) are not interchangeable, because if no vipakṣa can be found, then formulation of (ii) cannot be admitted; wheras if no sapakṣa is available, then (i) cannot be admitted . Both these are, however, regarded as cases of vyāpti. Both these types of vyāpti have, however, one property in common – viz., the property of being an object of the cognition which is contradictory to the cognition of deviation (vyabhicāra), which would ensue if there is any locus where H is present along with the absence of S. Thus, the property of being the object of knowledge which is opposed to the knowledge of deviation (vyabhicāradhīvirodhijñānaviṣayatva) is the common feature (anugama) that characterizes anvaya-vyāpti as well as vyatireka-vyāpti. We have discussed above the five types of 'defective reasons' (hetvābhāsa). Here again, the same term is being applied to different things that have apparently no common feature. Nevertheless, three definitions that are applicable to each of the hetvābhāsa-s have been formulated by Gaṅgeśa; one of them being as follows: If X is such that a veridical cognition of X prevents either an inference (anumāna) A or some cause of inference A, then X would be a hetvābhāsa with respect to A. [10]

These three definitions provide us with alternative anugama-s of the five types of hetvābhāsa. The Naiyāyika-s maintain that if the presence of the property S has already been ascertained in the locus P, then even if we are

[10] Yadvisayakatvena jñānasya anumiti-tatkāraṇa-anyatara-virodhitvam, tattvaṃ hetvābhāsatvam

aware of the presence of some property H that is pervaded by S in P, no inference of the form 'P has S' or 'S is present in P' will take place, unless we have a strong desire for inferring the presence of S in P. In accordance with this, the earlier Naiyāyika-s maintained that prior doubt regarding the presence of S in P, which they call pakṣatā, is a pre-condition of the inferential cognition 'P has S' or 'S is in P' Now such a doubt may assume various forms, e.g., (i) 'Does P possess S or not?' (ii) 'Is S present in P or not?', (iii) 'Is S counterpositive of an absence located in P or not?' and so on.Unless we can find here a common feature, it will be extremely difficult to express the causal connection between such a doubt and the said inferential cognition; because only any one, but not all of such doubts can be present before that inferential cognition.Here, again, Raghunātha Śiromaṇi has said that all such doubts are such that they are prevented from occurring by a cognition where the presence of S in P is ascertained [11]. Similar problems may also be raised about the avayava-s (components of inference).[12] Having thus discussed generally the Navya-Nyāya theory of inference, we show in the following figure no. 3 all the properties and relations that obtain in their prototypical sound inference 'the hill has fire as it has smoke on it'. In fact, there are six generated properties all related by different determining relations (nirūpya-nirūpaka-bhāva), shown in the figure no. 3 by (1) ..(6): (i) dhūmatva-niṣṭha-avacchedakatā, (ii) dhūma-niṣṭha-hetutā, (iii) vahnitva-niṣṭha-avacchedakatā, (iv) vahni-niṣṭha-sādhyatā, (v) parvatatva-niṣṭha-avacchedakatā, (vi) parvata-niṣṭha-pakṣatā. In an unsound inference, because of a defect in the probans, some of these relations are blocked (pratibaddha).

IV

Navya-Nyāya logic is a logic of terms and relations. There have been several partial attempts to symbolize Navya-Nyāya logic by using first order language (Bhattacharyya Sibajiban (1987), Ingalls (1951), Matilal (1968) Staal (1962), etc.). But these have neither increased the perspicuity of Navya-Nyāya language nor enhanced the power of Navya-Nyāya logic. We too are contributing our bit with the hope of getting a better understanding of the apparently formidable texts of Navya-Nyāya logic. Our endeavour, to begin with, is to glean the syntax of the Navya-Nyāya language from the brief overview mentioned above.

[11] yatra sādhyasya yādṛśasambandhāvagāhi-nirṇayanivartyaḥ yaḥ saṃśayaḥ sa tādṛśasambandhena sādhyānimitau pakṣatā

[12] For an in-depth discussion of this technique of anugama and its logical aspects, see B.K. Matilal (1968), pp. 83-86, and D. C. Guha (1968), pp. 281-293.

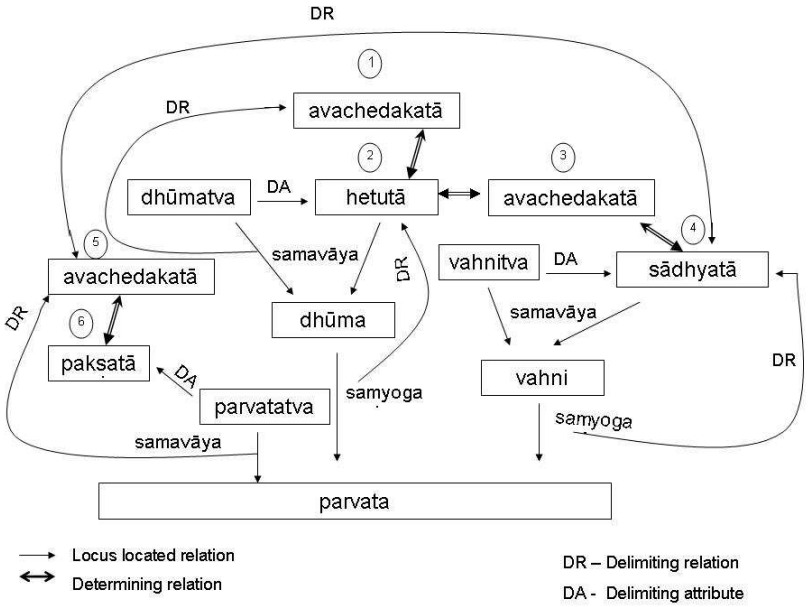


Figure 3: Parvato vahnimān dhūmāt (The hill has fire because it has smoke on it)

Navya-Nyaya syntax comprises the following.

Primitive terms: a,b,c, ... α, β, γ, ... $\alpha^{-1}, \beta^{-1}, \gamma^{-1}$, ...

Abstraction functors: Tv, Tā – such that if b is a term, then Tvb is also a term;

If α is a term, then Tā/Tva α is a term too.

Explanation: a, b, c, etc., stand for noun stems like ghaṭa (pot), vṛkṣa (tree), kapi (monkey). α, β, γ, ... stand for relational abstract expressions like kāraṇa (cause), ādhāra (locus), and $\alpha^{-1}, \beta^{-1}, \gamma^{-1}$, ...stand for inverse of them like kārya (effect), ādheya / vṛtti (located). Since ghaṭa (pot) is a term, ghaṭatva (pothood) is also a term; and since ādhāra (locus) ia term, so too is ādhāratā (locushood).

Formally, there is no need of introducing two sorts of primitive terms, yet we have taken two sorts simply to retain the intuitive difference between thing words and relational abstract expressions.

Complex term-forming operators: There are a number of 2-place operators used to form complex terms / sentence-surrogates, viz., L, D, A, C, P. Navya-Nyāya philosophers bring all of them under the category of 'sambandha' (relations). It will be obvious from the explanation given below that L, D, C, A, P are semantically distinct, and we are not offering any formal ditinguishing criterion. Besides these, there are standard logical particles – negation (N) conjunction ($\wedge$) and disjunction ($\vee$) . These logical particles too occur between two terms.

1. L such that a L b, α L a, α -1 L b
   Explanation: L is the locus-located relation (ādhāra-ādheya-bhāva), when a is located in b implying that locatedhood is in a and locushood is in b. For example, plum-in-a-cup should be understood as there is locatedhood-in- the- plum and locushood- in-the-cup.

2. D indicates the determiner-determined relation (nirūpya-nirūpaka-bhāva). D is such that α D β, while D -1 is β D -1 α . For example, while locushood determines locatedhood, locatedhood is determined by locushood and vice versa.

3. A is delimitation (avacchedaka-avacchinna-bhāva) such that α A a where α is an unmentioned qualifier. α A a can be embedded in a larger term, viz., α A a L βAb. For example, pot-delimited-by-pothood-located-in-the-floor- delimited-by-floorhood. A more complex term can be of the form Tv α A L -1 Tā γ D Tā γ -1L Tv β A b meaning the locatedhood-in-the-pot-delimited-by-potness-determining-the-locusness- in-floor-delimited-by-floorness.

4. Another relational operator is C or colocation (sāmānādhikaraṇya) is such that a C b when a and b L d or α C β when α and β L a. These two sentence- surrogates a C b and α C β are taken as particular, e.g., 'some Brahmins are wise'.

5. P or pervasion (vyāpti) is considered to be the most important relational operator which is directly related to the process of inference and hence plays a significant role in determing the characteristic features of the consequence relation a la Navya-Nyāya and in laying down the conditions of universal quantification. However, all accepted definitions of pervasion discussed by Gaṅgeśa are in terms of negation. So, we need to discuss first the constraints and rules concerning negation.

Negation :

Negation in Navya-Nyāya has been construed as term negation. Barring a few cases, sentential negation has always been transformed into term-negation or absence. There are mainly two types of absences – one denying the occurrence of something in a locus and in the other the identity between two negata are denied, i.e., their difference is highlighted. An absence, we have seen is always of something in some place in a specific relation, e.g., absence of pot on a table by the relation of contact. Here, the pot is the negatum or the counterpositive (henceforth CP), the table is the locus of the absence, the delimiting attribute or counterpositiveness is potness and the delimiting relation is contact, since the pot, when present, is on the table by the relation of contact. The relation between absence of pot and the table is known as the abhāvīya-svarūpa relation (henceforth AS). Symbolically expressed,

Absence of a = Na, CP(Na) = a, locus of = LC

The relation between the locus and CP = LCa R a

The delimiting relation of CPness of Na = LCa Ra = $\mathbf{R_1Na}$

The AS of Na = $\mathbf{R_2Na}$ = LC Na R Na.

LCa = LC Na by R_1Na, viz., LC_1 Na; L Na = LCa by R_2Na, viz., LC_2 Na.

Constraints on CP:

a A global or maximal property (kevalānvayī-dharma) which is present in every locus cannot be the CP of any absence because the locus of such absence is non-existent (aprasiddha).

b A purely fictitious property (alīka-dharma) is unnegatable because it does not exist in any locus.

Corollaries:

1. In LCa R a, if R is contact or inherence, a is always a positive entity, i.e., R_2Na cannot be identical with contact or inherence.
2. In LCa R a, if R is a temporal relation, a is always a non-eternal created entity.
3. LC_2 Na = N LC_1a

Explanation: Suppose, a = potness, then LCa = pot, LCa R a = inherence, LC_1 Na = pot, LC_2 Na = N LC_1 a = non-pot.. When LCa R Na is identity (a case of difference), LC_1 Na = potness, LC_2 Na = non-potness.

Laws governing Double Negation:

1. If LCNNa = LCa, when LCa is determined by R_1Na, then NNa = a (bhāvarūpa-samaniyata). This clause has been contested by Raghunātha who upheld that that there may be extensional equivalence between NNa and a, yet intensionally these two are to be distinguished. He, however, had no problem in admitting extensional equivalence between NNNa and Na.

2. If the first negation is a relational absence and the second one is difference, the NNa = a-ness. However, we must remember that Navya- Naiyāyika-s are not unanimous on this point.

These two are the most basic rules governing double negation. Navya-Nyāya logicians have shown great ingenuity in handling different varieties of double negation resulting from combination of different types of absences.

Rules for conjunction (samuccaya): $\dfrac{\alpha \wedge \beta La}{\alpha La \text{ and } \beta La}$

Rules for Disjunction

1. Rule for saṃśaya: $\dfrac{\alpha v \beta La}{\alpha La \text{ or } \beta La}$
2. Rule for anyataratva: $\dfrac{\alpha v \beta La}{\alpha La \text{ or } \beta La \text{ or } \alpha \wedge \beta La}$

Now we are in a position to go back to the relation of pervasion.The relational operator P holds between concrete as well as abstract properties. So both a P b and α P β are admissible terms in Navya-Nyāya. a P b holds if for any locus, LCNa $\rightarrow$ LCNb. In case of inference, pervasion holds between the *probans* and the *probandum* of the inference. An inference is sound if this pervasion relation holds. In fact, a place that contains the absence of fire (*probandum*) must be the locus of the absence of smoke (*probans*). Wherever α P β obtains, the corresponding sentence-surrogate is universally quantified. For example, Tv m P Tā h, i.e., mortality pervades humanity and so it is to be interpreted as 'all men are mortal'.

Navya-Nyāya logic has received various semantic interpretations in the hands of modern interpreters. Matilal (1998) had suggested a Boolean semantics for some fragments of Navya-Nyāya. Ganeri (2004) has offered a graph-theoretic semantice and Ganeri (2008) has developed a set-theoretic semantics.Without committing ourselves to any of these semantics, we

only point out that simple terms denote simple objects and complex terms denote complex objects. That is, the meaning of sentence-surrogates are not propositions but complex objects.

Now, we shall try to define the consequence relation in this logic of property-projection relying on the already given definitions of negation and pervasion. Their real concern had always been to select the right sort of projection-base and to frame appropriate rules for distinguishing between projectable and non-projectable properties[13]. Unlike Ganeri, while deriving the rules of negation we are not following the footsteps of Raghunātha, which is a minority view. So we retain all three rules of negation as proposed in Ganeri (2004). Besides, we are confining ourselves only to the propositional part of Navya-Nyāya Logic.

Let T α mean ' α is true'. Then,

R1 If NT α , then TN α [Rule for Absence]

R2 TNN α iff NTN α [Rule for Absence of absence]

R3 If TN α , then NT α [Exclusion Principle]

R3 might have created problems in case of partially locatable properties (avāpyavṛtti-dharma), hadn't there been the operator of delimitation. By delimiting the loci of partially locatable properties, simultaneous predication of α and N α in respect of the same thing can be avoided. As Matilal (1998) writes, 'Thus a device is used to reparse the partially locatable property as wholly locatable, so that the standard notion of negation is not "mutilated" in this system.' [14]

We have already mentioned that the soundness of an inference, according to Navya-Nyāya, depends on the relation of pervasion and the relation of pervasion, again, depends at least on satisfying the three conditions of pakṣa-sattva, sapakṣa-sattva and vipakṣa-asattva. Let us, therefore, formulate the properties of the consequence relation ' $\models$ ' as follows.

Let α be 'the probans m is located in x' and β be 'the probandum p is located in x', then

$\alpha \models \beta$ iff p pervades m.

The rules of pervasion then warrants the following:

[13] Sarkar (1997)
[14] P.144

$\alpha \models \beta$ iff Tα

iff for any assignment of values to x, T$\beta \rightarrow$ Tα

iff for any assignment of values to x, TN$\beta \rightarrow$ TNα

However, we must remember that ' $\models$ ' holds under the above-mentioned three conditions only in one particular world, i.e., in a specific model.

References

[Annaṃbhaṭṭa 1976] Annaṃbhaṭṭa (1976) Tarkasaṃgraha with Tarkasaṃgrahadīpikā, ed.&tr. Gopinath Bhattacharya, Progressive Publishers, Calcutta.

[Bhattacharya 1952] Bhattacharya Dinesh Chandra (1952) Vāngālīr Sārasvata Sādhanā:Vange Navya-Nyāya Carcā (in Bengali), Vangīya Sāhitya Parishad, Calcutta.

[Bhattacharyya 1958] Bhattacharya Dinesh Chandra (1958) History of Navya-Nyāya in Mithila, Mithila Research Institute, Darbhanga.

[Bhattacharyya 1978] Bhattacharya Gopikamohan (1978) Navya-Nyāya: Some Logical Problems in Historical Perspectives, Bharatiya Vidya Prakashan, Benaras.

[Bhattacharyya 1987] Bhattacharyya Sibajiban (1987) Doubt, Belief and Knowledge, Indian Council of Philosophical Research, New Delhi.

[Bhattacharyya 2009] Bhattacharyya Sibajiban (2009) 'An Introduction to Navya-Nyāya Logic' in 'Indian Logic', eds.J.N. Mohanty and Amita Chatterjee, The Development of Modern Logic,ed. Leila Haaparanta, Oxford University Press, Oxford.

[Das 2006] Das Ramesh Chandra (2006) Navya-Nyāya-bhāsāpradīpa of Maheśa Candra Nyāyaratna, Department of Special Assistance in Philosophy, Utkal University, Bhubaneswar.

[Ganeri 2004] Ganeri Jonardon (2004) 'Indian Logic', Handbook of History of Logic, Volume I, eds., Gabbay and Woods, Elsevier BV.

[Ganeri 2008] Ganeri Jonardon (2008) 'Towards a formal regimentation of the Navya-Nyāya technical language I & II', Logic, Navya-Nyāya & Applications: Homage to Bimal Krishna Matilal,Studies in Logic,

volume 15, eds. Mihir K. Chakraborti et al, College Publications, U.K. pp. 105-138.

[Upādhyāya 1892] Gaṅgeśa Upādhyāya (1892) The Tattvacintāmaṇi, Part I & II with the sub-commentaries of Mathurānātha Tarkavāgīśa, ed. Kāmākhyānātha Tarkavāgīśa, Bibliotheca Indica, Volume 98, Calcutta; Tattvacināmaṇi-Dīdhiti-Jāgadīśi and Gādādharī, Chowkhamba Sanskrit Series, Benaras.

[Guha 1968] Guha Dinesh Chandra (1968) Navya-Nyāya System of Logic, Bharatiya Vidya Prakashan, Benares.

[Ghosh 2010] Ghosh Partha (2010), Ed. Materialism and Immaterialism in India and the West, Center for Studies in Civilisations, New Delhi.

[Goble 2001] Goble Lou (2001), Ed. The Blackwell Guide to Philosophical Logic, Blackwell Publishers Limited, Oxford,

[Goekoop 1967] Goekoop C. (1967) The Logic of Invariable Concomitance in the Tattvacintāmaṇi, D. Reidel & Co., Dordrecht.

[Ingalls 1951] Ingalls, Daniel H.H. (1951) Materials for the Study of Navya-Nyāya Logic, Harvard University Press, Cambridge, MA.

[Jha 1984] Jha, Ganganatha. (1984) Trans. The Nyāyasūtra of Gautama (with the commentaries of Vātsyāyana, Uddyotakara, Vācaspati Miśra and Udayanācārya), 4 vols. Motilal Banarassidass Publishers Private Limited, Delhi.

[Matilal 1968] Matilal B.K. (1968) The Navya-Nyāya Doctrine of Negation, Harvard University Press, Cambride. MA, 1968

[Matilal 1986] Matilal B.K. (1986) Perception, Clarendon Press, Oxford.

[Matilal 1998] Matilal B.K. (1998) The Character of Logic In India, ed. J. Ganeri and H. Tiwari, SUNY, Albany.

[Nyāyaratna 1973] Nyāyaratna Maheśa Chandra (1973) Navyanyāyabhāṣāpradīpa, edited by Kalipada Tarkacarya, Sanskrit College, Calcutta, 1973.

[Mohanty 1989] Mohanty J.N. (1989) Gaṅgeśa's Theory of Truth, Motilal Banarasidass Publishers Private Limited, Delhi.

[Potter 1977] Potter K. H. (1977) Ed. Encyclopedia of Indian Philosophy, vol. 2, Nyāya-Vaiśeṣika, Motilal Banarasidass Publishers Private Limited, Delhi.

[Potter 1993] Potter K.H. and Bhattacharya, Sibajiban. (1993) Encyclopedia of Indian Philosophy, vol. 6, Navya-Nyāya, Motilal Banarasidass Publishers Private Limited, Delhi.

[Sarkar 1997] Sarkar T.K. (1997) 'Jaina Logic in Perspective', Essays in Indian Philosophy, ed. S. Saha, Allied Publishers Limited, Calcutta.

[Staal 1962] Staal J.F. (1962) 'Negation and the law of contradiction in Indian thought', Bulletin of the School of Oriental and African studies, 25.

[Thakur 2003] Thakur, Anantalal. (2003) Origin and Development of the Vaiśeṣika System, Centre for Studies in Civilizations, New Delhi, 2003.

[Visvanātha 1988] Visvanātha (1988) Kārikāvali-Siddhāntamuktāvalī, ed. Shri Shankar Ram Shastri, Chowkhamba Sanskrit Pratisthan, Delhi.

A Brief History of Chinese Logic

FENRONG LIU * and WUJING YANG †

1 Introduction

Chinese logic was born in the 6th to 3rd centuries B.C., an era of great cultural and intellectual expansion in Chinese history. The period was well-known for its various schools that held different thoughts and ideas, competing freely with each other, the so-called "contention of a hundred schools". The situation was described by the famous historian Sima Tan (died in 110 B.C.) in his book *On the Main Ideas of the Six Schools* (*lun liujia zhaozhi*, 论六家要旨), in which the six schools and their ideas were first presented and summarized. They are the Schools of Yin-Yang, Confucianism, Moism, Names, Legalism and Taoism. Four more schools were added later by Ban Gu (32–92 A.D.) in his book *The History of the Former Han Dynasty* (*hanshu*, 汉书), viz. the Schools of Agriculture, Diplomacy, "Minor-Talks", and the Miscellaneous School. One can imagine from these names how schools interacted with each other, while at the same time developing their own theories. Among them, according to Han Feizi (280–233 B.C.), Confucianism and Moism were the most popular and dominant ones.

Logical themes occur in many philosophical works in Ancient China, such as the oldest text *the Book of Changes* (*yijing*, 易经), *the Art of War* (*sunzi bingfa*, 孙子兵法), and works by the Confucians. But perhaps the greatest relevance and significance to logic is found in the School of Moism and the School of Names. The former is famous for its contributions to argumentation theory, *Bianxue* in Chinese. And the latter made fundamental contributions to the theory of names, *Mingxue* in Chinese.[1] Scholars of the Confucian School also proposed their own theories of names. *Mingbainxue* is a combination of these two theories, and it is considered to be Chinese logic.

*Department of Philosophy, School of Humanities and Social Sciences,Tsinghua University, Beijing, China.

†Department of Philosophy, Renming University of China, Beijing.

[1]"Bian" in Chinese means "argumentation", "Xue" studies, and "Ming" "names".

The School of Moism was founded by Master Mozi (墨子), who lived during the fifth century B.C.. Mozi was the first to challenge Confucianism by making reasoning the core of intellectual inquiry. As opposed to Confucian view that one should follow the fixed meaning of names and act on it, the Moists emphasize that one should define notions according to the actual situation. They are also in favor of approaching truth by argumentation: the term *Bianxue* reflects this point. The term Mozi is also used to refer to all works written by anonymous members of the Moist school. These texts cover a great variety of topics: epistemology, geometry, optics, economics, and so on.[2] Among them, there are six chapters of special logical interest, *Jing Shang* (经上), *Jing Xia* (经下), *Jing Shuo Shang* (经说上), *Jing Shuo Xia* (经说下), *Daqu* (大取) and *Xiaoqu* (小取). The collection of these six texts is usually called *The Moist Canons* ('*The Canons*', for simplicity). *Jing Shuo Shang* is an explanation to *Jing Shang*, the same with *Jing Shuo Xia* and *Jing Xia*. It is believed that *Daqu* was devoted to ethical issues, though there are major textual difficulties in understanding it. In this regard, *Xiaoqu* is much less problematic. It contains lots of logical topics, coherent and well-structured. We will introduce these topics soon in this paper.

The School of Names was founded by Deng Xi (560–501 B.C.), and both Hui Shi and Gongsun Long belong to this school.[3] Literally, this school is known for its theory of names. In particular, they had the following view of the relationship between names (*ming*, 名) and objects (*shi*, 实). Names are used to denote objects, so they must conform to the objects. If objects have changed, names must change too. Moreover, names cannot exist without referring to some objects, but objects can exist without necessarily having names. Similar to the sophists in the ancient Greece, this school was also famous for proposing strange propositions or paradoxes. For instance, Gongsun Long was famous for his statement and argument for "A white horse is not horse". His main point is that the notion of white horse comes from something white which describes color, and horse which describes shape, which is not the same as the notion which only describes shape. There are also other famous paradoxical propositions, for instance, "chicken have three feet" and "eggs have feathers". For "chicken have three feet", they claim that in addition to the left and right foot, there is an independent notion of 'chicken foot', so there are in fact three feet to

[2]For a new attempt of re-translation of Mozi, see I. Johnston, *The Mozi. A Complete Translation*, Hong Kong: The Chinese University Press, 2010.

[3]For a general introduction to this school of thought, see the Item "School of Names" in the *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/school-names/

a chicken. Also, common sense tells us that eggs have no feathers, but since one can get chicken from eggs, one can say that eggs have potential feathers. All these examples somehow go against our ordinary intuitions, so they triggered hot debates among different schools. But that is precisely why the School of Names used them, to expose problems in people's thinking and language use.

The School of Confucianism was founded by Confucius (551–479 B.C.). Its guiding idea is that 'rectification of names' is the first thing needed to achieve a harmonious society. They believe that chaos in real life is caused by wrong usage of names. Here is a famous conversation between Confucius and his student Zilu:

> "If names are not correct, language is not in accordance with the truth of things. If language is not in accordance with the truth of things, affairs cannot be carried on to success. When affairs cannot be carried on to success, proprieties and music will not flourish. When proprieties and music do not flourish, punishments will not be properly awarded. When punishments are not properly awarded, the people do not know how to move their hand or foot. Therefore a superior man (junzi) considers it necessary that the names he uses may be spoken appropriately and also that what he speaks may be carried out appropriately. What the superior man requires is just that in his words there may be nothing incorrect."[4]

Confucians hold that clearly identifying the intention and extension of a name is the precondition of correct acting. Once the system of names is fixed, the society and people know what is allowed to do and what is forbidden. "There is government, when the prince is prince, and the minister is minister; when the father is father, and the son is son."[5] Xunzi (313–238 B.C.), a follower of Confucius, was one of the early leaders in the consolidation of what came to be thought of as the Confucian tradition. Xunzi is also used to refer to his collected works, which address many topics ranging from economic and military policy to philosophy of language. In this paper, we will look especially at his theory of names.

The aim of this paper is to explain logic issues that were discussed by the above schools in the earlier period. In addition, we show how some of

[4] See the chapter Zilu from *The Analects* (*lunyu* 论语). The translation is from the online Chinese Text Project: http://ctext.org/analects.

[5] *Sishu zhangju jizhu* 四书章句集注, xinbian zhuzi jicheng, Beijing: Zhonghua Shuju, 1983.

those ideas developed over time in Chinese history. The paper is organized as follows: In Section 2, we start with the main theories of *Mingbianxue* in Pre-Qin Period. From the Han Dynasty onward, the School of Confucianism became dominant. We will see how logic issues were taken up and developed in later dynasties in Section 3, with a focus on contributions made in the Han, Wei-Jin and Song dynasties. In Section 4, we move to the 20th century and discuss some recent developments, highlighting comparative studies of Chinese logic, Indian and Western logic. Finally, we draw some conclusions.

2 Mingbianxue

2.1 Names

Names come together with objects, and relations between names and objects were a main concern for many schools. In the *Canons*, it says that

> "所以谓，名也。所谓，实也。名实耦，合也。" (A81)

What something is called by is its 'name'. What is so called is the 'object'. The mating of 'name' and 'object' is 'relating'.[6] Similar ideas were expressed in "*Yiming jushi*"(以名举实), [7] which says "one uses names to refer to objects". To give a name to some object, there are two basic things to consider, namely, "like" (*ruo*, 若) and "so" (*ran*, 然). These two things determine a "standard", called *fa* 法, namely, "that in being like which something is so".[8] So, in order to use one name consistently, we must follow *fa*. E.g., the name of "circle" can be applied to the compass, too, since it fits the same standard.[9]

Names can be of different kinds, as stated in "There are three kinds of names: unrestricted names, classifying names and private names" (*ming. da, lei, si.* 名。达，类，私。).[10] For instance, "thing" is "unrestricted", as any object necessarily requires this name. "Horse" is a "classifying" name, for anything that is "like the object" we necessarily use this name.

[6]NO 11. In this paper we follow Graham's numbering of the Canons. He made a hybrid text from Xiaoqu and parts of Daqu under the title "Names and Objects"(abbreviated "NO") and most of the remainder of Daqu as "Expounding the Canons"("EC", for short). "TC" and "HC" abbreviate Daqu and Xiaoqu, respectively. We will make some revisions of Graham's translation where necessary.

[7]NO 6 (HC 6A/9-6B/1).

[8]A 70.

[9]Ibid.

[10]A 78.

The name "Jack" is a "private" name, since the name stays confined to this single object.

Similarly, Xunzi divided names into four kinds in his chapter "rectification of names" in *Xunzi*: *Da gongming* (大共名), *Gongming* (共名), *Bieming* (别名) and *Da bieming* (大别名). Da gongming are names with the biggest extension, for instance, "human being" or "thing". Gongming are names with smaller extensions than those of Da gongming. Da bieming are the names with the smallest extension, for instance, the name "Confucius", or "Jack". Bieming are names with larger extension than those of Da bieming. In fact, Da gongming are similar to unrestricted names, and Da bieming are similar to private names in the Canons.

Both classifications of names presented here are made from an extensional point of view. Classifying names, Gongming or Bieming are what we would nowadays call *generic names*, while private names or Da bieming are simply *proper names*.

Most importantly, the Moists proposed principles regarding the distinction between any two classifying names. In fact, this follows from the notion of "standard". They say that proposing a "standard" is not arbitrary; we have to pick those properties which one object has and the other lacks. This view on the correct use of classifying predicates is elaborated below:

> "By referring arbitrarily one cannot know differences. Explained by: what they have. Although oxen are different from horses, it is inadmissible to use oxen having incisors and horses having tails as proof that oxen are not horses; these are things which they both have, not things which one has and the other lacks."[11]

To distinguish oxen from horses, having incisors and having tails are not the properties one should take. We will come back to this issue when discussing kind-based reasoning in Section 2.4. This is similar to "genus and differentia" as proposed by Aristotle. The genus is the kind under which the species falls, and the differentia states what characterizes the species within that genus. It is species that have essences which should be the basis of a correct definition. The notion of essence is similar to *fa* in Moist logic. Oxen and horses belong to the same kind "animal", and one should find a *fa* for each species that differentiates it from others within the same kind. We can fairly say that the theory of classifying names in

[11] B 66.

Ancient China had the same spirit as Aristotle's account of "genus and differentia".[12]

Concerning the relationship between names and objects, the Moists observe that different names can be used for the same kind of objects, and different objects can share the same name. An example for the former is that a kind of dogs has two different names, *quan* (犬) and *gou* (狗), they are both names for dogs. For the latter, according to a given standard, objects sharing the same name are not necessarily alike except in the respects covered by the standard. For example, pieces of stone and of wood, both of which fit the standard for "square", share the name "square", but are very different otherwise. Thus, the Moists were aware of the complex relations between syntactic names and semantical objects.

2.2 Kind (Lei): Tong and Yi

Kind (*Lei*) is one of the core concepts in the *Moist Canons*. As we have seen, classifying names are supposed to apply to kinds. We discussed not only rules of correctly using classifying names, but also principles governing the distinction between any two classifying names. Recall our example: "horse" is a "classifying" name, for anything "like the object" we necessarily use this name. More generally, the following view underlies the Moist account:

(a) For each kind, there are objects which belong to it, and in virtue of this, they are similar or the same.

(b) With each kind, there are some properties which are essential, being common to all the objects of that kind.

The similarity between objects of the same kind is called "the sameness of the kind" (*leitong*, 类同). Many different sorts of similarity are discussed in the *Canons*, but the sameness of the kind is clearly distinguished from other sorts, witness the text below:

> "*Tong (same)*. Identical, as unites, as together, of a kind. There being two names but one object is the sameness of 'identity'.

[12] A more concrete comparison can be found in Zhang Jialong and Liu Fenrong 2007, "Some thoughts on Mohist logic". In J. van Benthem, S. Ju and F. Veltman, eds., *A Meeting of the Minds–Proceedings of the Workshop on Logic, Rationality and Interaction*, Beijing, 2007.

> Not being outside the total is sameness 'as units'. Both occupying the room is the sameness of being 'together'. Being the same in some respect is sameness in being 'of a kind'."[13]

Objects of the same kind have common essential properties, called "*leitong*". In addition, the *Canons* also discusses difference in kinds. Again, there are many sorts of difference, but "difference in kinds (*leiyi*, 类异)" is the one that is relevant to our discussion here. It says:

> "*Yi (different)*. Two, not units, not together, not of a kind. The objects if the names are two necessarily being different is being 'two'. Not connected or attached is 'not units'. Not in the same place is 'not together'. Not the same in a certain respect is 'not of a kind'."[14]

Thus, what matters to a kind are its essential properties. They are the criterion by which a kind is identified. Moreover, according to these properties, we can determine whether an object is of that kind or not. It will become clear how this view is exploited concretely when we turn to the logical study of reasoning patterns. Lei (Kind), as the core notion in Moist logic, serves as a basis for much logical reasoning. Together with "reason (*gu* 故)" and "general law(*li* 理)", it forms the three basic components of a piece of reasoning.

2.3 Propositions and Logical Constants

As expressed in the Moist dictum "*yichi shuyi* (以辞抒意)"[15], propositions are used to elucidate ideas.[16] We express our ideas by means of various types of proposition. We now turn to the structure of logically complex propositions in Moist logic. The *Canons* discussed different types of proposition involving logical constructions like quantifiers, conditionals and modalities. Since there is no systematic categorical classification of propositions in the *Canons*, in what follows we are going to review it from a modern perspective. Along the way, we will pay attention to how Moists perceived differences between the propositional types, and especially, how they use logical constants to indicate these types in the language.

[13] A 86.
[14] A 87.
[15] NO 11.
[16] The translation by Graham was "propositions are used to dredge out ideas".

Quantifiers: all and some The universal quantifier is mainly expressed by the word "all (*jin*, 尽)". As explained in "all is none not being so." (*jin, moburan ye*. 尽，莫不然也。)[17] Written in a logical formula, it is $\forall x = \neg\exists\neg x$. Notice that here *jin* is defined in terms of "none (*mo*, 莫)" which is taken as a primitive; and thus, the universal quantifier is defined by a double negation. Besides *jin*, other words, like *ju* 俱, *zhou* 周, *ying* 盈, *bian* 遍, are also used to express the universal quantifier, they all mean "all". One can easily find propositions containing such words in the *Canons*. The negation of the universal quantifier is defined as well: in "some is not all." (*huoyezhe, bujin ye*. 或也者，不尽也。)[18] Put again in a logical formula, we get $\exists x = \neg\forall x$. This is not really what existential quantifiers mean nowadays ($\exists x = \neg\forall\neg x$.). Probably the Moist text is not a definition, but it wants to make the point that "some" (viewed as a part) differs from "all" (as the whole).

Disjunctions, conjunctions and conditionals The expressions "either... or... (*huo... huo...* 或...或...)" are used to express disjunction in the *Canons*. Of the many examples in the texts, we only give the following two for the purpose of illustration: "either call it ox, or call it non-ox"[19], and "either its body is gone or it is still here."[20]

Concerning conjunctions, there is no clear independent expression for this in the language, but the *Canons* have many propositions which express the idea that several things should hold at the same time. Probably, juxtaposition was seen as implicit conjunction.[21]

The conditional is defined in "the loan-named is not now so" (*jiazhe, jin buran ye.* 假者今不然也。[22]). Conditions or causes that lead to some phenomenon are called "reason (*gu*, 故)". There are two types of *gu*, "major reason (*da gu*, 大故)" and "minor reason (*xiao gu*, 小故)". The distinction between them is illustrated in the following text:

> "The gu of something is what it must get before it will come out. Minor reason: having this, it will not necessarily be so; lacking this, necessarily it will not be so. It is the unit, like

[17] A 43.
[18] NO 5 HC 6B/3-5.
[19] A 74.
[20] A 46.
[21] It has been proposed that 盈 (which really means "all") can be seen as a conjunction sign. See Zhou Yunzhi. *History of Chinese Logic*. Shanxi Education Press, 2002.
[22] NO 5 HC 6B/3-5.

> having a starting-point. Major reason: having this, it will necessarily be so; lacking this, necessarily it will not be so. Like the appearing bringing about the seeing."[23]

According to this explanation, major reason and minor reason are what we would nowadays call "sufficient and necessary condition" and "necessary condition", respectively. In the Moist texts, "if...then..." (*ruo...ze*, 若...则...) are often used to express conditionals[24].

Modalities Interestingly, modalities are considered in the *Canons* too. First, the word "*bi* 必" is used to express necessity. For instance, "There necessarily exists a winner in a disputation." (*wei bian wusheng, bi budang.* 谓辩无胜，必不当。[25]) Several tensed modalities are also considered. The word *qie* 且 is used to express the future tense. E.g., in "Going out in the future is not going out now." (*qie chumen, fei chumen ye*. 且出门，非出门也。), we can see two states of going out in the future or going out now clearly distinguished. Likewise, "*yi*已" is used to denote the past tense. As is clear from these examples, modalities are explicitly recognized as such in the text of the *Canons*.

Complex propositions One striking phenomenon is that the Canons are replete with complex propositions such as "riding a white horse is riding a horse", "killing a thief is not killing a man", etc. These are not simply constructed from basic propositions by means of the logical constants we have seen so far. In addition, they have rich variations in complex predicates. To conclude, we mention one logical issue here relevant to complex propositions, namely, extension of predicates, which is the basis of all correct reasoning with complex propositions[26]. To illustrate this, consider the following example from the *Canons*:

> "He loves people" requires him to love all people without exception, only then is he deemed to love people. "He does not love people" does not require that he loves no people at all; he does not love all without exception, and by this criterion is deemed not to love people. ... These are cases in which

[23] A 1.

[24] Sometimes "*ruo*" is omitted when it is clear from the context.

[25] B 35.

[26] See an analysis in Liu Fenrong and Zhang Jialong 2010, "New perspectives on Moist logic", *Journal of Chinese Philosophy*, to appear.

> something "applies without exception in one case but not in the other".[27]

This text shows that studying the correct application of certain predicates is exactly to spell out their extensions. This is a crucial topic in both classical and modern logic.

Remark The diversity of propositions considered by the Moists is marked by different indicators in the language, e.g. "*huo... huo...*" for disjunctions. The clear identification of these structuring expressions suggests that the Moists realized the distinction between logical and non-logical expressions. The former are the protagonists of modern logic, and they determine logical structures in general. Consider two examples we have seen: "*huo weizhiniu, huo weizhi feiniu.* (或谓之牛，或谓之非牛)", and "*qiti huoqu huocun* (其体或去或存)". They have the same logical form "*huo..., huo...*", but are about different subject matters.

We hope to have shown that the exploration of the meanings of logical constants by the Moists was innovative. In terms of related traditions, one might say that Moist logic seems closer in spirit to Stoic logic than to Aristotle's syllogistic.

2.4 Kind-based Inference Patterns

It was commonly held across different schools that one should infer from what one knows to what one does not know, to get new knowledge. The general term to denote this process is "*shuo* (说)", reasoning or providing proofs, as in "by means of inference bring out reasons" (*yishuo chugu*, 以说出故[28]). First, we would like to briefly address the sources of knowledge discussed in the *Canons*. It is said that "There are three different ways to get knowledge: viz. learning from others, reasoning from what one knows already, and consulting one's own experience." (*zhi: wen, shuo, qin.* 知，闻，说，亲。)[29] This clearly identifies the different ways of getting knowledge and besides, a nice example is also given to show these different sources plus their interplay. It goes as follows:

> Imagine that someone, say Jack, is standing outside of a room, and he sees an object which is white. From the very beginning

[27] NO 17.
[28] NO 11.
[29] A 80.

> then, he knows from his own observation that "the object outside of the room is white". But now, there is another object, inside the room, of a yet unknown color. Now Jack is told that the object in the room has the same color as the one outside. Now he knows that "the object in the room has the same color as the one outside", by learning from others. Finally he also knows that "the object in the room is white", via his own reasoning based on what he knows. This example illustrates exactly how *shuo* works for us when we acquire knowledge.

Now we get to a theme that modern logicians will recognize as being closest to their subject. To get to know something by means of *shuo*, we can appeal to many different kinds of reasoning. The remainder of this section is about reasoning patterns in the Moist texts. Our focus are the characteristics of these patterns and their validity. We will start with a simple pattern called *Xiao*, as explained in the following text from *Xiaoqu*:

> "The *xiao* consists of setting up the *fa* (standard). That which things are modeled after is that which is to be set up as the *fa*. When it conforms to the *xiao*, it is right. When it does not conform to the *xiao*, it is wrong."[30]

The name used for the reasoning here is called *Xiao* (效) which means "to imitate". The above text explained how the reasoning of *Xiao* is carried out. First, a general standard *fa* (法) is set up, which gives us general principles to follow in the inference. Next, according to the standard, we infer whether specific things conform to this standard or not. Thus, this reasoning goes from a general rule or standard to specific cases. It is similar to the following example which we are all familiar with: "All human beings are mortal, Socrates is a human being, so Socrates is mortal". In this example, the standard is "All human beings are mortal", and we infer that a specific human being Socrates conforms to this standard. In this sense, *Xiao* can be thought of as deduction.

We now continue with a few further central patterns in the *Xiaoqu*: "Illustrating (*pi*, 辟)", "Adducing (*yuan*, 援)", and "Inferring (tui, 推)". There is also a pattern of "parallelizing (*mou*, 侔)", but we do not address this complex issue in this paper.[31] We explain the reasoning by concrete examples, and then try to analyze it in terms of logical rules.

[30] NO 5.

[31] For a recent study on Moist reasoning with complex propositions, we refer to Liu and Zhang 2010, "New perspectives on Moist logic," *Journal of Chinese Philosophy*, to appear.

Illustrating (pi) "Illustrating is that, in order to make someone else know one thing, you refer to a different thing known by him already." (*piyezhe, jutawu yi mingzhi ye*. 辟也者，举他物而以明之也。)[32]

This pattern of reasoning is found in works much earlier than the Canons, like the *Book of Odes* around 1000 B.C.. The well-known sophist Hui Shi (380–305 B.C.) was famous for his talent in using this sort of reasoning in his arguments. The feature of illustrating is that two different things A and A' are used in the reasoning. The reason why one can get to know A by appealing to a different A' lies in the similarity between A and A', as introduced in the above.

To be more specific, consider an example from the book *Gongshu* (公輸) of Mozi. Mozi met the King of the State Chu. In order to convince the King that it is not right for the rich Chu to invade the poor State Song, he used a more obvious example. Namely, it is not right for rich people to leave their property behind and go robbing poor people. Since the King sees the injustice of the latter, he realizes that of the former, too. Clearly, in this example, (a) the wealthy State invading the poor State, and (a') the rich people robbing the poor, are of the same kind. As it is easy to see the injustice of (a'), one can then infer the injustice of (a) too. Notice that the purpose of illustrating is to make someone else know, not to make oneself know. In this sense, it is more like the process of explanation– and as such, it is a typical illustration of the interactive argumentative slant of the *Canons*.

Adducing (yuan) "Adducing means: if it is so in your case, why may it not be so in mine too?" (*yuanyezhe, yue: ziran woxidu bukeyiran ye*? 援也者，曰：子然我奚独不可以然也?)[33] Adducing is carried out in the following steps: one first quotes an opinion that the opponent accepts, then one argues that the opponent's opinion and one's own are the same or belong to the same kind. Then, it naturally follows that one's opinion should be accepted as well, if the opponent insists on his opinion. We mentioned one example in Section 2, when Gongsun Long defended his thesis "a white horse is not a horse". The argument used there is "Adducing". He asked why it would be a problem for him to say "a white horse is not a horse" if we accepted what Confucius said: "Chu's man is not a man".

Again, the basis of adducing is the earlier-discussed notion of kind. In the above example, "Chu's man is not a man" and "a white horse is not a

[32]NO 11.
[33]Ibid.

horse" are of the same kind, and so, if one of them true, the other should be true as well. In fact, both illustrating and adducing can be formalized into the following schema

(i) Object or statement A and A' are of the same kind (i.e. A has the kind-defining property P iff A' has that property P),

(ii) A has the property P,

(iii) Therefore, A' has the property P.

As we can see from the above examples, premise (i) is often omitted from the reasoning as being common knowledge. What is left then is a transition from some property of one object to another object or statement that is of the same kind. But of course, establishing the sameness in kind is an essential feature in practice.

Inferring (tui) "Inferring is using what is the same in that which he refuses to accept and that which he does accept in order to propose the former." (*tuiyezhe, yiqi suobuqu zhitongyu qisuoquzhe, yuzhiye.* 推也者，以其所不取之同于其所取者，予之也。)[34]

Consider the following scenario. If someone proposes a statement you disagree with, what you need to do is choose a statement that belongs to the same kind as what he proposed (and which he should therefore accept), but in fact he cannot accept it. In that case, he has to give up his initial statement. This describes precisely how inferring proceeds. Let us look at an example in the book *Gongmeng* (公孟) of *Mozi*. Gongmengzi does not think gods or ghosts exists, but nevertheless, he claims that *junzi* (君子) should learn how to pray. Mozi then says: "What Gongmengzi said is just like saying you have to learn how to treat your guests well, but there is no guest at all. This is also like having to make a fish net, but there is no fish." The absurdity of the last two cases is clear, so we conclude that what Gongmengzi said was wrong.

In this example, what Gongmengzi said about gods and what Mozi said about guests or fish are of the same kind. Clearly, Gongmengzi would not agree with the statement about guests and fish, so his statement about gods can also be rejected. The logical reasoning pattern here is this:

(i) Object or statement A and A' are of the same kind (i.e. A has property P iff A' has property P),

[34] Ibid.

(ii) A' does not have the property P,

(iii) Therefore, A does not have the property P.

This refutational style of reasoning is very common in practice if one wants to reject some statement proposed by others.

So far, we have seen that in illustrating, adducing and inferring, by comparing two objects or statements of the same kind, we infer that one has (or lacks) some property from the fact that the other object has (or lacks) that property. This sort of reasoning is often called "analogical inference" (*leibi tuili*, 类比推理). Kind-based analogical inference is the main reasoning pattern in Chinese logic, it was used widely in philosophical argumentation. We will see in Section 3 how it was developed further by later scholars. In the Western logical tradition, analogical reasoning is considered different from deductive and inductive inference. But for Chinese logic, the situation with analogical inference is more complex. Its view of reasoning patterns contains both deductive and inductive reasoning. In terms of applications, some patterns, e.g. Abducing and Inferring, are more often seen in refusing some claims, while Illustrating and Paralleling are used to infer positive conclusions. Also, what Chinese logic considers fundamental behind all reasoning is the notion of 'kind': all inferences are based on sameness and difference in kinds.

2.5 Argumentation

Both Xunzi and Mozi emphasized that to distinguish truth from falsehood, besides considering sameness and difference in kinds, we should also provide sufficient arguments. Xunzi says, "辨则尽故, *bianzejingu*"[35], in argumentation one should list all the reasons, and "以说出故, *yishuo chugu*"[36], by means of inference bring out reasons. As we have seen in Section 2.3, these include "major reason (*da gu*, 大故)" and "minor reason (*xiao gu*, 小故)".

Just as in Ancient Greece or Rome, disputation was popular during the Warring States period. The different schools criticized each other, trying to convince their King with new proposals. The Moists were not only concerned with this practice of disputation, but also with its meta-theory. We can find many illuminating discussions of this topic in the Canons. For instance, here is how they define a disputation: "disputation means

[35] Cf. the chapter of *Zhengming* in *Xunzi*.

[36] NO 6.

contending over claims which are the contradictory of one another" (*bian, zhengbi ye*. 辩，争彼也。)[37] To show what such contradictory claims are, one simple example is:

"One calling it 'ox'(P) and the other 'non-ox'(¬P) is contending over claims which are contradictories of each other" (*weizhiniu, huo weizhifeiniu, shi zhengbi ye*. 谓之牛，或谓之非牛，是争彼也。)[38]

Furthermore, the *Canons* propose basic principles regulating disputations. The first says that of two contradictory propositions, one must be false, they cannot be true at the same time. (*shi bujudang, bujudang bi huo budang*. 是不俱当，不俱当必或不当。)[39] This is clearly the logical Law of Non-Contradiction. Next, the *Canons* say that two contradictory propositions cannot be both false, one of them must be true (*weibian wusheng, bi budang, shuo zai bian*. 谓辩无胜，必不当，说在辩。)[40] This, of course, is the Law of Excluded Middle. There seems to be a consensus nowadays that the Moists explicitly proposed these two basic logical laws, though there are dissenting views.[41] Interestingly, it is the discourse function of logical laws, rather than their theoretical function, that is emphasized by the Moists.

The Moists also discussed the broader purpose of disputation in general. We conclude by citing their comprehensive and yet highly concise description in the following text:

> "The purpose of disputation is (1) by clarifying the portions of "is-this" and "is-not", to inquire into the principle of order and misrule; (2) by clarifying points of sameness and difference, to discern the patterns of names and of objects; (3) by settling the beneficial and the harmful, to resolve confusions and doubts. Only after that, one may by description summarize what is so of the myriad things, and by asserting seek out comparables in the multitude of sayings."[42]

Passages like this from the founding period of logic are intriguing, as modern logicians are becoming more interested in regaining argumentative

[37] A 74.
[38] Ibid.
[39] Ibid.
[40] B 35.
[41] D. Leslie. Argument by contradiction in pre-Buddhist Chinese reasoning, Australian National University, Canberra, 1964.
[42] NO 6 (HC 6A/9-6B/1).

multi-agent perspectives on logic, in addition to the dominant paradigm of reasoning as single-agent mathematical proof.[43]

2.6 Paradox

Finally, we mention one more striking analogy between Moist Logic and its counterparts elsewhere, in the form of two illustrations. Many paradoxes are discussed in the *Canons* - and these, of course, almost seem a hallmark of the profession of logic. This interest in paradoxes may lie in its direct connection to the earlier central concern with disputations, where one has to avoid being self-contradictory. Let us start with the first example, which is stated below:

> "To claim that all saying contradicts itself is self-contradictory. Explained by: what he says himself." (*yi yan wei jinbei, bei. shuo zai qiyan*.以言为尽悖，悖。说在其言。)[44]

Here is the implicit argument. Assume that "all saying contradicts", then the sentence "all saying contradicts" is false itself. What this means is that some statements are not contradictory. Thus, the Moists were aware of the phenomenon of self-reference, and its logical consequence of self-refuting statements. Clearly, this example is close to the paradox ascribed to the Cretan philosopher Epimenides in the sixth century B.C., who asserted that "Cretans are always liars." While this is not quite the famous Liar Paradox, which is contradictory whichever way one looks at it, it comes close.

We conclude with a second Moist paradox, which seems original without an obvious Western counterpart. It says:

> "That it is useful to learn. Explained by: the objector." (*xuezhiyi ye, shuo zai feizhe*. 学之益也，说在诽者。)[45]

This paradox seems to mix self-reference with pragmatics of speech acts.

Paradoxes have contributed greatly to the progress of logic. In this respect, too, the Moist logicians were on to something crucial, at the same time as their counterparts worldwide.

---

[43]Cf. R.Stalnaker. "Knowledge, belief and counterfactual reasoning in games," *Economics and Philosophy*, 12(2):133-16, 1996, R.H. Johnson, H.J. Ohlbach, Dov M. Gabbay, and J. Woods, editors. *Handbook of the Logic of Argument and Inference: The Turn Toward the Practical*. Amsterdam: North-Holland, 2002. J. van Benthem. *Logic Dynamics of Information and Interaction*. Cambridge University Press, 2010.

[44]B 71.

[45]B 77. It has been claimed that this passage is directed against Taosim as teaching that all intellectual endeavour is useless.

Remark Compared with Western logic, *Mingbianxue* is more concerned with practical issues. The main purpose of Confucian's "rectification of names" is to serve the government, and even with the School of Names and the Moists, their view on disputation is practical. This may has something to do with the social situation at that time, different states fight against each other, and schools must come up with good theories to help their king to win. So not surprisingly, pure scientific exploration is often mixed with concerns about practical matters –somewhat ominous for the fate of the *Mingbianxue*. After the unification of China by the state of Qin, in order to unify all thought and political opinion, the Emperor Qin Shihuang ordered a burning of all historical books except the history of Qin, and scholars were suppressed as well. Thus the Hundred Schools of Thought were marginalized except for the school of Legalism. Later on, in the Han dynasty, the Emperor Wu espoused Confucianism as the orthodox state ideology, proscribing all non-Confucian schools of thought. Even so, though the political environment remained unfavourable to the logic-oriented Moist School ever since, many of its ideas survived under later dynasties. We will see how in the next section.

3 Later Development of Chinese Logic

The main development of Chinese logic in more recent times took place in four periods: the Han, Wei-Jin, Song and Qing Dynasties. In what follows, we discuss a few representative scholars or works from each.

In the Han Dynasty, the mainstream of intellectual activities was reflection on and synthesis of different earlier schools. The Masters of Huannai (*huannanzi*, 淮南子) was one, edited by the King of Huainan Liu An (179–122 B.C.). The book consists of 21 chapters, with ideas from many schools. As far as logic is concerned, it further developed the theory of analogical reasoning. The main ideas are the following. To carry out an analogical inference correctly, one must first "know the kinds (*zhilei*, 知类)." To know the kinds means to know the sameness and difference of the kinds. After one knows the kinds, one can reason on the basis of it (*yileituizhi*, 以类推之). The book presents many examples to show when one can infer with kinds, and when one cannot (*lei buke bixu*, 类不可必推). Here is an example. A small horse with big eyes cannot be called a big horse, but a big horse with blind eyes can be called a blind horse. Here big eyes with a small horse do not affect its physical capacity, so we cannot add 'big' to horse. By contrast, blind eyes do affect the physical capacity

of a horse, so we can call the horse a blind horse. The two situations look similar, but are essentially different. Here is one more example in the book. One may die because of a small injury in one's fingers, but one may survive even if one's arm was cut. So one cannot simple conclude that big injury leads to death, and one can survive all small injuries. The book advocates care in analogical reasoning. It also lists mistakes in analogical inference and analyzes possible reasons for such errors. Again the point stressed is the importance of Zhilei: one has to recognize the essential properties of kinds, and the necessary relations between different kinds.

Like the Pre-Qin period, Wei-Jin (220–420 A.D.) is one more era in Chinese history known for its free intellectual atmosphere. Ji Kang (224-263 A.D.), Wang Bi (226-249 A.D.) and Ouyang Jian (267–300 A.D.) were prominent scholars. The relation between names, language and objects was a core issue that was extensively discussed. Wang Bi's view is called "言不尽意 *yan bujinyi*", language is not adequate to express meaning – and in complete contrast to this, Ouyang Jian argued for "言尽意 *yan jinyi*", language is adequate to express meaning. These discussions extended the tradition of the School of Names. But the theory of argumentation was taken further, too. Ji Kang stated explicitly that the purpose of a disputation is to find the natural rule of things. One has to think carefully and distinguish what is right from what is wrong, and one cannot rely on what was said before. Ji Kang proposed several concrete strategies for disputation, such as trying to avoid affirming two contradictory statements. One should take all the cases of the issue under discussion into account, not only one or the other. In particular, to reject an opponent's claim, he proposed a method very similar to "reduction to absurdity". These strategies abound in his works *Essay on Nourishing Life* (*yangsheng lun*, 养生论) and *On the Absence of Sentiments in Music* (*shengwu aile lun*, 声无哀乐论).

As for the *Moist Canons*, a very important contribution was *Annotated Moist Canons* (*mobian zhu*, 墨辩注) of Lu Sheng[46]. For unknown reasons, this book got lost. What is available is its preface, which was found in *The History of Jin* (*jinshu*, 晋书)[47]. The preface contains only 294 Chinese characters, but it summarized the main lines of *Mingbianxue*. In the preface, for the first time, Lusheng mentioned the textual organization of the *Canons*, and he proposed reading them according to the following rule:

[46]It is commonly believed that he lived between the 3th century and the first part of the 4th century.

[47]This is about the history of West Jin (265–316) and East Jin (316–420), with 21 scholars involved.

for each section, one uses *Shuo* to interpret *Jing* (*yinshuo jiujing, gefu qizhang*, 引说就经，各附其章). To understand the importance of this contribution, one has to know a bit about the Chinese history of printing. The *Canons* were first printed on bamboo slips (*zhujian*, 竹简), but later they were copied on silk (*boshu*, 帛书). With the bamboo slips, for each section, *Shuo* comes after each *Jing*, and thus it was naturally divided into two bamboo slips bound together, that can be read from right to left. In the transition from bamboo slips to silk printing, the texts in *Jing* and in *Shuo* were mixed up, so that people could no longer understand them when copied on silk. This observation by Lu Sheng turned out extremely helpful to later researchers in understanding the *Canons*.

In the Song Dynasty, Chinese philosophy reached its peak. The dominant philosophy was called *Lixue* (Studies on *Li*): the main concern of the philosophy was to find Li for everything. Many works of the period contained discussions of logical issues. We only give a few examples. Based on iconographic and cosmological concepts, Shao Yong (1011–1077) took an "image-number study" approach to study the *Book of Changes*. He wrote an influential article on cosmogony, *Book of Supreme World Ordering Principles* (*huangji jingshi*, 皇极经世), to argue that numbers are the origin of the universe, and everything else can be derived from them. In particular, he placed the Hexagrams of the *Book of Changes* into a binary order (the Fu Hsi Ordering). These ideas reached Europe in the 18th century. Leibnitz was deeply impressed when he saw them in 1701, and his views on a universal language and binary arithmetic were influenced by it. One more example is logical inference. Besides Li, "*gewu zhizhi* (格物致知)" was another core notion to *Lixue*, which means "to study the phenomena of nature in order to acquire knowledge". When explaining Gewu zhizhi, Zhu Xi (1130–1200) talked about inductive and deductive inference:

> "There are two ways of getting knowledge, one is to explore from the bottom to the top, the other is to explore from the top to the bottom . . . Gewu is to study many things to get general knowledge, Zhizhi is to infer from general knowledge to concrete things."[48]

The method from top to bottom is what we would call deduction, and that from bottom to top induction. *Li* often acted as a general rule of deduction to understand things in the world.

[48] *Zhuzi Yulu*, edited by Li Jingde, appeared in 1270. It is a collection of conversation between Zhu Xi and his decedents, a valuable resource to understand the ideas of Lixue.

Moving one more historical period, at the end of Ming Dynasty, the philosopher Fu Shan (1607–1684) started annotating the chapter *Daqu* in the *Canons* – a starting point of a different approach to their study. The Qian-Jia Textual Research School of Thought in the Qing Dynasty (1644–1911) followed Fu Shan's ideas, and systematically went back to the classic works from the Pre-Qin period. The Canons received unprecedented attention. Most of these works are purely textual studies, trying to restore the original text of the classics. Zhang Huiyan (1761–1802) re-arranged the four chapters of the Canons: For each section *Shuo* follows *Jing*, as Lu Sheng had suggested. Building on this, Wang Niansun (1744–1832), Wang Yinzi (1766–1834)[49] and Sun Yirang (1848-1908) began to study the Canons text very carefully, proof-reading and annotating every sentence. Sun Yirang wrote a book *Mozi Jiangu* (墨子间诂) after 30 years' effort. The book immediately became the most important reference in the research of the Canons even since it appeared in 1898. After that, most of the logic texts of the Moists became accessible.

One point to realize here is that key works of Indian and Western logic had been introduced to China by that time.[50] Sun Yirang points out in his book that there are principles in the *Canons* that are similar to Aristotle's deductive reasoning, Bacon's induction, and Indian Hetuvidyā, which paved the way for the comparative studies carried out by Liang Qichao and Zhang Taiyan in the early 20th century. We will see these in the next section.

Looking back along this long history, although *Mingbianxue* was not a popular subject after the ancient unification of China, its logical themes were develop steadily by many scholars, and logical skills were explicitly discussed in the philosophical literature. Of course, we only gave a glimpse of this long period, and a more systematic study is urgently needed.

[49] Wang Niansun and Wang Yinzhi are father and son.

[50] Hetuvidyā was first introduced to China in the 6th century. It became very popular in the Tang Dynasty and expanded its influence to other Asian countries, e.g. Japan and Korea. However, there was little development during the Song, Yuan and Ming Dynasties. Only in the late Qing Dynasty, several scholars got interested in Yogācāra, and studies on Hetuvidyā were resumed.
The first translation of Euclides' Elements of Geometry by the Jesuit Matteo Ricci and a Chinese scientist Xu Guangqi appeared in 1607. In the early 20th centuries, further logic textbooks were translated. The Chinese version of Mill's *A System of Logic* appeared in 1905, and of Jevons' *Elementary Lessons on Logic* in 1907. For more details, cf. Song Wenjian, *Introduction and Studies of Logics* (*luojixue de chuanru yu yanjiu* 逻辑学的传入与研究), Fuzhou: Fujian Renmin Press, 2005.

4 Logic Studies in the Early 20th Century

After the introduction of Indian logic, and especially Western logic at the end of the 19th and beginning of the 20th century, Chinese logic attracted more and more attention. In this concluding section, we briefly look at what happened in the early 20th century.

Liang Qichao (1873–1929) published an article on *Lunlixue of Mozi* (墨子论理学)[51] in 1904, where he took notions from Western logic to interpret the *Moist Canons*. He said "what is called logic in the West is what is called bian by the Moists" and "The notions of *ming*, *ci* and *shuo* are concepts, propositions and inference in Western logic". Concerning reasoning patterns, Liang Qichao observed that Aristotle's syllogism consists of three parts, a major premise, a minor premise and a conclusion. The situation in Indian logic is similar, there are three parts, too, called pratijn*ã*, (the proposition or conclusion), hetu (the reason), and ud*ã*harana (the example). In Chinese logic, the three parts are *ci* (the proposition), *gu* (the reason) and *lei* (the kind). On the basis of his comparative studies, Liang claimed there is Chinese logic.[52]

In 1917, Hu Shi finished his dissertation *The Development of the Logical Method in Ancient China* (*xianqin mingxueshi*, 先秦名学史), which explored the logic of several schools in the Pre-Qin period. In particular, he used the three categories "*gu*, *lei* and *fa*" to understand logical inference in Chinese logic. Hu's work was influential in the West, many scholars first learnt about Chinese logic from it.

Zhang Taiyan (1869–1936) compared the three logics in a more sophisticated way. He agreed with Liang that logical inference in all three traditions consists of three steps. But he pointed out that what is different is the order of the steps. In Western logic, the proposition comes first, then the reason, finally the example. In Indian logic, the example and the reason come before we get to the conclusion. In Chinese logic, reason and proposition come first, then the conclusion. Take a classical example we used before, in Aristotle's logic we have:

> All human beings are mortal,
> Socrates is a human being,
> So Socrates is mortal.

[51] "Lunlixue" was the Chinese term being used to translate "logic". This translation was adopted in Japan first.

[52] The issue of whether there is a Chinese logic in a technical sense is still controversial nowadays, with answers heavily depending on authors' own notion of logic.

In Indian logic, one would have the following:

> Socrates is mortal,
> Socrates is a human being,
> So all human beings are mortal.

In Chinese logic, it would go as follows:

> Socrates is a human being,
> All human beings are mortal,
> So Socrates is mortal.

According to Zhang Taiyan, these different orders reflect different ways of thinking. He also discussed the difference in the context of argumentation. He found that the Indian style of inference best serves the purpose of real argumentation, as a combination of induction and deduction. Judging logics by their application in argumentation influenced later research.

Other scholars in the early 20th century, like Tan Jiefu (1887–1974) also contributed to comparative studies. While parts of these early works were rigid and unimaginative by today's standards, this comparative phase has proved very fruitful, becoming a powerful stream of work by logicians in mainland China and worldwide.[53] We hope to review these achievements in detail on some other occasion.

[53]Here are some works in this line: Angus. C. Graham. *Later Moist Logic, Ethics and Science*. The Chinese University Press, 1978. Chad Hansen, *Language and Logic in Ancient China* (Ann Arbor: University of Michigan Press, 1983). Christoph Harbsmeier, "Language and Logic, " in *Science and Civilization in China*, vol.7, ed. Joseph Needham (Cambridge: Cambridge University Press, 1998). Cheng Chung-ying, "Inquiries Into classical Chinese Logic," *Philosophy East and West* 15, no. 3/4 (1965): 195–216; "Logic and language in Chinese thought," in *Contemporary Philosophy: A Survey*, ed. Raymond Klibansky (Florence: Institute Internationale di Philosophia, 1969, 325–347); "Kung-Sun Lung: White horse and other issues," *Philosophy East and West* 33, no. 4 (1983): 341–354. Zhang Chunbo and Zhang Jialong, "Logic and language in Chinese philosophy," in Brian Carr, editor, *Companion Encyclopedia of Asian Philosophy*. London: Routledge, 1997, 620–635. Zhang Jialong, editor, *History of Chinese Logical Thought*, Changsha: Hunan Education Press, 2004. Zhou Yunzhi, *History of Chinese Logic*. Taiyuan: Shanxi Education Press, 2002. Cui Qingtian, *Comparative Studies on Moist Logic and Aristotle's Logic*, Beijing: Renmin Press, 2004. Sun Zhongyuan, *Studies on Chinese Logic*. Beijing: Shangwu Yinshuguan, 2006.

What we want to emphasize at this point is the broader significance of the comparative studies made by the early pioneers that we discussed. They crossed between cultures, and while misunderstandings did occur, the result was a meeting of traditions.[54]

5 Conclusion

In this paper we have walked, lightly, from the 6th century B.C. all the way to the 20th century. First, we explained the main theories of Chinese logic in its golden age of the Pre-Qin period. We then sketched how these thoughts (especially, theory of names and kind-based reasoning) developed later on, with a focus on the Han, Wei-Jin, Song and Qing Dynasties – though a more systematic investigation is called for. Finally, we briefly looked at the first serious meeting of traditions: the comparative studies on Chinese logic facing its Indian and Western counterparts in the early 20th century, an encounter that was crucial to modern Chinese logic.

Acknowledgement We thank the guest editors for their effort in putting together this volume. We thank Jeremy Seligman and Johan van Benthem for their very useful comments and corrections.

[54]This theme of history of logic and cultural communication will be the subject of an upcoming workshop in Amsterdam on "History of Logic in China": http://www.sciencehistory.asia/history-logic-china

PART II

Mathematical Logic and Foundations

Model Theory

ANAND PILLAY *†

1 Introduction

Contemporary or modern (mathematical) logic was born at the end of the 19th century. Its origin is connected with mathematics rather than philosophy, and my article will likewise be informed by a mathematical culture although I will try make connections with philosophy and the philosophy of mathematics. Although mathematical logic emanates from a so-called Western intellectual tradition, it is now, like mathematics as a whole, a world subject with no essential national or cultural distinguishing marks.

Unfortunately I am not knowledgeable about philosophical and (early) mathematical traditions in the Indian subcontinent, so will not be able to make any serious comparative analyses. Also I am not trying here to give a proper history of model theory with appropriate references, bibliography, credits etc., but rather a description of how I see the subject now, with some minor commentary on historical developments. Also I will only be able to give a hint of the main technical notions and definitions in the subject. So I will point the reader towards a few basic texts, reviews, and historical accounts of the subject, where more details and as well as a detailed bibliography can be found, such as Hodges' textbook and history [3], [4] and Marker's textbook [5]. Another survey [7] by myself contains more technical details than the current article, and my book [6] from 1996 contains an exhaustive technical treatment of some of the themes I will discuss, but assuming a prior acquaintance with model theory. The volume [2] is a good reflection of the state of model theory around the beginning of the modern era (1971). It also contains an informative historical article by Vaught on model theory up to 1945. Finally the book [1] gives a readable account of some of the machinery behind one of the major modern successes of the applications of model theory (mentioned at the end of Section 6).

*University of Leeds

†Supported by EPSRC grant EP/F009712/1, and also by a Visiting Professorship at the University of Paris-Sud in March-April 2010.

Among the strands in the early history of logic were identifications of correct standard forms of argument (the syllogisms) but also, with Gottfried Leibniz, the rather bold idea that one might in principle be able to settle all disputes by mechanical logical means. These were complemented by considerations of the nature of mathematical truths compared to empirical truths (e.g. Kant), as well as the beginnings of the mathematicization of logic (e.g. Boole).

So "logic" here is supposed to refer to intrinsic reasoning or truths, independent of experience. For example the statement that a thing is equal to itself is a truth of logic rather than experience, although philosophers such as Hegel (and also I guess many Indian philosophers) have commented on the vacuity and even conditionality of such truths. Likewise the fact that from "P implies Q", and P, we can deduce Q, is supposed to be valid on purely logical grounds, independent of which statements P and Q denote.

Rather than try to base all knowledge on logic, Frege and Russell, among others, attempted to show that all or at least major parts of *mathematical knowledge* can be founded on logic. Once one starts to investigate seriously such claims, one is forced to define one's terms, and find a formal framework within which to carry out the project. And this, in a sense, was behind the birth of modern logic. But another crucial factor was that dominant mathematicians of the time, such as Hilbert and Poincaré, were very caught up in "foundational" problems, not only around whether mathematics could be reduced to logic, but also about the justifications of the use of "infinitistic" methods and objects, outside the scope of normal intuition. As it turned out Gödel's work in the 1930's showed that not only the Frege-Russell-Whitehead project, but also a "second level" program of Hilbert to "reduce" infinitistic to finitistic methods, were doomed.

In spite of this failure of the logicist and Hilbert programs, the efforts of these late 19th and early 20th century logicians left a lasting impact on mathematics (and also philosophy). Firstly set theory as a universal language for mathematics was largely accepted (even though not all mathematical truths could be settled on the basis of accepted axioms about sets), and this contributed towards the possibility of mathematicians from different subdisciplines being able, at least in principle, to communicate in a precise and effective manner with each other. And of course the search for additional axioms for sets led to the rich subject of contemporary set theory. Moreover the "defining of one's terms" issue mentioned above led to precise mathematical treatments of notions such as truth, proof, and algorithm. It is interesting that model theory (truth), proof theory (proof) and recursion theory (algorithm), together with set theory, remain the four principal

and distinct areas of contemporary mathematical logic. In any case modern mathematics, its language, and unity, are closely bound up with logic, although paradoxically logic has been somewhat marginalized within contemporary mathematics. Nevertheless, mathematical logic is now undoubtedly regarded as a bona fide part of mathematics and the various areas and subareas have their own internal programs and aims, which are continually being modified. But one can ask to what extent these investigations can have impacts on mathematics as a whole, as was the case at the beginning of the 20th century. I will try to convey something both of the "inner movement" of model theory, as well as its actual and potential wider impacts. To read this article profitably will require some mathematical background, but as mentioned above I will try to comment on the "philosophical" content and impact too.

2 Truth

The notion *truth in a structure* is at the centre of model theory. This is often credited to Tarski under the name "Tarski's theory of truth". But this "relative", rather than absolute, notion of truth was, as I understand it, already something known, used, and discussed. In any case, faced with the expression "truth in a structure" there are two elements to be grasped. Truth of what? And what precisely is a structure? An illuminating historical example concerns the independence of Euclid's "axiom of parallels" from his other axioms. A statement equivalent to this axiom of parallels is

(AP): given any line ℓ and point p not on ℓ there is exactly one line through p which is parallel to (does not intersect) ℓ.

The independence statement is that (AP) is *not* a logical consequence of a certain collection $\mathcal{A}$ of other axioms involving points and lines (such as that any two distinct points lie on a unique line). This was shown by finding a "model" of the set $\mathcal{A}$ of axioms in which moreover the statement (AP) is false. The kinds of things here that are (or are not) true are statements such as (AP) or the statements (axioms) from $\mathcal{A}$. And the relevant structure or "model" consists of one collection P of objects which we call "points", another collection L of objects, called "lines", and a relation I of "incidence" between points and lines, thought of as saying that p lies on ℓ. Note that (AP) can be expressed, in a somewhat convoluted manner, as follows:

(*) for any p and for any ℓ such that *not* $pI\ell$, [there is ℓ' such that ($pI\ell'$ and it is not the case that there exists p' such that $p'I\ell$ and $p'I\ell'$) and for

any ℓ'' such that $pI\ell''$ and it is not the case that there exists p' such that $p'I\ell''$ and $p'I\ell''$, $\ell'' = \ell'$].

So the structure constructed (a model of non Euclidean geometry) was one where the statements in $\mathcal{A}$ are true and the above statement (*) is *false*.

Already there is a considerable degree of abstraction in my presentation. The intuitive geometric notions of point and line are replaced by purely formal sets and relations. This is a typical example of a *structure* in the sense of model theory, logic, or universal algebra (or even Bourbaki), namely a universe of objects, together with certain relations between them. In the example the objects come in two sorts, "points" and "lines" and the only relation is I. Moreover statements such as (*) above, have a rather definite logical form. They involve the basic "variables" p, ℓ, as well as expressions (logical connectives) such as "and", "not", "for all", "there exists", as well as "equality". To check the truth or falsity of such an axiom in a structure, the "for all" and "there exists" connectives should range over objects in the structure at hand, and it is this kind of proviso which typifies "truth in a structure" as opposed to "absolute truth".

So at the basic level, model theory is concerned with two kinds of things, structures and formal sentences (or statements), as well as the relation (truth or falsity of a sentence in a structure) between them. Traditionally the expressions *syntax* (for formal statements) and *semantics* (for the interpretation of sentences in structures) were a popular way of describing model theory. The formal sentences in the example above belong to what is called *first order logic*, because the *for all*, and *there exists* expressions (or quantifiers) range over objects or elements of the underlying set of the structure (rather than subsets of the underlying sets for example). Higher order and/or infinitary logic involve quantifying over subsets or subsets of the set of subsets etc, and/or infinitely long sentences or expressions. There are also other variants, involving cardinality or probability quantifiers for example. These higher order or infinitary logics were extremely popular in the 1960's and 1970's, and are still the subject of substantial research. However we will, in this article, concentrate on the first order case.

So, summarizing, a *structure* M is a set X equipped with some distinguished family $\mathcal{R}$ of *relations* on X, namely subsets of $X, X \times X, X \times X \times X$ etc. We also allow a family $\mathcal{F}$ of distinguished functions from $X \times X \times ... \times X$ to X. There are two typical kinds of examples. First of a combinatorial nature such as *graphs*. A graph is a set X (of "vertices") equipped with a binary relation $R \subset X \times X$, representing adjacency. Secondly, the structures of algebra, such as groups, rings, fields etc. For example a group is a set X equipped with a function $m : X \times X \to X$ satisfying the group axioms

(associativity, and existence of an identity and inverses). Corresponding to a structure M is a formal first order language $L(M)$ within which one can express properties which may or may not be true in the structure M. For example, in the case of graphs the property that every element is adjacent to another element can be expressed by:

for all x there is y different from x such that $R(x, y)$,

or more formally

$$\forall x \exists y (x \neq y \wedge R(x, y))$$

Likewise in the case of groups the basic group axioms can be expressed in a first order manner, and by definition a group is a structure (with a single distinguished binary function) in which these axioms are true.

Commonly the notion that a (formal) sentence σ is true in a structure M, is also expressed by saying that M is a *model* of σ, as discussed at the beginning of this section. The formal notation is $M \models \sigma$.

What is called a *theory* in logic is some collection of sentences belonging to some first order language. An example of such is $Th(M)$ for M a given structure, namely the collection of all sentences in $L(M)$ which are true in the model M.

If M and N are structures for a common first order language (for example M, N are both graphs) it makes sense to ask whether M and N are *isomorphic*, meaning that there is a bijection between the underlying sets X, Y say of these structures which interchanges the distinguished relations. Being isomorphic means being the same to all intents and purposes. A weaker notion is *elementarily equivalence* meaning that any first order sentence true in M is true in N (and vice versa). The question of when elementarily equivalent implies isomorphic is a pervasive problem in model theory which will be discussed subsequently.

I mentioned at the beginning of this paragraph the idea that “truth in a structure" is a kind of relative rather than absolute truth. However I should make it clear that this is neither a notion of “truth in a possible world", nor “truth relative to a point of view", nor “approximate truth", although model-theoretic tools have been used to explore these latter notions.

3 Decidability

I want to distinguish at the beginning between those first order *theories* which I will call *foundational* and those which I will call *tame*. The foundational theories (such as the accepted axioms of set theory in the language with a "membership relation") are those which purport to describe all or large chunks of mathematics, and are connected to the origin of modern logic as described in section 1. Gödel proved that in general such foundational theories are *undecidable*. Namely there is no algorithm to decide whether or not a given (formal) statement, is or is not a consequence of the axioms. Among the important foundational theories is $Th(\mathbb{N})$ where the structure $\mathbb{N}$ consists of the set of natural numbers equipped with addition and multiplication. Undecidability of $Th(\mathbb{N})$ amounts to there being no algorithm or effective method for deciding which (first order) statements about $\mathbb{N}$ are *true*. The proof of this rests on Gödel's insight that arithmetic, namely the structure $\mathbb{N}$, is rich enough to represent reasoning and computation in a "first order" manner. So for example any effective procedure for deciding which first order statements or sentences are true in $\mathbb{N}$ would yield an effective procedure for deciding whether or not for any given computing device and any given input, there is a well-define output (which is known to be impossible). At the opposite end of the spectrum are the "tame" theories and/or structures, which are as a rule decidable. A typical example is real plane geometry. The real plane $P = \mathbb{R}^2$ is just a flat surface, as usually understood, stretching to infinity in all directions. The relevant structure has two sorts of objects, the set P of points of the plane, and the set L of straight lines in P, equipped with a single relation $I(p, \ell)$ expressing that the point p is on the line ℓ. It is a fact that the structure $M = (P, L, I)$ is decidable. Already one sees a distinction between "geometry", represented by the structure M, and arithmetic, represented by the structure $\mathbb{N}$. In addition to the real numbers there are other number systems which belong to geometry, such as the complex numbers and the p-adic numbers. And again the number systems themselves (fields), or plane geometry over those number fields, are decidable structures.

The distinction between "foundational" and "tame" theories is heuristic rather than mathematically precise. But model theory does have a number of precise notions other than decidability, which separate these classes of theories, and more generally provide other meaningful dividing lines between classes of first order theories and structures. Contemporary model theory has tended to concentrate on the tame region of mathematics, al-

though exploration of the borderline or middle ground between tame and foundational is a fascinating topic.

4 Foundations revisited

As mentioned in the introduction the two main programs to build mathematics on, or recover mathematics from, logic, namely axiomatic or set-theoretic (Frege, Russell, Whitehead), and proof-theoretic (Hilbert), failed. But as one might expect, these programs have been preserved or resurrected in more modest fashions. The proof theory/set theory/recursion theory nexus has been the main environment for such endeavours. One of the popular programs is what is called “reverse mathematics", developed by Harvey Friedman and Steve Simpson among others. To go into detail here would be be too technical for the mature of this article. But briefly the idea is to recover certain parts of mathematics from certain parts of logic (and vice versa) at the level of theorems and axioms. The logical environment here is what is called *second order arithmetic*, although it is actually a first order theory. The kind of axioms considered are *set existence* axioms of a logical nature. It was recognized rather early that theorems of mathematics, such as the existence of solutions of differential equations, depend on such logical axioms of various levels of strength. The point of reverse mathematics is that often one can in turn derive the logical axiom from the mathematical theorem. So here the strength or content of an axiom of logic is expressed by an accepted theorem of mathematics. This gives a new sense in which logic explains mathematics, mathematics is recovered from logic, or even logic is recovered from mathematics. This subject of reverse mathematics has not been uncontroversial, but nevertheless the subject has had a pervasive influence around the proofs/sets/computability side of mathematical logic. One of the things I want to discuss is a kind of reverse mathematics at the level rather of logical properties and mathematical objects. The logical properties will come from model theory, and the mathematical objects from some basic kinds of geometry. The whole relationship will exist within “tame" mathematics, far from the foundational theories discussed earlier. This “model-theoretic" reverse mathematics was the creation of Boris Zilber. But there are a couple of provisos. First the relationships between logical properties and geometry, were just conjectural. Secondly these conjectured relationships turned out to be false. In the next section I will describe this model-theoretic reverse mathematics.

5 Categoricity

A natural property of a structure M for a first order language L is *categoricity*, which means that whenever N is elementarily equivalent to M then in fact N is isomorphic to M. Namely M is completely determined by the first order sentences which are true in M. Unfortunately (or fortunately) because of the *compactness theorem* of first order logic, a structure M will be categorical if and only if it (or rather its underlying set) is *finite*. (The compactness theorem states that a set Σ of first order sentences has a model if and only if every finite subset of Σ has a model.) As model theory typically deals with infinite structures, the next best thing is the notion of categoricity with respect to a cardinal number. So here some acquaintance with basic set theory, cardinal numbers and ordinal numbers, is required. The smallest infinite cardinal is $\aleph_0$ the cardinality of the set of natural numbers. The next bigger after that is $\aleph_1$. The cardinal numbers are all of the form $\aleph_\alpha$ for some ordinal α. As soon as M is infinite, there will (by the compactness theorem) be structures elementarily equivalent to M of any infinite cardinality. For κ an (infinite) cardinal, we will say that the structure M is κ-categorical if whenever M_1, M_2 are structures elementarily equivalent to M, both of cardinality κ, then M_1 and M_2 are isomorphic. By definition the property that M is κ-categorical, is a property of the first order theory $Th(M)$ of M.

It turns out that the case when $\kappa = \aleph_0$ is very special and in some sense a "singularity". The study of $\aleph_0$-categorical structures is equivalent (by considering automorphism groups) to the study of a certain class of infinite permutation groups, often called "oligomorphic" permutation groups. The model-theoretically more interesting notion is κ-categoricity, for *uncountable* κ, namely $\kappa > \aleph_0$ (or $\kappa = \aleph_\alpha$ for $\alpha > 0$). In this context we have the celebrated theorem of Michael Morley that a structure M is κ-categorical for *some* uncountable κ just if M is κ-categorical for *any* uncountable κ. Bearing in mind Morley's Theorem we use the expression *M is uncountably categorical* for "M is κ-categorical for some uncountable κ".

One of the key "number systems" in mathematics is the field $\mathbb{C}$ of complex numbers. We view this as a structure $(\mathbb{C}, +, \times)$ namely $\mathbb{C}$ equipped with addition and multiplication as distinguished functions. For different reasons related to *definability* which will be discussed later, this structure is sometimes identified with the subject *algebraic geometry*, the study of sets of solutions of systems of polynomial equations. What is relevant to our current discussion is that $(\mathbb{C}, +, \times)$ is an uncountably categorical structure: any structure elementarily equivalent to it will be an *algebraically closed*

field of characteristic 0, $(F, +, \times)$, the isomorphism type of which is determined by its transcendence degree, which coincides with its cardinality if F is uncountable.

Another basic example of an uncountably categorical structure is a *vector space* V over a countable field F. The structure is $(V, +, \{f_r : r \in F\})$ where $f_r : V \to V$ is scalar multiplication by r. The structures elementarily equivalent to this are precisely the vector spaces over F, each of which is determined by its F-dimensions, which again agrees with its cardinality in the uncountable case. Again for definability reasons, this structure (or class of structures) is sometimes identified with linear geometry over F (sets of solutions of linear equations).

A third basic example is the set $\mathbb{Z}$ of integers (positive AND negative) equipped with the successor function f which takes x to $x + 1$. $Th(\mathbb{Z}, f)$ contains the information that the underlying set is infinite and that f is a bijection such that for each n $f^n(x) \neq x$ for all x (where f^n denotes f iterated n times). We leave it to the reader to check that again this structure is uncountably categorical. One can not really see any natural geometry attached to this structure.

The thrust of what came to be called Zilber's conjecture was that, in a technical sense which I do not want to go in to now, the above three structures (or rather their theories) are the *only* examples of uncountably categorical structures. So Zilber's conjecture was saying that some very fundamental structures of mathematics can be characterized by logic, namely through the notion of uncountable categoricity of their first order theory, and so in a sense this class of structures is "implicitly defined" by logic. This conjecture and in fact the general point of view giving rise to it, presents another possible fundamental relationship between logic and mathematics.

Zilber's conjecture turned out to be false. Ehud Hrushovski, in the late 1980's, found a combinatorial method for constructing new uncountably categorical structures which do not fit into the three cases described above. For now let us say that the first example above (the complex field) has a model-theoretic property called *nonmodularity*, the second example has a property *modularity and nontriviality* and the third a property *triviality*. What Hrushovski's examples gave were nonmodular structures which were not "essentially" algebraically closed fields. Zilber has since attempted to preserve at least the spirit of his original conjecture by trying to show that these new examples of Hrushovski also have a geometric origin and correspond to some classical mathematical objects. But what for me is more interesting is the fact that the original Zilber conjecture is valid in a

range of very interesting and rich contexts, and carries with it new insights as well as analogies between different parts of mathematics. Some such examples will be discussed below.

6 Definable sets

An interest in the *definable sets* in a structure M has always been present in model theory. But since the 1980's the study of definable sets has moved to centre stage in the subject. In section 2, I introduced and discussed the notion σ is true in M, notationally, $M \models \sigma$, where M is a structure for a language L and σ is a first order sentence of that language. In particular I mentioned the sentence

$$\forall x \exists y (x \neq y \wedge R(x, y))$$

in the language of graphs expressing that every element is "adjacent" to another element. However consider the expression

$$\exists y (x \neq y \wedge R(x, y))$$

which I will denote $\phi(x)$. It does not really make sense to ask whether this expression is true or false in a structure $M = (X, R)$, because it depends on what x refers to. But it *does* make sense to ask, given a structure M together with an element $a \in X$, if $\phi(x)$ is *true of* a in M, which in this specific case means to ask whether a is adjacent to some other element in M. We write $M \models \phi(a)$ to mean that $\phi(x)$ is *true of* a in M. The set of such a, is a typical example of a *definable set* in M. The expression $\phi(x)$ above is called a *first order formula*, and x is called a *free variable* in the expression, because it is not controlled or quantified by a "for all" or "there exists". Likewise we can speak of formulas $\psi(x_1, x_2, .., x_n)$ of a first order language L, in any number of free variables. If M is a structure for such as language, then the set defined by ψ in M is, by definition:

$$\{(a_1, .., a_n) \in M^n : M \models \psi(a_1, .., a_n)\}$$

Sets as above, which are collections of finite tuples of the underlying set of the structure M, are precisely what we call *definable sets* in the structure M. There is a natural way of saying that a map (or function) between two definable sets is definable. Hence from a structure M we obtain a *category* $Def(M)$, the category of definable sets in M.

It has become useful to think of definability in a "geometric" rather than "combinatorial" way. For example consider the circle with centre $(0, 0)$ and radius 1 in the real plane. It is defined in the structure $(\mathbb{R}, +, \times, 0, 1)$ by the formula

$$\phi(x_1, x_2) : x_1^2 + x_2^2 = 1$$

Note that the formula $\phi(x_1, x_2)$ does not contain any "for all" or "there exists". It is a *quantifier-free* formula. On the other hand the formula

$$\psi(x) : (\exists y)(x = y^2)$$

does have a quantifier, and moreover defines, in $(\mathbb{R}, +, \times, 0, 1)$ the set of nonnegative elements of $\mathbb{R}$.

Many structures M of a "tame" nature often have a "quantifier-elimination" property that definable sets in the structure can be defined by formulas with not so many quantifiers ("for all", "exists"). This enables one to get a handle on $Def(M)$. In the case of $(\mathbb{C}, +, \times, 0, 1)$ there is a full quantifier elimination, in the sense that all definable sets are defined without quantifiers. The consequence is that $Def((\mathbb{C}, +, \times, 0, 1))$ is "essentially" the category of "complex algebraic varieties". In the case of $(\mathbb{R}, +, \times, 0, 1)$ there is a relative quantifier elimination yielding that the category of definable sets is precisely the category of "semialgebraic" sets. Each of these categories (algebraic varieties, semialgebraic varieties) corresponds to a whole subject area of mathematics. These quantifier elimination results are associated with Abraham Robinson and Alfred Tarski. Moreover in the case of $(\mathbb{R}, +, \times, 0, 1)$ the relative quantifier elimination result lies at the foundations of *semialgebraic geometry*.

The (so far undefined) properties of nonmodularity, triviality, etc. from section 5, can be expressed or seen in the behaviour of definable families of definable sets. For example nonmodularity of $(\mathbb{C}, +, \times)$ is seen via the 2-dimensional family of lines in $\mathbb{C}\times\mathbb{C}$ (a 2-dimensional definable family of definable 1-dimensional subsets of $\mathbb{C} \times C$). Among other rich mathematical structures M where $Def(M)$ is tractable, are *differentially closed fields*, and *compact complex manifolds* (proved by Robinson, and Zilber, respectively) The mathematical sophistication increases here. But in both these cases the Zilber conjecture from section 5 is true, in suitable senses. Moreover, without going into definitions and extreme technicalities, the property of "nonmodularity" has definite mathematical meaning and consequences in these examples. Differentially closed fields are "tame" structures appropriate for or relevant to the study of ordinary differential equations in regions in the complex plane. Definable sets are essentially solution sets of

differential equations. And the property of nonmodularity (of a definable set) is related to the complete integrability of the corresponding differential equation. For compact complex manifolds, a definable set is essentially a compact complex analytic variety, and nonmodularity is related to it being "algebraic" (biholomorphic to a complex algebraic variety).

Among the celebrated applications of model theory to other parts of mathematics is Hrushovski's proof of a certain "number theoretic-algebraic geometric" conjecture, the Mordell-Lang conjecture for function fields of positive characteristic, which makes essential use of the validity of the Zilber conjecture in "separably closed fields" (as well as using other model-theoretic techniques).

7 Miscellanea

What I have given so far is a discussion of a few themes in contemporary model theory, influenced by my own preoccupations. Here I will attempt to rectify the balance, mentioning other trends and themes (some of which are also close to my own work and interests).

Well into the 1970's it was a pretty common belief within the mathematical logic community that model theory consisted essentially of a collection of tools and techniques related to the fundamental notions of semantics and syntax, possibly enhanced by a few basic theorems. (This may also be suggested by the previous sections of the present paper.) In spite of the strength of logic and model theory in the Soviet Union, Poland, and other countries in Eastern and Western Europe, it must be said that in the 1950's and 1960's the (emerging) subject was dominated by two schools, one around Alfred Tarski in Berkeley, and the other one around Abraham Robinson in Yale. Both stressed the potential applications of model theory within other parts of mathematics (although we should note that already in 1940 the Soviet logician Malt'sev was applying the compactness theorem to obtain results in group theory). In the case of Robinson the intention was very clearly reflected in his pioneering work around nonstandard analysis, the theory of model companions, and applications to complex analysis, among other things. It was a little less clear what Tarski had in mind, in spite of his early and fundamental work on definable sets in the field of reals. But undoubtedly the group around Tarski, including Vaught, Morley and Keisler, set the stage for later developments in "pure" model theory. The 1960's and 1970's also saw a close relationship developing between model theory and set theory, with for example an intense investigation of infinitary and/or

non first order logics, where Tarski and his group had a major influence. In fact around this time the conventional wisdom was that the future of model theory lay in its connection with set theory, in spite of Morley's work on categoricity (from the mid 1960's). It was Saharon Shelah who, building on the work of Morley, showed that (first order) model theory could be a subject with its own coherent and internal program. With the benefit of hindsight I would say that he raised the question of whether there could be a meaningful classification of first order theories (not explicitly involving decidability properties). Shelah tended to look for dividing lines among first order theories, as well as "test questions" which would be answered one way on one side of the dividing line and another way on the other side. The test questions which Shelah asked typically had a strong set-theoretic content, possibly resulting from the surrounding mathematical culture and influences (such as Tarski). One such test question, coming naturally out of Morley's work, was, for a given theory T, what could be the function $I(T, -)$ which for a cardinal κ gives the number of models of T of cardinality κ, up to isomorphism. Shelah's investigation (and solution) of this problem involved a series of dividing lines among first order theories, the first of which was "stable versus unstable". The property of "stability" for a first order theory T (or structure M) vastly generalizes the property of uncountable categoricity from section 5. A rough definition of stability of T is that no linear ordering is definable on any infinite set in a model of the theory T (so the real field $(\mathbb{R}, +, \times)$ is *unstable*). Shelah and other model theorists developed a considerable machinery for constructing structures, classifying structures, and also studying and classifying definable sets in structures, under a general assumption of stability. This is called stability theory. Although Zilber's conjectures were not originally formulated within the generality of stable theories, it is stable theories that provide the right environment for these notions. The integration of these different points of view is often called geometric stability theory (or even geometric model theory).

There are two conclusions to this part of the story. Firstly, that in spite of the heavily set theoretic appearance of Shelah's work in model theory (up to and including the present) it actually has a strong geometric content with amazing mathematical insights. Secondly, it is now uncontroversial that model theory exists as a subject in and for itself, and part and parcel of the subject is its strong connections to other parts of mathematics.

An important and respected tradition in model theory, to which both Robinson and Tarski contributed seriously, and which is already referred to above, is the model-theoretic and logical analysis of specific concrete

structures and theories. But the issue is which notions or "bits of theory" are guiding the analysis. Decidability and quantifier elimination were historically major such notions. Valued fields have been studied logically for a long time. Again the mathematical sophistication increases here, but I just want to comment that fields equipped with a valuation are another context in which"infinitesimals" appear in mathematics. The work of Ax-Kochen-Ershov on the first order theory of Henselian valued fields (late 60's), followed by Macintyre's quantifier elimination theorem for the field $\mathbb{Q}_p$ of p-adic numbers (mid 70's) represented and led to another major interaction of model theory with algebraic geometry and number theory. More recently, this logical analysis of valued fields has been increasingly informed by notions from stability theory, even though the structures under discussion are unstable.

In the early to mid 1980's, several model theorists (including myself) tried to develop a theory, analogous to stability theory, based on abstracting definability properties in the *unstable* structure $(\mathbb{R}, +, \times)$. This came to be called o-minimality. This has been another successful area with close contacts to real analytic geometry. But even here, the connection with stability theory has recently turned out to be much more than an "analogy".

I have restricted myself in the bulk of this article to first order logic and model theory, where the syntax is of a restricted form. But more general logics, involving infinitely long expressions, and/or quantifiers other than "there exists" and "for all", continue to be investigated. At the same time, "finite model theory", the study of the connection between semantics and syntax when we restrict ourselves to finite structures, has seen a fast development and is now integrated into computer science.

References

[1] E. Bouscaren (editor), *Model Theory and Algebraic Geometry: An introduction to E. Hrushovski's proof of the geometric Mordell-Lang conjecture*, Lecture Notes in Mathematics 1696, Springer, Berlin 1998.

[2] L. Henkin et al. (editors), *Proceedings of the Tarski Symposium*, American Math. Soc,. Providence RI 1974.

[3] W. A. Hodges, *A shorter model theory*, Cambridge University Press, 1997.

[4] W. A. Hodges, *The history of model theory*, http://wilfridhodges.co.uk/history07.pdf

[5] D. Marker, *Model theory: an introduction*, Graduate Texts in Mathematics, Springer, 2002.

[6] A. Pillay, *Geometric Stability Theory*, Oxford University Press, 1996.

[7] A. Pillay, *Model Theory*, Notices of the American Mathematical Society, vol. 47, no. 11, December 2000, p. 1373 - 1381.

A Taste of Set Theory for Philosophers

JOUKO VÄÄNÄNEN [*†]

Contents

[*]Institute for Logic, Language and Computation, University of Amsterdam

[†]Research partially supported by grant 40734 of the Academy of Finland and by the EUROCORES LogICCC LINT programme.

1 Introduction

Originally set theory was a theory of infinity, an attempt to understand infinity in exact terms. Later it became a universal language for mathematics and an attempt to give a foundation for all of mathematics, and thereby to all sciences that are based on mathematics. So what is set theory?

Set theory is a very general but still entirely exact theory of objects called sets. It is useful in a number of fields of philosophy, like logic, semantics, philosophy of mathematics, philosophy of language and probably several others, but it is also useful in mathematics, computer science, cognitive science, linguistics, and even in the theory of music. It can be used anywhere where one needs an exact mathematical approach to objects that can be thought of as collections of something.

Even high school mathematics includes simple operations on sets, like union and intersection. College mathematics usually includes set theoretical concepts like ordered pair, cartesian product, relation, function, and so on. Elementary logic courses include such set theoretical concepts as finite sequence and relation. All the concepts mentioned so far are very useful for any philosophy student. Why? Because all these basic mathematical concepts can be given a uniform exact account. In this account any true properties of those concepts can be proved with a simple argument involving only a few lines.

The remarkable thing about set theory is that not only basic mathematics but indeed *all* mathematics can be represented as properties of sets. Thus we can define in set theory the natural numbers, the real numbers, the complex numbers, the Euclidean spaces $\mathbb{R}^n$, the Hilbert space, all the familiar Banach spaces, etc. Moreover, everything mathematicians prove about these objects can be proved from a few relatively simple axioms concerning sets. Therefore it is said that set theory can serve as a universal language of mathematics, indeed a foundation of mathematics. This gives set theory a special place in the philosophy of mathematics.

Of course, a representation of all mathematics in set theory is meant to be taken *only* as a representation. The fact that real numbers can be defined as sets does not mean that real numbers *are* sets. The point is that it is in principle possible to think of real numbers as sets. It is important to note that the goal of set theorists is not to convince other mathematicians that what mathematicians are doing is really set theory. The point of set theory as a universal language of mathematics is that set theory offers a common ground where any unclear argument can be scrutinized. If some argument in mathematics seems to use something that has not been stated, we can

start the process of reducing the argument to the first principles in set theory. If this process is successful, then the argument can be considered valid without question. In this process it becomes clear whether, for example, the Axiom of Choice, a very powerful construction principle for abstract objects, is used. Also, some mathematical results depend on principles, such as the Continuum Hypothesis, that go beyond what is usually considered a priori true. Then the mathematical result can be stated as an implication: if the Continuum Hypothesis is assumed then this or that holds.

2 Elementary set theory

In this section sets are just collections of objects. We shall later define more exactly what this means. We use lower case letters $a, b, c, ...$ to denote sets. Since sets are collections, they have elements i.e. members of the collection in question. If a is a set and b is an element of a, then we write $b \in a$ and read this "b is an element of a" or "b is in a". Two sets are equal if they have the same elements. A set a is a *subset* of another set b, in symbols $a \subseteq b$ if all elements of a are also in b. The simplest sets the *singleton* set $\{a\}$ which has a as its only element, the *unordered pair* $\{a, b\}$ which has a and b and nothing else as its elements, and the *empty* set $\emptyset$, which has no elements whatsoever. Note that $\{a, b\} = \{b, a\}$ and that there is only one empty set, because any two sets without elements have the same elements and are therefore equal.

The most important non-trivial sets are: (1) The set $\{0, 1, 2, ...\}$ of natural numbers, denoted $\mathbb{N}$, (2) the set of rational numbers, denoted $\mathbb{Q}$, and (3) the set of real numbers, denoted $\mathbb{R}$. When we proceed deeper into set theory below we can actually *define* these sets, but let us take them for the moment as given.

We can form new sets from the ones we already know by means of set theoretic operations like the *union* $a \cup b$, *intersection* $a \cap b$, and *power set* $\mathcal{P}(a) = \{b : b \subseteq a\}$. There are a couple of others, and when one learns to use ordinals, there are the transfinite operations on sets.

Already with the simple operations $\cup$ and $\cap$ we get the following important concept: Let X be any set. Then obviously $\mathcal{P}(X)$ is closed under $\cup$ and $\cap$. Also, we can form the complement $-a = \{x \in X : x \notin a\}$ of any subset a of X. Finally, let us denote $\emptyset$ by 0 and X by 1. We have arrived at the structure

$$(\mathcal{P}(X), \cap, \cup, -, 0, 1)$$

which is a familiar algebraic structure, namely a *Boolean algebra*, because

it satisfies the identities

$a \cap (b \cap c)$	=	$(a \cap b) \cap c$	Associativity law
$a \cup (b \cup c)$	=	$(a \cup b) \cup c$	
$a \cap (b \cup c)$	=	$(a \cap b) \cup (a \cap c)$	Distributivity law
$a \cup (b \cap c)$	=	$(a \cup b) \cap (a \cup c)$	
$a \cap b$	=	$b \cap a$	Commutativity law
$a \cup b$	=	$b \cup a$	
$a \cup -a$	=	1	Law of complements
$a \cap -a$	=	0	
$a \cup (a \cap b)$	=	a	Absorption law
$a \cap (a \cup b)$	=	a	
$-(a \cap b)$	=	$-a \cup -b$	De Morgan law
$-(a \cup b)$	=	$-a \cap -b$	

These are all easy to prove, even by just looking at a picture, as in Figure 1.

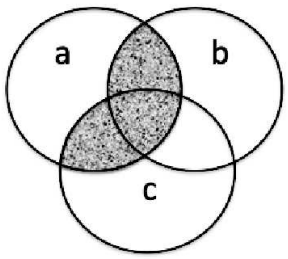


Figure 1: $a \cap (b \cup c) = (a \cap b) \cup (a \cap c)$.

An important role in applications of set theory is played by the concept of a an ordered pair (a, b) of two sets a and b. The characteristic property of ordered pairs is: $(a_0, a_1) = (b_0, b_1)$ if and only if $a_0 = b_0$ and $a_1 = b_1$. The cartesian product of two sets a and b is $a \times b = \{(x, y) : x \in a, y \in b\}$. It is the idea of set theory that everything is defined in terms of the sole primitive symbol $\in$. This is by no means necessary but since it is possible it is tempting and is usually done. The most common definition for the ordered pair (x, y) in terms of $\in$ is $\{\{x\}, \{x, y\}\}$.

A function from a set a to another set b is any subset f of $a \times b$ such that for each $x \in a$ there is exactly one $y \in b$ such that $(x, y) \in f$. Then

we write $f : a \to b$ and $y = f(x)$. In this definition of the concept of a function one notices a characteristic feature of set theory: the concept of a function is extremely general. We do not require that there is some "rule" which tells us how to compute $f(x)$ for a given x. All we require is that exactly one y such that $(x, y) \in f$ exists. Set theory uses classical logic so for a y such that $(x, y) \in f$ to exist it suffices that non-existence leads to a contradiction. There is also *constructive set theory* (see below) where intuitionistic logic is used and existence means more than deriving contradiction from non-existence.

A set a is *finite* if it is of the form $\{a_0, ..., a_{n-1}\}$ for some natural number n. This means that the set a has at most n elements. A set which is not finite is *infinite*. Finite sets have the following properties: $\emptyset$ is finite. If a and b are finite, then so is $a \cup b$. If a is finite and $b \subseteq a$, then also b is finite. If a and b are finite, then so is $a \times b$. If a is finite, then so is $\mathcal{P}(a)$.

With the above concepts one can already develop a lot of mathematics. One can define the integers as ordered pairs (n, m) of natural numbers with the intuitive meaning that (n, m) denotes the integer $n - m$. One can define the rationals as ordered pairs (r, q) of integers with the intuitive meaning that (r, q) denotes the rational r/q. One can define the reals as sets a of rationals, bounded from above, with the intuitive meaning that $a \subseteq \mathbb{Q}$ denotes the real $\sup(a)$.

3 Cardinal and ordinal numbers

A set is infinite if it is not of the form $\{a_1, ..., a_n\}$ for any natural number n. Set theory was developed to deal with problems of infinite sets and indeed there are some paradoxical phenomena related to infinite sets. A famous anecdotal example is *Hilbert's Hotel*: Imagine a hotel the rooms of which are numbered by all natural numbers. Suppose the hotel is full but a tourist comes in and asks for a free room. The reception can ask the person in room 0 to move to room 1, the person in room 1 to move to room 2, ..., the person in room n to move to room $n + 1$, etc. This process leaves room 0 empty and the tourist can take it. There are many further variations of this anecdote. For example, one can fit infinitely many new tourists into a hotel which is already full. A vast extension of this idea, coupled with the so called Axiom of Choice, is the Banach-Tarski Paradox: The unit sphere in three-dimensional space can be split into five pieces so that if the pieces are rigidly moved and rotated they suddenly form two spheres of the original size (see Picture 2). The trick is that the splitting exists only in

the abstract world of mathematics and can never actually materialize in the physical world. Conclusion: infinite abstract objects do not obey the rules we are used to among finite concrete objects. This is like the situation with sub-atomic elementary particles, where counter-intuitive phenomena, such as entanglement, occur.

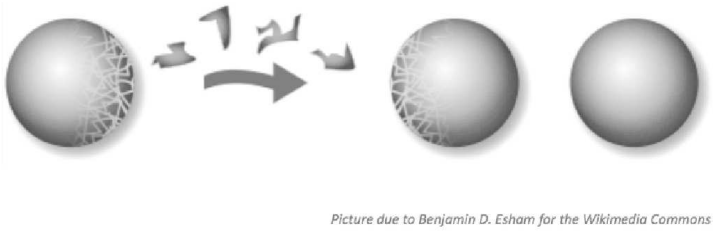


Figure 2: The Banach-Tarski Paradox.

3.1 Equipollence

Equipollence of two sets means the existence of a bijection between the sets. A bijection is a mapping which is both one-to-one and onto. In other words, a bijection between two sets a and b is a function $f : a \to b$ so that for every $y \in b$ there is a unique $x \in a$ such that $f(a) = b$. Still in other words the equipollence of a and b means the existence of functions $f : a \to b$ and $g : b \to a$ such that for all $x \in a$ we have $g(f(x)) = x$ and for all $y \in b$ we have $f(g(y)) = y$. In set theory it is thought that if two sets are equipollent, then they have the same number of elements. Because the sets may be infinite, it is not a priori clear what it means to say that the sets have the same number of elements. However, if there is a bijection between the sets, it is quite credible to argue that whatever we mean by the number of elements of an infinite set, equipollent sets should get the same number.

For finite sets equipollence means indeed that the sets have the same number of elements. For infinite sets we have to give up the idea that *the part is smaller than the whole*, since for example the set of natural numbers $\{0, 1, 2, ...\}$ is equipollent with its proper part $\{1, 2, 3, ...\}$, as the bijection $n \mapsto n + 1$ demonstrates. The part may not be smaller than the whole but at least it cannot be greater than the whole. And in some cases the part is

smaller than the whole. Cantor proved that the set $\mathbb{N}$ of natural numbers is not equipollent with the set $\mathbb{R}$ of real numbers. This can be seen as follows: Suppose there were a bijection $f : \mathbb{N} \to \mathbb{R}$. Then there is an onto function $g : \mathbb{N} \to [0, 1]$. Let us construct a real number on $[0, 1]$ as follows. The number is $0.d_1d_2d_3...$ where $d_i = 1$ if the real number $g(i)$ has the decimal expansion $0.e_1e_2e_3...$, where $e_i \neq 1$. Otherwise $d_i = 0$. In this way we obtain a real number $r \in [0, 1]$. Since g is onto, there is $n \in \mathbb{N}$ such that $g(n) = r$. Let us look at d_n. We have $d_n = 1$ if and only if $d_n \neq 1$, a contradiction. Hence no such f can exist. So $\mathbb{N}$ is less than the whole $\mathbb{R}$ in harmony with our intuition. This result is due to Cantor. He went on to prove that the set $\mathbb{Q}$ of all rational numbers is equipollent with $\mathbb{N}$ and hence not equipollent with $\mathbb{R}$. Moreover, he showed that the set $\mathbb{A}$ of all algebraic numbers is also equipollent with $\mathbb{N}$ and hence not equipollent with $\mathbb{R}$. We get the surprising conclusion that there are fewer algebraic numbers than real numbers, hence many (if not most) of the real numbers must be transcendental. This was a remarkable conclusion by Cantor because at the time when the observation was made, very few transcendental numbers were known. Thus by purely abstract set theoretic methods Cantor had proved the existence of many many transcendental numbers.

Technically speaking, a bijection between two sets a and b is a function $f : a \to b$ which is *one-one* i.e. $\forall x \in a \forall y \in a(f(x) = f(y) \to x = y)$ and *onto* i.e. $\forall y \in b \exists x \in a(f(a) = b)$. With this definition, sets a and b are *equipollent*, $a \sim b$, if there is a bijection $f : a \to b$. Then $f^{-1} : b \to a$ is a bijection and $b \sim a$ follows. The composition of two bijections is a bijection, whence

$$a \sim b \sim c \Longrightarrow a \sim v.$$

Thus $\sim$ divides sets into equivalence classes. Each equivalence class has a canonical representative (a cardinal number, see Subsection "Cardinals" below) which is called the *cardinality* of (each of) the sets in the class. The cardinality of a is denoted by $|a|$ and accordingly $a \sim b$ is often written $|a| = |b|$. One of the basic properties of equipollence is that if

$$a \sim c, b \sim d \text{ and } a \cap b = c \cap d = \emptyset,$$

then

$$a \cup b \sim c \cup d.$$

Indeed, if $f : a \to c$ is a bijection and $g : b \to d$ is a bijection, then $f \cup g : a \cup b \to c \cup d$ is a bijection. If the assumption $a \cap b = c \cap d = \emptyset$ is dropped, the conclusion fails, of course, as we can have $a \cap b = \emptyset$ and

$c = d$. It is also interesting to note that even if $a \cap b = c \cap d = \emptyset$, the assumption $a \cup b \sim c \cup d$ does not imply $b \sim d$ even if $a \sim c$ is assumed: Let $a = \mathbb{N}, b = \emptyset, c = \{2n : n \in \mathbb{N}\}$ and $d = \{2n + 1 : n \in \mathbb{N}\}$. However, for finite sets this holds: if $a \cup b$ is finite, $a \cup b \sim c \cup d, a \sim c, a \cap b = a \cap d = \emptyset$ then $b \sim d$. We can interpret this as follows: the cancellation law holds for finite numbers but does not hold for cardinal numbers of infinite sets.

A basic fact about equipollence, and indeed the starting point of all of set theory, is the result of Cantor that no set is equipollent with its power set. Let us see why this is so. Suppose a set a is equipollent with $\mathcal{P}(a)$. Thus there is a bijection $f : a \to \mathcal{P}(a)$. Let $b = \{x \in a : x \notin f(x)\}$. Then $b \in \mathcal{P}(a)$ so there is some $x \in a$ such that $b = f(x)$. Is x in b or not? If $x \in b$, then $x \notin f(x)$, a contradiction, since $f(x) = b$. Therefore we must conclude $x \notin b$. But then $x \notin f(x)$, whence $x \in b$, a contradiction again. So no such f can exist. It is remarkable that with this simple short argument one can make the far-reaching conclusion that there are an unending sequence of greater and greater cardinalities, namely one needs only follow the sets $\mathbb{N}$, $\mathcal{P}(\mathbb{N})$, $\mathcal{P}(\mathcal{P}(\mathbb{N}))$, $\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))$,...

There are many more interesting and non-trivial properties of equipollence that we cannot enter into here. For example the Schröder-Bernstein Theorem[1]: If $a \sim b$ and $b \subseteq c \subseteq a$, then $a \sim c$.

3.2 Countable sets

Countable sets are the most accessible infinite sets. They are the infinite sets that we can actually list, or rather, we can start listing a countable set and if we lived forever, we would list the entire set. So this is in sharp contrast to sets like $\mathbb{R}$, the set of all reals. Even if one lived forever, one could not list all real numbers. The quintessential example of a countable set is the set $\mathbb{N}$ of all natural numbers. Any set that is indexed by the natural numbers as $\{a_n : n \in \mathbb{N}\}$ is likewise called countable. And now we have already exhausted the class of countable sets! There are no others.

Countable sets already manifest the paradoxical feature of infinity that *the part need not be less than the whole*, for even the simplest countable set $\{0, 1, 2, ...\}$ is equipollent with its proper subset $\{1, 2, 3, ...\}$ via the bijection $n \mapsto n + 1$. By considering the bijection $n \mapsto 2n$ we can see that $\{0, 1, 2, ...\}$ is equipollent with the set of even numbers $\{0, 2, 4, 6, ...\}$. In fact, *all* infinite countable sets are equipollent: Suppose $A = \{a_n : n \in \mathbb{N}\}$ and $B = \{b_n : n \in \mathbb{N}\}$ are two infinite sets. Let $f(a_0) = b_0$. If $f(a_n) \in B$ has been defined, let

[1]The original formulation says: If there is a one-one function $a \to b$ and another $b \to a$ there is a bijection $a \to b$, see e.g. [12, p. 27].

$f(a_{n+1})$ be b_m with the smallest m such that $b_m \notin \{f(a_0), ..., f(a_n)\}$. Since B is infinite, such an m must always exist. Moreover, every b_m gets chosen at some point, for obviously $b_m \in \{f(a_0), ..., f(a_m)\}$.

Intuitively there are much more rational numbers than integers. Therefore it is a bit surprising that the set of all rational numbers is actually countable. Let us see how we can arrive at this conclusion. We can identify the rational number n/m (in lowest terms) with the ordered pair (n, m) of natural numbers. So let us first show that if a and b are countable, then so is $a \times b$. If either set is empty, the cartesian product is empty. So let us assume the sets are both non-empty. Suppose $a = \{a_0, a_1, ...\}$ and $b = \{b_0, b_1, ...\}$. Let

$$c_n = \begin{cases} (a_i, b_j), & \text{if } n = 2^i 3^j \\ (a_0, b_0), & \text{otherwise.} \end{cases}$$

Now $a \times b = \{c_n : n \in \mathbb{N}\}$, whence $a \times b$ is countable. So if we identify a rational number n/m (in lowest terms) with the pair (n, m), then there is some k such that $(n, m) = c_k$, and we have identified the set of all (non-negative) rational numbers with an infinite subset of $\mathbb{N}$, so in particular it is countable.

We showed above that the cartesian product of two countable sets if countable. A similar, and very useful fact is the following: a countable union of countable sets is countable. The empty sets do not contribute anything to the union, so let us assume all the sets in our countable family are non-empty. Suppose A_n is countable for each $n \in \mathbb{N}$, say, $A_n = \{a^n_m : m \in \mathbb{N}\}$ (we use here the Axiom of Choice to choose an enumeration for each A_n). Let $B = \bigcup_n a_n$. We want to represent B in the form $\{b_n : n \in \mathbb{N}\}$. If n is given, we consider two cases: If n is $2^i 3^j$ for some i and j, we let $b_n = a^i_j$. Otherwise we let $b_n = a^0_0$. Now indeed $B = \{b_n : n \in \mathbb{N}\}$.

One of the reasons why countable sets are so important is that sets defined by induction are usually countable. Examples of such sets are abundant in logic, most notably the set of terms and and the set of formulas in a countable vocabulary. Any formal language based on a countable vocabulary generates a countable set of expressions. More generally, in a countable vocabulary the set of all strings of symbols of a fixed finite length is countable, and hence so is the set of all finite strings of symbols, as it is the union of a countable family of countable sets.

A powerful application of the above idea is the Löwenheim-Skolem Theorem of first order logic: Every countable first order theory has a countable model. There are reasons to believe—although this view is also con-

tested[2]—that first order theories represent the best axiomatizations that we can ever get. Thus we are stuck with countable models whether we want it or not. For set theory this is called Skolem's Paradox. The paradox is that we can prove in set theory that the set of all reals is uncountable, but still set theory itself has countable models. That is the paradox. The solution of the paradox is that what seems countable from outside may not seem countable inside. More exactly, if we have a countable model of set theory, we can be sure that the mapping from the natural numbers onto the model is not an element of the model. This is a rough awakening to the reality that everything in set theory is relative. There are no signs that this would be the fault of set theory. It is even true of number theory vis a vis Gödel's Incompleteness Theorem.

3.3 Ordinals

The ordinal numbers introduced by Cantor are a marvelous general theory of measuring the *potentially infinite* on the one hand, and the actually infinity on the other hand. They are intimately related to inductive definitions and occur therefore widely in logic. It is easiest to understand ordinals in the context of games, although this was not Cantor's way. Suppose we have a game with two players I and II. It does not matter what the game is, but it could be something like chess. If II can force a win in n moves we say that the game has *rank* n. Suppose then II cannot force a win in n moves for any n, but after she has seen the first move of I, she can fix a number n and say that she can force a win in n moves. This situation is clearly different from being able to say in advance what n is. So we invent a symbol ω for the rank of this game. In a clear sense ω is greater than each n but there does not seem to be any possible rank between all the finite numbers n and ω. We can think of ω as an infinite number. However, there is nothing metaphysical about the infiniteness of ω. It just has infinitely many predecessors. We can think of ω as a tree T_ω with a root and a separate branch of length n for each n above the root as in the tree on the left in Figure 3.

Suppose then II is not able to declare after the first move how many moves she needs to beat I, but she knows how to play her first move in such a way that after I has played his second move, she can declare that she can win in n moves. We say that the game has rank $\omega + 1$ and agree that this is greater than ω but there is no rank between them. We can think of $\omega + 1$ as the tree which has a root and then above the root the tree T_ω,

[2]See e.g. [19].

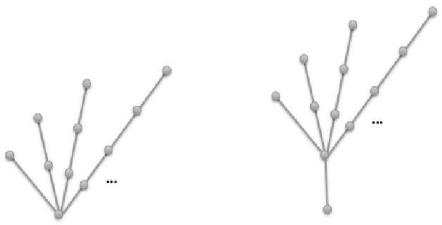


Figure 3: T_ω and $T_{\omega+1}$.

as in the tree on the right in Figure 3. We can go on like this and define the ranks $\omega + n$ for all n.

Suppose now the rank of the game is not any of the above ranks $\omega + n$, but still II can make an interesting declaration: she says that after the first move of I she can declare a number m so that after m moves she declares another number n and then in n moves she can force a win. We would say that the rank of the game is $\omega + \omega$. We can continue in this way defining ranks of games that are always finite but potentially infinite. These ranks are what set theorists call ordinals.

3.4 Cardinals

Historically cardinals (or more exactly cardinal numbers) are just representatives of equivalence classes of equipollence. Thus there is a cardinal number for countable sets, denoted $\aleph_0$, a cardinal number for the set of all reals, denoted $\mathfrak{c}$, and so on. There is some question as to what exactly are these cardinal numbers. The Axiom of Choice offers an easy answer, which is the prevailing one, as it says that every set can be well-ordered. Then we can let the cardinal number of a set be the order type of the smallest well-order equipollent with the set. Equivalently, the cardinal number of a set is the smallest ordinal equipollent with the set. If we leave aside the Axiom of Choice, some sets need not have have a cardinal number. However, as is customary in current set theory, let us indeed assume the Axiom of Choice. Then every set has a cardinal number and the cardinal numbers are ordinals, hence well-ordered. The α^{th} infinite cardinal number is denoted $\aleph_\alpha$. Thus $\aleph_1$ is the next in order of magnitude from $\aleph_0$. The famous *Continuum Hypothesis* is the statement that $\aleph_1 = \mathfrak{c}$. Equivalently, for every set A of reals, either A is countable of the cardinal number of A is

$\mathfrak{c}$. For Borel[3] sets of real numbers it is true that there is no cardinality between $\aleph_1$ and $\mathfrak{c}$. If we assume large cardinals[4], it is even true that for sets of reals definable with real parameters there is no cardinality between $\aleph_1$ and $\mathfrak{c}$. So it is not *so* far-fetched to suggest that maybe the same holds for *all* sets of reals. On the other hand, the tenet of set theory is that properties of *definable* sets are different from the properties of *arbitrary* sets. So maybe indeed the "regular" sets of reals—for some sense of "regular"—obey the Continuum Hypothesis but when we enter the absurd and unintuitive world of totally undefinable—arbitrary—sets of reals, the Continuum Hypothesis fails.

4 Axiomatic set theory

After the above tour of basic concepts of set theory we can return to the beginning and ask what is it that we are doing. This is all the more important because, as we have indicated, a lot of mathematics can be developed in set theory, if not all of mathematics. So the philosophical question arises, what is set theory based on? The most commonly held view is that set theory is the most fundamental theory in mathematics and it is not possible to base set theory on anything even more primitive.

So how do we really know what is true of sets and what is not? This question is crucially important also because most of the sets we encounter in set theory are infinite and unquestionably abstract. They seem to exist only in their own abstract world which cannot be seen by eyes, binoculars or microscopes, cannot be touched by hand, and cannot be observed by listening, tasting or smelling. It is often said that we can observe sets only by thinking of them, but this seems an inadequate answer. The most commonly held view is that we simply accept certain simple facts about sets as axioms and then use rules of logic to derive more complicated facts. The axioms are accepted because of their intuitive appeal and because of their usefulness. From the axioms that we present below one can derive virtually all of mathematics, and that is ultimately the most important reason for accepting them. They simply seem to give a "house" for mathematics to live in.

[3]The class of Borel sets is the smallest class of sets containing the open sets and closed under complements and countable unions, see [12, p. 132].

[4]Large cardinals are "large" cardinals that have special properties that are used in proofs. Their existence cannot be proved, so they have to be just assumed. However, they seem quite necessary in modern set theory, see e.g. [12, p. 275]

Technically speaking, the axioms are first order sentences in a vocabulary which has just one binary predicate symbol $\in$ in addition to identity.

The simple idea of sets as collections of objects is too loose in a closer analysis. This can be seen from the many paradoxes it has led to. The most important is *Russell's Paradox*: Consider the set R of sets that are not elements of themselves. If $R \in R$, then $R \notin R$, and if $R \notin R$, then $R \in R$. This paradox shows that we cannot allow just any collection to be a set. Current thinking is that sets are in a sense "small" enough collections to be considered as sets. According to this thinking, arbitrary collections are called *classes*. A class that is not a set is called a *proper class*.

In the axiomatic approach paradoxes like Russell's Paradox are avoided because sets and proper classes are kept away from each other. Technically speaking, objects in the axiomatic approach, that is, the range of all quantifiers, is sets. Classes are treated via formulas. A formula $\varphi(x)$, with perhaps parameters, is identified with the class $\{a : \varphi(a)\}$ of sets that satisfy $\varphi(x)$. So even if we think of our formulas as talking only about sets, we can talk about classes by talking about formulas defining the classes.

There is an intuitive model of set theory which goes beyond the simple idea that sets are "collections" of objects. According to this intuition sets have been created in stages. Elements of a set are, or have been, created before the set itself. This intuition does not mean that sets have really been created by someone, it is just a metaphor. The concept of an ordinal can be used to make the intuitive idea of stages more exact. The more exact version is called the *cumulative hierarchy* of sets. For this end, let $V_0 = \emptyset$, $V_{\alpha+1} = \mathcal{P}(V_\alpha)$ and $V_\nu = \bigcup_{\alpha<\nu} V_\alpha$ if ν is a limit ordinal. Finally, let $V = \bigcup_\alpha V_\alpha$. This is the intuitive model of set theory. Strictly speaking, it is not model in the sense of model theory because its domain is a proper class.

Now we present the axioms of set theory. They are called the Zermelo-Fraenkel axioms, denoted ZFC. When we discuss the axioms it is good to keep in mind the intuitive model offered by the cumulative hierarchy.

1. **Axiom of Extensionality:** Sets which have the same elements are equal i.e.

$$\forall x \forall y(\forall z(z \in x \leftrightarrow z \in y) \rightarrow x = y).$$

   This axiom seems obvious but it is actually a deep axiom. It demonstrates that we do not want there to be anything else about sets than their elements. The elements form an aggregate we call a "set" but we do not *care* what it is that pulls these elements together. The opposite attitude would be to think that there is much more to a set than

its elements, e.g. the *way*, whatever it means, how the elements are connected together into a set.

2. **Axiom of Pair:** From any two sets a and b we can form a new set $\{a, b\}$ which has exactly a and b as elements i.e.

$$\forall x \forall y \exists z \forall u (u \in z \leftrightarrow (u = x \vee u = y)).$$

Note that $\{a, b\}$ is *not* the union of a and b - however big sets a and b are the set $\{a, b\}$ has at most two elements, so in particular it is always finite. It is perfectly possible that $a = b$ and then $\{a, b\} = \{a\}$. We can form sets like $\{\mathbb{N}, \mathbb{Q}\}$, $\{\mathbb{Q}\}$ and $\{\mathbb{N}, \mathbb{Q}, \mathbb{R}\}$. Such sets are not particularly common or useful, but their existence in set theory is a manifestation of the basic tenet: whenever we have a set, we consider it as a "completed" totality, something we can use to build new sets.

3. **Axiom of Union:** For any set a we can form the *union* $\bigcup a$ of a, which consists of all sets which are elements of elements of a i.e

$$\forall x \exists y \forall z (z \in y \leftrightarrow \exists u (u \in x \wedge z \in u)).$$

Often sets are given in the form $a = \{a_i : i \in I\}$, that is, a is the range of the function $i \mapsto a_i$. Then $\bigcup a$ is the set $\bigcup_{i \in I} a_i$. This is a basic operation in mathematics and many applications of set theory.

4. **Axiom of Power set:** For any set a we can form the *power set* $\mathcal{P}(a)$ of a which consists of all sets which are subsets of a i.e

$$\forall x \exists y \forall z (z \in y \leftrightarrow \forall u (u \in z \rightarrow u \in x)).$$

One often hears criticism of this axiom but often also for a wrong reason. The problem with this axiom is *not* that it says that "all" subsets of a, whatever that means, exist. It says that those subsets which *do* exist can be collected together. The opposite of this axiom would be to think that some power sets are so large that they are proper classes. For example, we could think that, opposite to what the power set axioms says, the set of all reals, which is essentially the power set of $\mathbb{N}$, is a proper class. This is a coherent idea, but it does not mean that we have missed some subsets. We have all the subsets that we have, but we just cannot pull all of them together into a set. A smooth theory of the reals seems to require the power set axiom, but there are also alternative approaches.

5. **Axiom Schema of Subsets:** For any set a we can form a new set by taking the intersection of a and any class. In particular we can form new sets of the form $\{x \in a : \varphi(x)\}$ where $\varphi(x)$ is any formula. More exactly, for any formula $\varphi(x, \vec{y})$ we have the following axiom:

$$\forall x \forall x_1 ... \forall x_n \exists y \forall z (z \in y \leftrightarrow (z \in x \wedge \varphi(z, \vec{x}))).$$

Sets of the form $\{x \in a : \varphi(x)\}$ are very common in mathematics, for example $a \cap b = \{x \in a : x \in b\}$. Combined with the axioms of pair, union and power set, the Axiom of Subsets is very powerful indeed. This axiom has the impredicative element that the formula $\varphi(x)$ in $\{x \in a : \varphi(x)\}$ can have quantifiers and because these quantifiers range over the entire universe of sets the set $\{x \in a : \varphi(x)\}$ itself is also in the range of the quantifiers. We can remove this impredicativity by requiring that all quantifiers in $\varphi(x)$ are *bounded* i.e. of the form $\forall y \in z$ or $\exists y \in z$. However, this limits the applicability of the axiom seriously and leads to completely different kind of set theory, the so called Kripke-Platek set theory (see [4]).

6. **Axiom Schema of Replacement:** Suppose a is a set. If there is a way to associate to every element i of a a new set a_i, then we can form a new set $\{a_i : i \in a\}$, that is, a set which has all the a_i, where $i \in a$, as elements, and nothing else. More exactly, for any formula $\varphi(x, \vec{y})$ we have the following axiom:

$$\begin{aligned}\forall x \forall x_1 ... \forall x_n (\forall u \forall z \forall z' ((u \in x \wedge \varphi(u, z, \vec{x}) \wedge \varphi(u, z', \vec{x})) \rightarrow z = z') \\ \rightarrow \exists y \forall z (z \in y \leftrightarrow \exists u (u \in x \wedge \varphi(u, z, \vec{x})))).\end{aligned}$$

This axiom introduced by Fraenkel is needed e.g. in transfinite recursion.

7. **Axiom of Infinity:** This axiom simply says that there is an infinite set. More exactly,

$$\exists x (\exists y (y \in x \wedge \forall z \neg (z \in y)) \wedge \forall y (y \in x \rightarrow \exists z (y \in z \wedge z \in x))).$$

There are many ways to write this axiom, all equivalent, given the other axioms. The particular formulation here yields the set $A = \{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, ...\}$. It is easy to see on the basis of the Axiom of Extensionality that all elements of this set A are different.

8. **Axiom of Foundation:** This axiom says that every set has an element which is minimal with respect to $\in$, that is

$$\forall x \exists y (x \cap y = \emptyset).$$

> This is the most useless axiom (of set theory) that anyone ever invented. In fact there are reasons to claim that no-one ever used this axiom! However, since the intuitive idea of sets is that they were "created" in stages, with elements of a set having been created before the set itself, then of course every set has an $\in$-minimal element, namely the one that was 'created first". Since we do not really think sets were created—creation being a mere metaphor—there is hardly any mathematical example where this axioms turns up. Set theorists count it in for their internal aesthetic reasons. Its usefulness is not based in what it gives but rather in that we can live without the circular sets it excludes.

5 Axiom of Choice

The Axiom of Choice is one of the axioms of set theory but we treat it here separately from the others because it is of a slightly different character. The Axiom of Choice states that if a set a of non-empty sets is given, then there is a function f such that $f(x) \in x$ for all $x \in a$. That is, the function f picks one element from each of those non-empty sets. There are so many equivalent formulations of this axiom that books have been written about it. The most notable is the Well-Ordering Principle: every set is equipollent with an ordinal (see e.g. [12, p. 45]).

The Axiom of Choice is the only axiom of ZFC which brings arbitrariness or abstractness into set theory, often with examples that can be justifiably called pathological, like the Banach-Tarski Paradox (see above). Every other axiom states the existence of some set and specifies what the set is. The Axiom of Union says the new set is the union $\bigcup_{i\in A} B_i$, the Axiom of Power Set says the new set is the powerset $\{B : B \subseteq A\}$, the Axiom of Subsets states that the new set is of the form $\{b \in a : \varphi(b)\}$.

Because of the abstractness brought about by the Axiom of Choice it has received criticism and some authors always mention explicitly if they use it in their work. The main problem in working without the Axiom of Choice is that there is no clear alternative and just leaving it out leaves many areas of mathematics, like measure theory, without proper foundation.

A basic problem with an axiom like the Axiom of Choice is that it has a formulations which are rather obvious, like the formulation above, and equivalent formulations which are completely unbelievable, like the Well-Ordering Principle. If one thinks of formulations that make it look obvious, one would like to accept it, but when one looks at the unbelievable conse-

quences one would like to reject it. So which way to go?

It is sometimes wrongly believed that the problem of the Axiom of Choice is in that no-one knows which element to choose from each non-empty set. This is not the point. If a set a is non-empty, i.e. it is not the case that every set is not in a, then by the laws of logic there must be a set in a. This does not require the Axiom of Choice as it is simply a consequence of provability of $\neg\forall x\neg A \rightarrow \exists xA$. The problem is how to make *infinitely* many such choices.

6 Independence results

In set theory it is relatively easy to formulate questions that have turned out to be impossible to decide on the basis of the axioms. The most famous of these is the Continuum Hypothesis, already proposed by Cantor. The Continuum Hypothesis claims that every uncountable set of reals is equipollent with the entire set of reals $\mathbb{R}$ (see discussion on Continuum Hypothesis in Section 3.4).

The undecidability of a sentence on the basis of any axioms, set theory or not, can be proved by producing two models of the axioms, one where the sentence is true and another where it is false. In the case of the Continuum Hypothesis such two models have indeed been produced (see e.g. [12, chapters 13 and 14]). The two models, one due to Kurt Gödel and the other due to Paul Cohen, have led to an extensive study of models of set theory, and a profusion of different kinds of models have been uncovered. Most of these models are constructed by a method called *forcing*. This highly interesting method has turned out to be of relevance also outside set theory.

The basic idea of forcing is that instead of trying to build directly a model where something we are interested in is true, we settle with something less. We settle with contemplating what finite pieces of information, called *conditions*, "force" to be true, if ever a model based on them was constructed. For example, if we have a name $\dot{A}$ for a set of natural numbers, then the condition $\{0 \in \dot{A}, 1 \notin \dot{A}\}$ forces $\dot{A}$ to contain 0 but not 1, and this condition leaves it open whether e.g. 2 is in $\dot{A}$ or not. We form a particular infinite sequence of conditions called a generic sequence and build a model, called a generic model from that sequence. Remarkably, a sentence is true in the generic model if and only if some condition in the generic sequence forces it to be true. This can be done in such a manner that the Continuum Hypothesis is forced to be true or false in the generic model according to our

will. If we want the Continuum Hypothesis to be true we use one kind of condition and if we want it to be false we use another kind of condition. For more on forcing see [12, Chapter 14] and [15, Chapter VII].

Forcing has turned out to have a connection to both modal and intuitionistic logic. This connection arises from the fact that we can think of the set of forcing conditions as the frame of a Kripke structure. For example, a condition p is said to force $\neg\varphi$ if and only if no extension of p forces φ. This is exactly the same as the definition of the truth of a negated sentence of intuitionistic logic at a node of a Kripke structure.

The philosophical importance of forcing is manifold. It represents a useful *weak* truth definition, and as such one which can be used in different parts of philosophical logic. It uncovers a huge gap in what the axioms of set theory decide leading to the philosophical question, whether there is ultimately any true universe of mathematical objects. Skeptics say that Gödel's Incompleteness Theorem casts a doubt on the existence of mathematical objects, and Cohen's forcing, especially the independence of the Continuum Hypothesis, was the last blow which to many people totally shattered the idea of a platonist reality of mathematics. The opposite view is that mathematical objects form a definite unique reality of their own and the results of Gödel and Cohen merely manifest an inherent underdetermination of the axioms of set theory in uncovering what is true in this invisible world and what is not.

7 Some recent work

7.1 Descriptive Set Theory

A set A is said to be *definable* if there is a formula $\varphi(x)$ such that A is the set of sets b that satisfy $\varphi(x)$. Since there are only countably many formulas there can be only countably many definable sets. However, if we allow parameters, we get more definable sets. Typical parameters that are sometimes allowed are on the one hand ordinal numbers and on the other hand real numbers. Descriptive Set Theory is an important sector of set theory which concentrates on sets that are definable with real parameters. The basic ideology is that the arbitrariness or pathology brought by the Axiom of Choice is only manifested in the realm of undefinable sets. The sets we actually work with are a fortiori definable—otherwise we could not talk about them! Seminal results of Martin-Steel-Woodin ([17]) show that assuming so called large cardinals, phenomena like the Banach-Tarski Paradox do not occur among definable sets. In other words, large cardinals

remove the negative effect of arbitrariness that the Axiom of Choice brings to set theory. The abstract arbitrary sets are there, and are needed for the general theory, but they do not disturb the world of definable sets with their paradoxical counter-intuitive properties. Current work in Descriptive Set Theory further emphasizes this and at the same time brings set theory closer and closer to classical analysis, topology and measure theory (see e.g., the paper [5]).

7.2 Non well-founded set theory

The Non-well-founded set theory of Peter Aczel ([2]) takes on the empirical fact that the Axiom of Foundation is not really a necessary axiom. So non-well-founded set theory replaces the Axiom of Foundation with its ultimate strongest possible denial: any combination of circularity in the $\in$-relation is manifested by some sets. Circularity comes up naturally in computer science: the state of a program may very well come back to itself. Of course, the common sense view is that then the program is in a loop and can be "dismissed" as a program with a bug. However, another common sense view is that most programs can enter a loop, and some programs, like operating systems, are even expected to come back to the same state time after time. It has turned out that non-well-founded set theory can be used to model conveniently processes in computer science (see e.g., the paper [3]).

7.3 Constructive set theory

Constructive set theory drops classical logic from set theory. As a result, $\neg\forall x\neg\varphi(x)$ is not anymore a guarantee for $\exists x\varphi(x)$. For us to assert $\exists x\varphi(x)$ we have to have a construction of an x and a proof that $\varphi(x)$. At first sight this seems to have devastating consequences for set theory. However, if we just adopt constructive logic but do not change the axioms we do not gain much ([11]). To really make a difference in the direction of constructive mathematics, one has to rethink the axioms. One approach gaining popularity is the *Constructive Zermelo Fraenkel Set Theory CZF* (see [1]). The goal of CZF is to offer a simple intuitive foundation for constructive mathematics in the same way as ZFC offers one for classical mathematics.

8 Historical Remarks and Further Reading

Set theory was launched by Georg Cantor (see [6] and [7]) in 1874. There are many elementary books providing an introduction to set theory, for example [8], [9], [18], [16]. Textbooks covering a wide spectrum of modern set theory are [13] and [14]. A colossal recent source of advanced set theory is [10].

References

[1] Peter Aczel. The type theoretic interpretation of constructive set theory. In *Logic Colloquium '77 (Proc. Conf., Wrocław, 1977)*, volume 96 of *Stud. Logic Foundations Math.*, pages 55–66. North-Holland, Amsterdam, 1978.

[2] Peter Aczel. *Non-well-founded sets*, volume 14 of *CSLI Lecture Notes*. Stanford University Center for the Study of Language and Information, Stanford, CA, 1988. With a foreword by Jon Barwise [K. Jon Barwise].

[3] Peter Aczel. Final universes of processes. In *Mathematical foundations of programming semantics (New Orleans, LA, 1993)*, volume 802 of *Lecture Notes in Comput. Sci.*, pages 1–28. Springer, Berlin, 1994.

[4] Jon Barwise. *Admissible sets and structures*. Springer-Verlag, Berlin, 1975. An approach to definability theory, Perspectives in Mathematical Logic.

[5] Howard Becker and Alexander S. Kechris. *The descriptive set theory of Polish group actions*, volume 232 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1996.

[6] Georg Cantor. *Contributions to the founding of the theory of transfinite numbers*. Dover Publications Inc., New York, N. Y., 1952. Translated, and provided with an introduction and notes, by Philip E. B. Jourdain.

[7] Joseph Warren Dauben. *Georg Cantor*. Princeton University Press, Princeton, NJ, 1990. His mathematics and philosophy of the infinite.

[8] Keith Devlin. *The joy of sets*. Springer-Verlag, New York, second edition, 1993. Fundamentals of contemporary set theory.

[9] Herbert B. Enderton. *Elements of set theory*. Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1977.

[10] Matthew Foreman and Akihiro Kanamori (Eds.). *Handbook of Set Theory*. Springer-Verlag, Berlin, 2010.

[11] Harvey Friedman. The consistency of classical set theory relative to a set theory with intuitionistic logic. *J. Symbolic Logic*, 38:315–319, 1973.

[12] Thomas Jech. *Set theory*. Perspectives in Mathematical Logic. Springer-Verlag, Berlin, second edition, 1997.

[13] Thomas Jech. *Set theory*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2003. The third millennium edition, revised and expanded.

[14] Akihiro Kanamori. *The higher infinite*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, second edition, 2003. Large cardinals in set theory from their beginnings.

[15] Kenneth Kunen. *Set theory*. North-Holland Publishing Co., Amsterdam, 1980. An introduction to independence proofs.

[16] Kenneth Kunen. *Set theory*, volume 102 of *Studies in Logic and the Foundations of Mathematics*. North-Holland Publishing Co., Amsterdam, 1983. An introduction to independence proofs, Reprint of the 1980 original.

[17] Donald A. Martin and John R. Steel. A proof of projective determinacy. *J. Amer. Math. Soc.*, 2(1):71–125, 1989.

[18] B. Rotman and G. T. Kneebone. *The theory of sets and transfinite numbers*. Oldbourne, London, 1966.

[19] Stewart Shapiro. The "triumph" of first-order languages. In *Logic, meaning and computation*, volume 305 of *Synthese Lib.*, pages 219–259. Kluwer Acad. Publ., Dordrecht, 2001.

Understanding, Formal Verification, and the Philosophy of Mathematics

JEREMY AVIGAD [*†]

Abstract

The philosophy of mathematics has long been concerned with determining the means that are appropriate for justifying claims of mathematical knowledge, and the metaphysical considerations that render them so. But, as of late, many philosophers have called attention to the fact that a much broader range of normative judgments arise in ordinary mathematical practice; for example, questions can be interesting, theorems important, proofs explanatory, concepts powerful, and so on. The associated values are often loosely classified as aspects of "mathematical understanding."

Meanwhile, in a branch of computer science known as "formal verification," the practice of interactive theorem proving has given rise to software tools and systems designed to support the development of complex formal axiomatic proofs. Such efforts require one to develop models of mathematical language and inference that are more robust than the the simple foundational models of the last century. This essay explores some of the insights that emerge from this work, and some of the ways that these insights can inform, and be informed by, philosophical theories of mathematical understanding.

[*]Professor in the Department of Philosophy and the Department of Mathematical Sciences, Carnegie Mellon University

[†]This essay is based on a talk presented in Paris in January, 2010, to a fellows seminar in Mic Detlefsen's *Ideals of Proof* project; and in February, 2010, in the University of Pittsburgh's *Center for Philosophy of Science* colloquium series. I am grateful to members of both audiences for comments and discussion after the talk. Comments from Anil Gupta, Tom Ricketts, and Mark Wilson in particular have influenced the presentation in Section "The philosophy of mathematics". Section "Mathematical ease and difficulty" incorporates material I presented in other workshops in Paris in May of 2009 and January of 2010, where I received helpful comments from Andrew Arana, Alan Baker, Karine Chemla, Evelyn Fox Keller, Paolo Mancosu, and Marco Panza, among others. Discussions with Penelope Maddy have also influenced the writing of that section. Finally, I am grateful to an anonymous referee for corrections and helpful suggestions. This work has been partially supported by NSF grant DMS-0700174 and a grant from the John Templeton Foundation.

1 Introduction

Since Plato and Aristotle, the philosophy of mathematics has been concerned with clarifying the nature of mathematical objects, and determining the appropriate means of justifying claims of mathematical knowledge. But in our daily mathematical practice, we often employ normative judgments that go beyond assessments of justification and correctness. For example, mathematical questions can be interesting, or not; questions can be natural; concepts can be fruitful or powerful; some proofs provide better explanations than others; some historical developments are important; and some observations are insightful. Even though our ways of expressing these judgments are often vague and imprecise, the evaluations matter to us a great deal. They bear on the kinds of mathematics we decide to do and they way we go about doing it, they way we teach and communicate mathematics, and the kinds of mathematics we praise and support. In other words, such judgments amount to normative assessments in one of our core scientific practices, and so deserve philosophical attention.

Intuitively, what unifies these kinds of judgments is that they evaluate the extent to which pieces of mathematics—concepts, proofs, questions, conjectures, theories, and so on—contribute to our *understanding*. This last word is often used to invoke an aura of mystery and ineffability, suggesting that once we look beyond well-worn questions of justification our modes of hard-nosed philosophical analysis break down entirely, leaving us with nothing to do but shake our heads in wonder. The point of this essay is to argue to the contrary. Specifically, I will consider some recent scientific advances in the formal modeling of mathematical reasoning and proof, and argue that these advances give us some leverage in making sense of such normative assessments.

The outline of this paper is as follows. In Section 2, I briefly explore some of our (vague) intuitions as to what we are talking about when we talk about mathematical understanding. In Section 3, I shift gears and discuss some technical developments in a branch of computer science known as "formal verification." More specifically, I will discuss efforts in interactive theorem proving, which involves the use of computational proof assistants to construct complex mathematical proofs. I will describe some of the ways the field has been forced to model mathematical language and inference and, in each case, I will consider what these efforts have to tell us about mathematical understanding. Finally, in Section 4, I will try to bring the vague intuitions and the technical work closer together, and explore some of the ways that philosophical work can inform, and be informed by,

the technical developments. In particular, I will consider the ways that better a philosophical understanding of mathematical methods and abilities, mathematical concepts, and mathematical ease and difficulty can both help us come to terms with the kinds of evaluations expressed above, and support scientific work in fields that rely implicitly on notions of mathematical understanding.

I am by no means the only one now trying to make sense of normative assessments in ordinary mathematical practice; see, for example, the collection of essays *The Philosophy of Mathematical Practice* [21] and the references there for an overview of some of the recent work in the area. This essay draws on and expands prior writings of my own, including [2, 3].

2 Understanding

Section 1 drew a distinction between traditional philosophical concerns regarding correctness and justification on the one hand, and a broader class of normative assessments on the other. I will now pose two philosophical "problems" that help make some of the issues salient. The first, which I will call "the problem of multiple proofs," goes like this. On the standard account, the value of a mathematical proof is that it warrants the truth of the resulting theorem. Why, then, is it often the case that a new proof of a theorem is often highly valued? For example, Furstenberg's ergodic-theoretic proof [8] of Szemerédi's theorem [29] is recognized as a seminal element of ergodic Ramsey theory, and Tao [30] notes that a host of proofs of the theorem have been published since then, each one providing new insights. Clearly the proof of a theorem does *something* more than establish that the final result is true; can we say, in precise terms, what that something is?[1]

The second philosophical problem, which I will call "the problem of conceptual possibility," is as follows. It is often said that some mathematical advance was "made possible" by a prior conceptual development. For example, Riemann's introduction of the complex zeta function and the use of complex analysis made it possible for Hadamard and de la Vallée Poussin to prove the prime number theorem in 1896. What is the sense of

[1] The question may bring to mind Georg Kreisel's "unwinding program," which involves the use of formal methods to extract additional information from mathematical proofs. This has become a fruitful and active branch of proof theory, which now generally goes by the name of "proof mining" (see [6, 19]). With regard to the evaluation of informal proofs, information that is "implicit" in such proofs is certainly an important part of the story; see, for example, the discussion in [2].

"possibility" here? It is certainly not physical possibility, that is, the claim that someone like Chebyshev was physically incapable of writing down a proof in 1850. And it is not hard to make the case that nor is it a matter of logical possibility, that is, the fact that Chebyshev's axioms and principles of inference were not strong enough to entail the desired conclusion (see, for example, [1]). An intuitive answer is that Chebyshev did not have the right definitions in place, but that just pushes the problem back to explaining why he could not have written down those definitions. In other words, answering the question requires us to adopt a viewpoint in which writing down a good definition can be a hard thing to do.

In both cases, the answer seems to have something to do with understanding: new proofs provide us with a better understanding of the theorem they prove, as well as the surrounding mathematics; and historical developments provide us with understanding that supports further advances. Indeed, informal talk about mathematical understanding arises in a number of scientific and academic pursuits. Educational research aims to determine the ways of communicating understanding to students efficiently and effectively; psychology and cognitive science aim to explain how subjects acquire mathematical understanding; the history of mathematics focuses on events that have have furthered our understanding; in formal verification and automated reasoning, one tries to get computers to understand the mathematics we give them.

The purpose of this essay is to explore the prospect of developing a philosophical theory that can help us come to terms with these various notions of understanding. In order to avoid misunderstandings, I would like to make three points clear.

First, I am not claiming originality or priority in raising these issues. For example, one can find the problem of multiple proofs neatly expressed in the writings of Wittgenstein [37, III–60]:

> It might be said: "—that every proof, even of a proposition which has already been proved, is a contribution to mathematics". But why is it a contribution if its only point was to prove the proposition? Well, one can say: "the new proof shews (or *makes*) a new connexion"

A number of people working in the philosophy of mathematics today have come at these issues in various ways, and it is not possible for me to provide an adequate survey of such work here. For further reading, see the suggestion at the end of Section 1.

Second, I would like to emphasize that the "problems" I have raised are *not* great mysteries, set out only for us to marvel at how deeply inscrutable mathematical understanding is. On the contrary, we have a lot of good intuitions to start with, and it is easy to begin enumerating reasons why we might prefer one proof to another, or why we might value a historical development. The point is simply that, until recently, these issues have not received serious philosophical attention, and so the language we use to discuss them is still vague and imprecise. The challenge is to sharpen our intuitions so that they can better support rational discussion and scientific inquiry.

Finally, let us not get too hung up on the word "understanding." In ordinary usage, the word has social and even moral connotations; for example, we praise children for understanding and hold criminals liable only insofar as they have understood the consequences of their actions. Here I am only concerned with much more focused issues having to do with the methodology of mathematics. I have invoked the word "understanding" because it is often used in our informal talk about these issues, but what I am arguing for is the importance and promise of a certain type of philosophical analysis, rather than an exhaustive and univocal analysis of the notion of understanding as it applies in every domain. I do not mind if you prefer to characterize the project I am describing here as developing a theory of "mathematical values," "mathematical competence," "mathematical ability," or something of that sort. In short, I wish to focus on the phenomena, not the word.

Let us begin with some straightforward observations. Mathematics is hard; mathematical solutions, proofs, and calculations involve long sequences of steps, that have to be chosen and composed in precise ways. The problem is not that there are too few options, but too many. For example, at each stage in a deduction or calculation there are arbitrarily many facts we can interpolate from our background knowledge ($2+2=4$, $4+4=8$, . . .), most of which will be no help at all. From among all the options available to us, we have to settle one initial step that may plausibly take us closer to our goal, and then another, and then another. To compound matters, even the best among us have limited cognitive capacities; we can only keep so many pieces of information in mind at one time, and anticipate only a small number of consequences of the definitions and facts before us. It should strike you as something of a miracle that we are able to do mathematics at all. And yet, somehow, we are; being mathematically competent means being able to proceed reasonably, if imperfectly, under these circumstances. What I would like to understand are the complex mechanisms that make it

possible for us to do so.

One way of posing the challenge is to note that whereas logic and traditional foundational research aims to determine what is *allowed* in a mathematical argument or calculation, this falls short of determining which steps are *appropriate*, or likely to be *fruitful*, in a given situation. This distinction was neatly expressed by Poincaré in 1908, in *Science et méthod* [24, Book II, Chapter II][2]:

> Logic teaches us that on such and such a road we are sure of not meeting an obstacle; it does not tell us which is the road that leads to the desired end.

In other words, logic tells us how to verify that a proof of a given theorem or the solution to a given problem is correct, but it does not tell us how to find such a solution or proof in the first place. Something more is needed to explain how we find manage to select a fruitful path from among a bewildering array of useless options:

> Discovery consists precisely in not constructing useless combinations, but in constructing those that are useful, which are an infinitely small minority. Discovery is discernment, selection.

While the image of finding a selecting a path towards our goals provides a helpful metaphor, I find literary metaphors helpful as well. For example, Herman Melville's *Moby Dick* is largely a story of humankind's attempts to come to terms with a chaotic and indifferent universe; this image accords well with mathematics since, after all, mathematics doesn't really care whether we understand it or not. One of the most difficult aspects of doing mathematics is sitting down to a blank sheet of paper, and trying to figure out where to begin. Blankness, as a metaphor, comes up often in *Moby Dick*; for example, in the final pages, the great white whale presents a "blank forehead" to the ship. The following passage is taken from an entire chapter, Chapter 42, devoted to a discussion of the color white.

> But not yet have we solved the incantation of this whiteness, and learned why it appeals with such power to the soul; and more strange and far more portentous... and yet should be as it is, the intensifying agent in things the most appalling to mankind.

[2]The next two passages are also quoted in [3].

> Is it that by its indefiniteness it shadows forth the heartless voids and immensities of the universe, and thus stabs us from behind with the thought of annihilation, when beholding the white depths of the milky way? Or is it, that as in essence whiteness is not so much a colour as the visible absence of colour; and at the same time the concrete of all colours; is it for these reasons that there is such a dumb blankness, full of meaning, in a wide landscape of snows–a colourless, all-colour of atheism from which we shrink?

Melville also offers us a grand and eloquent account of what happens when we get an unfiltered glimpse of the infinity of possibilities before us, with the story of Pip, one of the ship's deck hands, who falls overboard while his fellow shipmates sail off in chase of a whale.

> The sea had jeeringly kept his finite body up, but drowned the infinite of his soul. Not drowned entirely, though. Rather carried down alive to wondrous depths, where strange shapes of the unwarped primal world glided to and fro before his passive eyes; and the miser-merman, Wisdom, revealed his hoarded heaps; and among the joyous, heartless, ever-juvenile eternities, Pip saw the multitudinous, God-omnipresent, coral insects, that out of the firmament of waters heaved the colossal orbs. He saw God's foot upon the treadle of the loom, and spoke it; and therefore his shipmates called him mad. So man's insanity is heaven's sense; and wandering from all mortal reason, man comes at last to that celestial thought, which, to reason, is absurd and frantic; and weal or woe, feels then uncompromised, indifferent as his God.

These passages give us a good sense of what we are up against. Vast and complex, mathematics offers us great riches, but, at the same time, threatens to overwhelm us. A theory of mathematical understanding should explain how we cope with the complexity and maintain our sanity while exploring the wonders before us.

3 Formal verification

The phrase "formal verification" refers to a branch of computer science which uses formal methods to verify correctness. This can mean verifying the correctness of hardware and software design, for example, to ensure

that a circuit description, an algorithm, or a network or security protocol meets its specification. But it can also mean verifying that a proof of a mathematical theorem is correct. There is a lot of overlap between these two pursuits, but also a number of differences in emphasis. Here I will focus on the latter type of verification.

"Interactive theorem proving" provides on important approach. Working with an interactive proof assistant, users enter enough information for the system to confirm that there is a formal axiomatic proof of the theorem that the user has asserted. In fact, many systems enable one to extract a formal proof object—a complex piece of data representing a fully detailed axiomatic proof—which can be manipulated and verified independently of the system that constructed it.

There are a number of such systems currently in use; those in which substantial portions of mathematics have been formalized include Mizar, HOL, HOL light, Isabelle, Coq, and ACL2 (see [35] for an overview). The technology is still young, and it will be a while before such systems are commonly used in mathematical circles. But initial achievements make it clear that the potential is there. Notable theorems of mathematics that have been formalized to date include the four-color theorem [10], the prime number theorem [5, 16], Dirchlet's theorem on primes in an arithmetic progression [15], and the Jordan curve theorem [13]. At the time of writing of this article, two very ambitious projects are well underway: Thomas Hales is heading a project [14] to verify a proof of the Kepler conjecture, which asserts that there is no way of filling space with spheres that can beat the density of the familiar lattice packing; and Georges Gonthier is heading a similar project [12] to verify the Feit-Thompson theorem, which asserts that every finite group of odd order is solvable. Once again, I do not have sufficient space to provide an adequate overview of the field and its history; a good starting point for that is the December 2008 issue of the *Notices of the American Mathematical Society*, a special issue on formal proof, with articles by Hales, Freek Wiedijk, John Harrison, and Gonthier. My goal here is to consider some of the issues that arise with respect to formal verification, and what they have to tell us about mathematical understanding.

3.1 Understanding mathematical language

To start with, an interactive proof system relies on an underlying formal framework, which specifies a language in which assertions are to be expressed and the admissible rules of inference. There are a number of frameworks currently in use. Zermelo-Fraenkel set theory has long been

recognized as a powerful foundational framework for mathematics, and, for example, the Mizar system uses a variant thereof. In the language of set theory, everything is a set; but one introduces definitions that allow one to recognize some sets as being natural numbers, some as being real numbers, some as being functions from the natural numbers to the reals, and so on. Other proof systems, in contrast, use frameworks that take such "typing" information to be built into the basic language. For example, HOL, HOL light, and Isabelle use a formulation of higher-order logic in Church's simple type theory, in which every term is assigned such a type. What makes simple type theory "simple" is that the type of a variable cannot depend on a parameter. On the other hand, in many mathematical contexts, it is natural to let a variable x stand for an element of the vector space $\mathbb{R}^n$, where n is another variable ranging over the natural numbers. Some systems, like Coq, use a more elaborate type theory, where $\mathbb{R}^n$ can be represented as a type. Adding rules to manipulate these more elaborate complicates the underlying logical framework. But, as we will see below, this kind of information is fundamental to mathematical reasoning, and making it part of the underlying logical framework means that one can build general-purpose mechanisms to handle such information into the core of the system itself. (Coq is moreover based on a constructive logic, and computational aspects of the mathematics in question play a central role in the formalization process.)

As an example of how mathematics is expressed in such systems, here is Hales' statement of the Jordan curve theorem in HOL light:

```
!C. simple_closed_curve top2 C ==>
  (?A B. top2 A /\ top2 B /\
    connected top2 A /\ connected top2 B /\
  ~(A = EMPTY) /\ ~(B = EMPTY) /\
   (A INTER B = EMPTY) /\ (A INTER C = EMPTY) /\
       (B INTER C = EMPTY) /\
      (A UNION B UNION C = euclid 2)
```

Here, the exclamation point denotes a universal quantifier, and the question mark denotes an existential quantifier. The predicate "top2" denotes the standard topology on the Euclidean plane. In ordinary mathematical language, the expression above asserts that if C is a simple closed curve in the Euclidean plane, then the entire plane can be written as a disjoint union $A \cup B \cup C$, where A and B are connected open sets.

While one can get used to such typography and notation, it is less pleasant than reading ordinary mathematical text. As part of his MS thesis work

at Carnegie Mellon, Steve Kieffer implemented a parser for an extension of set theory designed by Harvey Friedman, and entered hundreds of definitions from Suppes' *Set theory* and Munkres' *Topology* (see [18]). For example, here is his rendering of Munkres' definition of the topology X generated by a basis $\mathscr{B}$:

```
DEFINITION MunkTop.13.2: 2-ary function Basisgentop.
If TOPBASIS[\mathscr{B},X] then
Basisgentop(\mathscr{B},X) \simeq
  (!\mathscr{T} \subseteq \wp(X))(
  (\forall U \subseteq X)(U \in \mathscr{T} \iff
    (\forall x \in U)(\exists B \in \mathscr{B})(
      x \in B \wedge B \subseteq U))).
```

And here is his definition of a certain topology, the K-topology, on the real numbers:

```
DEFINITION MunkTop.13.3.c: 0-ary function Krealtop.
Krealtop \simeq Basisgentop(
  Stdrealtopbasis \cup
  {V \subseteq \mathbb{R} :
    (\exists W \in Stdrealtopbasis)(
      V = W \less {Incl_{FrR}(1_{N}/n) :
              n \in \mathbb{N}})}, \mathbb{R}).
```

These may not look like much, but they do come close to the structure of ordinary mathematical language. To make this point, Kieffer added a feature which allows the user to specify natural-language equivalents for the symbols in the language, and implemented a simple heuristic to determine when to use symbols or the expanded language. With these in place, the definitions above were rendered as follows:

Definition: If $\mathscr{B}$ is a basis for a topology on X then *the topology on X generated by $\mathscr{B}$* is the unique $\mathscr{T} \subseteq \wp(X)$ such that for every $U \subseteq X$, $U \in \mathscr{T}$ if and only if for every $x \in U$, there exists $B \in \mathscr{B}$ with $x \in B$ and $B \subseteq U$.

Definition: *The K-topology on* $\mathbb{R}$ is the topology on $\mathbb{R}$ generated by the standard basis for a topology on $\mathbb{R}$ union the set of $V \subseteq \mathbb{R}$ such that there exists W in the standard basis for a topology on $\mathbb{R}$ such that $V = W \setminus \{1/n : n \in \mathbb{N}\}$.

The prose is not literary and tends to have a run-on feel, but it is not terribly far from ordinary mathematical text. What this seems to suggest is that our conventional modeling of mathematical language is on the right track. And insofar as this modeling captures the structure of mathematical language, it follows that when we read and write mathematical assertions, we are implicitly able to recognize and make use of this structure. Thus:

> Understanding mathematical language, involves, in part, being able to identify the fundamental logical and mathematical structure of an assertion, that is, recognize logical connectives and quantifiers, function application, predication, and so on.

3.2 Understanding mathematical proof

Those who work in interactive theorem proving are attuned to the fact that representing mathematical arguments requires not only an "assertion language," but a "proof language" as well. This fact is often glossed over in conventional logic texts, where a formal mathematical proof typically amounts to little more than a sequence of assertions. But ordinary textbook proofs have a lot more structure than that. It is sometimes helpful to think of ordinary mathematical proofs as being higher-level descriptions of low-level formal axiomatic proofs, or recipes for constructing such proofs. In fact, in the field of interactive theorem proving, it is common to refer to the user's input as "code."

For example, here is a formal proof, in the Isabelle proof assistant, of the statement that if n is any natural number not equal to 1, then n is divisible by a prime.

```
lemma prime_factor_nat: "n ~= (1::nat) ==>
  EX p. prime p & p dvd n"
  apply (induct n rule: nat_less_induct)
  apply (case_tac "n = 0")
  using two_is_prime_nat apply blast
  apply (case_tac "prime n")
  apply blast
  apply (subgoal_tac "n > 1")
  apply (frule (1) not_prime_eq_prod_nat)
  apply (auto intro: dvd_mult dvd_mult2)
done
```

The first line contains a statement of the lemma to be proved. This statement becomes the current goal; subsequent lines then apply formal rules

and procedures that reduce the goal to simpler ones. For example, the first line applies a form of induction, which then requires the user to prove the statement for a given natural number, n, assuming that it holds of smaller ones. The second line splits on cases, depending on whether n is 0 or not; the first case is easy dispensed with using the previously established fact that 2 is prime. (The procedure "blast" is a generic automated routine that fills in the details.) If n is not 0, the fact that it is not 1 implies that it is greater than 1, in which case on applies the previously established fact that any number greater than 1 that is not prime can be written as a product of two strictly smaller numbers, at which point the inductive hypothesis applies.

What makes this "proof script" hard to read is that the text only gives the *instructions* that are used to act on the current goals; one has to "replay" the proof with the assistant to see the goals evolve. Fortunately, Isabelle also allows one to use a proof language called Isar [34] (modeled after Mizar's proof language [27]), which makes intermediate goals explicit. Here is a proof of the same lemma in Isar:

```
lemma prime_factor_nat:
  fixes n :: nat
  assumes "n ~= 1"
  shows "EX p. prime p & p dvd n"
proof (induct n rule: less_induct_nat)
  fix n :: nat
  assume "n ~= 1" and
    ih: "ALL m < n. m ~= 1 --> (EX p. prime p & p dvd m)"
  then show "EX p. prime p & p dvd n"
  proof -
  { assume "n = 0"
    moreover note two_is_prime_nat
    ultimately have ?thesis by auto }
  moreover
  { assume "prime n" then have ?thesis by auto }
  moreover
  { assume "n ~= 0" and "~prime n"
    with `n ~= 1` have "n > 1" by auto
    with `~prime n` and not_prime_eq_prod_nat obtain m k
     where "n = m * k" and "1 < m" and "m < n"
      by blast
    with ih obtain p where "prime p" and "p dvd m"
```

```
     by blast
    with ‘n = m * k‘ have ?thesis by auto }
  ultimately show ?thesis by blast
qed
```

Other proof languages are designed with different desiderata in mind. For example, here is a proof written in a language called Ssreflect [11], which is designed to be used with the Coq proof assistant. In the following theorem, known as the Burnside normal complement theorem, p denotes a prime number and S is assumed to be a Sylow p-subgroup of G. I have only shown the first few lines of the proof.

```
Theorem Burnside_normal_complement :
  'N_G(S) \subset 'C(S) -> 'O_p^'(G) ><| S = G.
Proof.
move=> cSN; set K := 'O_p^'(G); have [sSG pS _] := and3P sylS.
have [p'K]:  p^'.-group K /\ K <| G
  by rewrite pcore_pgroup pcore_normal.
case/andP=> sKG nKG; have{nKG} nKS := subset_trans sSG nKG.
have{pS p'K} tiKS: K :&: S = 1
  by rewrite setIC coprime_TIg ?(pnat_coprime pS).
suffices{tiKS nKS} hallK: p^'.-Hall(G) K.
  rewrite sdprodE //= -/K; apply/eqP;
  rewrite eqEcard ?mul_subG //=.
  by rewrite TI_cardMg //= (card_Hall sylS) (card_Hall hallK)
  mulnC partnC.
```

The language is not for the faint-hearted. It is, however, remarkably efficient for writing proofs, allowing one to combine a number of small steps naturally into one line of code.

While there are striking differences between these various proof languages, there are also many features in common. Ordinary mathematical proofs call upon us to perform many different types of reasoning. At any point in a proof, we may be unwrapping hypotheses or establishing small facts that set the context for the subsequent proof; we may be applying previous lemmas or theorems and checking that the side conditions are satisfied; we may be unfolding a definition, or naming an object asserted to exist; we may be carrying out a calculation; and so on. Any proof language that aims to capture ordinary mathematical argumentation has to have mechanisms that allow one to carry out these steps, and, conversely, the formal mechanisms that are designed to allow one to do this efficiently helps shed light on what is necessary to read and write ordinary mathematical proofs.

> Understanding mathematical proof involves, in part, being able to recognize contextual cues, explicit or implicit reliance on local assumptions, background knowledge, recently established facts, and so on; and to determine whether inferences are a matter of calculation, unwrapping definitions, applying a lemma, etc.

3.3 Understanding mathematical domains and structures

Let us engage in with a little exercise. Suppose we know that z be a complex number satisfying $|z| \leq 1$, and we want to bound the absolute value of e^z. We might start expanding e^z as a Taylor series as follows:

$$|e^z| = \left|\sum_{i=0}^{\infty} \frac{z^i}{i!}\right| \leq 1 + |z| + \left|\sum_{i=2}^{\infty} \frac{z^i}{i!}\right| \leq \ldots$$

In this expression, what type of object is i? z^i? 1? What does the division symbol denote? The symbol $\leq$? The summation symbol?

On inspection, we see that the variable i indexes a sum, so it ranges over the nonnegative integers. Since z is a complex number, so is z^i. We then divide z^i by the integer $i!$; this is possible because $i!$, an integer, can also be viewed as a complex number, and we can divide complex numbers. But taking the absolute value returns a real number; thus the symbol "1" here denotes the corresponding real number. Indeed, The ordering relation doesn't make sense on the complex numbers; so $\leq$ *has* to be viewed as a comparison between real numbers. As far as the summation symbol is concerned, keep in mind that in the expression $\sum_{i=0}^{\infty} \frac{z^i}{i!}$, i is a dummy variable, which is to say, writing $\sum_{j=0}^{\infty} \frac{z^j}{j!}$ does not change the value of the expression. One way to analyze the notation is to view the inner expression as denoting the *function* which maps any integer, i, to $\frac{z^i}{i!}$. Summation then becomes a higher-order operator, which takes a function as an argument.

What is interesting is that we are typically not mindful of these subtle issues when reading and working with expressions like the one above. Mathematical competence involves being able to recognize these facts implicitly and use that information in appropriate ways. When it comes to formalizing such proofs, it turns out to be remarkably difficult to spell out such details precisely. For example, one can take integers to be a different sort of object than complex numbers, in which case one has make use of the standard embedding of the integers in the complex numbers to make sense of the expression; or one can take integers to be complex numbers

satisfying the additional property of being integral, in which case, one has to rely on closure properties of operations like addition and multiplication to keep track of which objects in an expression have this property.

A good deal of technology has been borrowed from the theory of programming languages and the theory of automated reasoning to cope with this. For example, *type inference* involves determining, in a given context, what type of object a given expression denotes. *Overloading* is the act of using the same symbol for more than one purpose, such as using $\cdot$ as the multiplication symbol in more than one group, or using + for the natural numbers and the reals. *Polymorphism* and *type classes* provide means of making use of the fact that operations like addition have common properties (like $x + y = y + x$) in different instantiations. A *coercion* is a means of casting of a value of one type to another, for example, viewing an integer i as a real number in contexts where the latter is expected. *Implicit arguments* provide ways of systematically leaving out information when it can be inferred from the context, for example, writing $g \cdot h$ for multiplication in a group when the appropriate instance of group multiplication can be inferred. Coercions and implicit arguments are often insert automatically using *unification* and *matching* algorithms, which find ways of instantiating variables to get two terms to agree.

The kinds of algebraic reasoning that require such inferences are ubiquitous in mathematics. For example, when manipulating an expression $\sum_{i<n} a_i$, it may not matter whether the summation symbol is taken to mean addition in the integers, the complex numbers, or an abelian group. All the following laws hold in any commutative monoid:

$$\sum_{i<n+1} a_i = \left(\sum_{i<n} a_i\right) + a_n$$

$$\sum_{i\in S\cup T} a_i = \sum_{i\in S} a_i + \sum_{i\in T} a_i \quad \text{if } S \cap T = \emptyset$$

$$\sum_{i\in S}(a_i + b_i) = \sum_{i\in S} a_i + \sum_{i\in S} b_i$$

Also,

$$c \cdot \sum_{i\in S} a_i = \sum_{i\in S} c \cdot a_i$$

holds if $\cdot$ distributes over +. In fact, these laws still hold when the summation operator is instantiated not only by summation in the integers or complex numbers, but also as various types of products ($\prod_{i\in S} a_i$), boolean operations or meets and joins in a lattice ($\bigvee_{i\in S} a_i$, $\bigwedge_{i\in S} a_i$), the minimum and

maximum operations on the natural numbers ($\min_{i \in S} a_i$, $\max_{i \in S} a_i$), unions and intersections of sets ($\bigcup_{i \in S} a_i$, $\bigcap_{i \in S} a_i$), or the least common multiple or greatest common divisor functions on the integers ($\mathrm{lcm}_{i \in S} a_i$, $\gcd_{i \in S} a_i$).

Moreover, algebraic reasoning often requires us to view the same object in multiple ways. For example, if F is a field with a subfield E, then F can simultaneously be viewed as a field, a vector space over E, and an algebra over E. If H and K are subsets of a group G that are closed under the group operations, then H and K are also groups in their own right. An expression like $H \cap K$ can therefore be viewed as describing the intersection of the two sets, so that an element g is in $H \cap K$ if and only if g is in both H and K. But $H \cap K$ is also a group, containing the identity of G and having group operations that arise by restricting those of G. Proof assistants need to be able to handle these multiple views, just as we do when we do mathematics. Indeed, that ability is a fundamental part of mathematical competence:

> Understanding mathematical conventions regarding domains and types involves being able to resolve ambiguities and infer type information from the context; being able to recognize concrete domains as implicitly embedded in other domains; being able to recognize concrete and abstract structures as instances of more general classes of structures; and so on.

3.4 Understanding mathematical inference

So far, we have considered only some of the most basic aspects of mathematical competence, namely, the ability to parse and understand general mathematical language, and keep track of the kinds of objects at play in a mathematical proof. We have not even begun to consider even the mildest forms of mathematical reasoning proper.

Spelling out every textbook inference in terms of elementary logical steps is tedious and difficult, and most interactive proof assistants employ various methods to fill in small gaps automatically. One can get a sense of such methods from the two-volume *Handbook of Automated Reasoning* [26], or John Harrison's excellent introductory textbook *Practical Logic and Automated Reasoning* [17]. Once again, here I only have space to offer a cursory glance at the field. Some broad categorizations can be used to characterize different approaches to the problem. To start with, one can distinguish between decision procedures and search procedures. The former are algorithms that are guaranteed (at least, in principle) to terminate when called on a class of inferences, and determine whether or not the inference is valid (ideally, with some kind of formal certificate or proof of

validity when the answer is positive). Alas, thanks to Gödel, we know that many classes of inferences are undecidable; to handle such inferences, we can design procedures which search for a proof that an inference is valid, but may not halt if not. One can also distinguish between methods that are domain-general—that is, generic strategies that are designed to work in a wide-range of contexts—and methods that are domain-specific, that is, targeted toward very particular kinds of inferences. Finally, one can distinguish between "principled" search methods, which, for example, guarantee completeness and rely on fundamental theoretical considerations, and "heuristic" methods, that is, algorithms which one has tinkered with and modified to ensure that they work will in practice, often at the expense of having a clean theoretical characterization of their behavior.

When it comes to domain-general methods, one finds systems designed for propositional theorem proving; first-order theorem proving; higher-order theorem proving; and equality reasoning, among others. Each of these is a vast industry, and the references above provide a good entry to the literature. As of late, there has also been interesting research on general ways of combining different procedures in effective ways, such as including some domain specific procedures in general frameworks for proof search. "Nelson-Oppen" methods, which provide ways of combining decision procedures for domains with restricted overlap, represent one important approach.

Research in domain-specific methods is equally active. For example, linear arithmetic packages can determine the solvability of linear equalities and inequalities in the reals, integers, or combinations of these domains. Problems involving nonlinear inequalities are much more difficult, but there has been a lot of work on handling manageable fragments of the theory of real closed fields, or reasoning in the presence of transcendental functions. Interactive proof assistants have also begun to incorporate techniques from computer algebra systems; for example, methods based on Buchberger's method of *Groebner bases* can be used to solve a number of algebraic problems.

This barely scratches the surface. Automated reasoning is a vibrant field, and despite the great progress that has been made in recent years, there is still a lot we do not understand. When it comes to ordinary mathematical reasoning, one can summarize the state of affairs by saying that automated methods do especially well on large, homogeneous problems, where the search space can be kept under control and the inferences reduced to large but relatively straightforward calculations; but we are still unable to capture straightforward mathematical inferences that chain heterogeneous bits of

background knowledge together in various ways. The ability to do so is an important part of our mathematical competence:

> Understanding mathematics involves being able to carry out straightforward mathematical inferences in specific mathematical domains, even when those inferences are difficult to spell out in formal axiomatic terms.

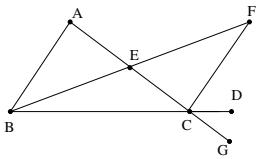
3.5 Understanding mathematical diagrams

Finally, let us briefly consider diagrammatic reasoning, which plays an important role in mathematics. Since the end of the nineteenth century, diagrams have been used only sparingly in professional mathematical texts, and conventional attitudes hold that all rigorous mathematical content should be borne by the text. Nonetheless, diagrams are often used to accompany and illustrate a mathematical argument, and some arguments can be nearly unintelligible until one has drawn a good diagram. Moreover, a good diagram can help guide the writing of a proof, and sometimes a diagram can be entirely convincing in and of itself.

Until recently, it has been common for philosophers of mathematics to dismiss diagrammatic reasoning as being merely a heuristic, psychological artifact of mathematical practice, outside the philosopher's purview. But as of late, a number of philosophers have begun to take visualization and diagrammatic reasoning more seriously [9, 22, 31]. And, it turns out, diagram use if often governed by implicit norms and conventions that can be studied and analyzed.

Consider, for example, Proposition 16 in Euclid's *Elements*.

Proposition I.16. *In any triangle, if one of the sides be produced, then the exterior angle is greater than either of the interior and opposite angles.*



Proof Let ABC be a triangle, and let one side of it BC be produced to D. I say that the exterior angle ACD is greater than either of the interior and opposite angles CBA, BAC.

Let AC be bisected at E, and let BE be joined and produced in a straight line to F. Let EF be made equal to BE, let FC be joined, and let AC be drawn through to G.

Then, since AE is equal to EC, and BE to EF, the two sides AE, EB are equal the two sides CE, EF respectively; and the angle AEB is equal to the angle FEC, for they are vertical angles. Therefore the base AB is equal to the base FC, the triangle ABE is equal to the triangle CFE, and the remaining angles equal the remaining angles respectively, namely those which the equal sides subtend; therefore the angle BAE is equal to the angle ECF.

But the angle ECD is greater than the angle ECF; therefore the angle ACD is greater than the angle BAE. Similarly also, if BC be bisected, the angle BCG, that is, the angle ACD, can be proved greater than the angle ABC as well. Therefore etc. Q.E.D. □

Later in the *Elements* (Proposition 32 of Book I), Euclid shows that in fact the external angle is equal to the sum of the internal angles, but that depends on facts about parallel lines that are established with the help of Proposition 16. Notice that the last paragraph of the proof simply asserts that angle ECD is greater than angle ECF, presumably because the latter is properly contained in the former. But what justifies this last claim? The diagram "makes it clear," but it is just such "intuitive" uses of the diagram that were called into question during the nineteenth century, with the rise of the axiomatic method.

With some effort, we can show that the desired conclusion is, indeed, warranted by diagrammatic information that is set forth in the proof. For example, points E and F are on the same side of line CD since B is on both lines and, by the construction, E is between B and F. Similarly, we can show that D and F must be on the same side of line CA, since they are both opposite from point B. But these two facts essentially say that F is "inside" the angle formed by CD and CA, which is implies that angle ECF is properly contained in angle ECD.

What is interesting about the *Elements* is that such arguments are never carried out, whereas other inferences are spelled out in great detail. Ken Manders has observed [22] that in a Euclidean proof, topological facts (the inclusion of one angle in another, the intersection of lines, the fact that one point lies between two others along a line, and so on) are often "read off from the diagram," whereas metric facts, such as the congruence of angles are segments, are always justified explicitly in the text. Inspired by his analysis, Ed Dean, John Mumma, and I [4] designed a formal system that exhibits these features, and hence is capable of representing Euclid's arguments more faithfully. Our project involved, in particular, undertaking a careful study of the diagrammatic inferences that occur in the first four

books of the *Elements*, and characterizing the norms and conventions that determine the kinds of information that one is able to read off from the diagram in a Euclidean proof. Understanding Euclidean geometry means, in part, being able to distinguish the valid diagrammatic inferences from invalid ones. More generally:

> Understanding mathematical diagram use involves being able to represent information in a diagram appropriately, and draw valid inferences from the information so represented.

4 The philosophy of mathematics

At this stage, it would be reasonable for you to ask, "what does all this have to do with the philosophy of mathematics?"

To answer this question in a constructive way, it will be helpful to set some ground rules. The question is not meant to spark a turf war, with mathematicians, computer scientists, and philosophers squabbling over who is allowed to make pronouncements over mathematical understanding. Nor is asking whether issues in formal verification have any role in philosophy a matter of passing value judgment on the former; mathematical logic and computer science are important fields of inquiry in their own right, independent of their interaction with philosophy. Rather, let us take the question above to ask what role distinctly philosophical methods can play in relation to the methods of mathematical logic and computer science, and the extent to which philosophical inquiry can inform, and be informed by, work in mathematical logic and software engineering. Towards the end of *The Problems of Philosophy* [28], Bertrand Russell highlighted the role of philosophy in sharpening concepts and clarifying vague intuitions in order to make further scientific inquiry possible. Here I will argue that the philosophy of mathematics can play just such a role here, in helping us come to terms with what exactly we are talking about when we try to talk about mathematical understanding in various scientific contexts. In other words, I am claiming that a better philosophical framework for reasoning about mathematical understanding can support such scientific work, as well as address the kinds of philosophical "problems" that I described in Section 2. The next three sections suggest three ways that such a philosophical framework would be useful.

4.1 Mathematical methods and abilities

Take another look at the pronouncements on understanding that I used to summarize the conclusions of each subsection of Section 3. What do they have in common?

You will notice that the phrase "being able to" occurs in each; in other words, in each case I have characterized an aspect of mathematical understanding in terms of "being able to" perform certain tasks. Informally, we often explain our ascriptions of understanding by describing the associated abilities. For example, if I tell you that my calculus students don't understand integration by parts and you ask me what I mean, I am likely to respond by giving examples of what they can and cannot do.

This provides a helpful way of thinking about mathematics. On traditional foundational accounts, mathematical knowledge is viewed as a collection of propositions. In the context of a formal background theory, we formulate definitions and prove theorems; our knowledge then amounts to knowing that our terms have been defined in thus-and-such a way, and knowing that thus-and-such a theorem is a consequence. But once we have fixed our definitions and axiomatic framework, all the consequences are determined, and the growth of mathematical knowledge is then simply a matter of cranking out these consequences. If we think of mathematical understanding, more broadly, in terms of a body of methods and abilities, new modes of analysis are opened up to us. Rather than a collection of facts, mathematics becomes something much richer and more interesting, namely, a way of thinking and confronting the mathematical challenges we face. It is not just a matter of knowing *that* certain statements are true, but, rather, a matter of knowing *how* to proceed appropriately in mathematical contexts.

Providing a theory of mathematical understanding then amounts to giving an account of the relevant methods and abilities. Such an account can be used to address the philosophical problems raised in Section 2, providing us with better means to explain what we obtain from different proofs of a theorem and why certain historical developments are so important.

There is a straightforward model that can be invoked. Doing mathematics means undertaking various tasks, such as solving problems, proving theorems, verifying inferences, developing theories, forming conjectures, and so on. "Reasoning" involves a passage though various epistemic states en route to our goals. "Understanding" then consists of the methods, techniques, procedures, protocols, tactics, and strategies that make this passage possible. As Section 3 suggests, this involves all of the following:

- being able to recognize the nature of the objects and questions before us
- being able to marshal the relevant background knowledge and information
- being able to traverse the space of possibilities before us in a fruitful way
- being able to identify features of the context that help us cut down complexity

Emphasizing the word "method" means focusing on the procedures that carry us from one state to another; emphasizing the word "ability" means focusing on the net result of the transformation.

But we face a number of problems when we try to fill out the details and develop ways of talking about "methods" and "abilities" in more scientific terms. The notion of a "method" has the connotations of an algorithm, which is to say, a specific way of going about something. But often we only care about what it is that the method accomplishes, and not the particular details of how it is accomplished. That is, different methods can give rise to the same ability; you and I may multiply three-digit numbers in different ways, and, in some contexts, it might only matter that we can both carry out the multiplication. On the other hand, there is a compositional aspect to our mathematical abilities, in that some abilities can be explained in terms of others. For example, my ability to solve a problem may depend on my ability to apply a certain lemma, which may in turn depend on my ability to expand a definition appropriately. Or my ability to carry out a calculation may depend on the ability to recognize that certain background conditions obtain. These features then push us to think of methods in terms of algorithms and subroutines, which, again, may push us to overly specific descriptions of what they are doing.

There are other features of mathematical abilities and methods that pose challenges. For example, the identity criteria are murky; when should we take two descriptions of a method or an ability to denote the same object? Moreover, methods are inherently fallible. For example: one can show that a subgroup H of G is normal in G by showing that it is a characteristic subgroup of another normal subgroup of G; but this is not the only way to show that H is normal in G, and not every normal subgroup can be identified in this way. Thus we need a way of describing methods that are appropriate, though at the same time imperfect, in a given context.

In sum, the challenge is to develop a language for talking about mathematical methods and abilities that is well-suited to studying the issues

raised in Sections 2 and 3. The computational models employed in the field of formal verification provide a good starting point, but, ultimately, we need to focus on the features of mathematics that render it intelligible, rather than proof assistants and their implementation. The goal, then, is to find a level abstraction that is appropriate for talking about mathematical understanding.

4.2 Mathematical concepts

Developing a better language for talking about mathematical methods and abilities may have some side benefits as well. Consider, for example, the notion of a mathematical *concept*. Conventional psychological approaches to the notion of concept, involving prototypes and exemplars, don't do a good job of characterizing mathematical concepts and the role the play in mature mathematical reasoning. For example, some objects may fall more distinctly under the concept of "table" than others, and among various tables, some are more prototypical than others. In contrast, mathematical concepts can have sharp boundaries. There is a precise modern definition of what it means to be a "group," and any particular mathematical object either is or is not an instance of the group concept. To be sure, there are more natural or common instances of groups; but that naturality does not make them any more group-ish than contrived examples.

Yet mathematical concepts have a number of properties that make it hard to pin them down. For example, mathematical concepts, like the group concept, can evolve over time. Moreover, understanding a concept admits degrees: an undergraduate understanding of the group concept is different from that of a graduate student working in group theory, which, in turn, differs from that of the leading experts in the field. Various things "improve our understanding" of a concept, and not just seeing more of them. For example, representation theory, the method of representing elements of groups as linear transformations of vector spaces, gives us a better understanding of groups. When we consider the historical record, we often recognize "implicit uses of a concept" in the forerunners of our modern theories. For example, Euler's work on power residues (and, particularly, residues modulo a prime) provides a good example of an implicit use of the group concept years before the concept had been defined or axiomatized.

Can we come up with precise ways of thinking and talking about mathematical concepts that accord with these informal observations? One solution may be to think of mathematical concepts as collections of abilities

bundled around a central token (see footnote 18 of [3]). Surely one important ability that can be associated with any mathematical concept is the ability to state the corresponding definition and apply it correctly. As a result, we can still follow the traditional Fregean route by saying that an object "falls under a concept" if it satisfies the associated definition. But now we can analyze the notion of "understanding a concept" more generally as possessing the associated abilities. For example, understanding the group concept involves knowing the definition of a group; knowing common examples of groups, and being able to recognize implicit group structures when it is fruitful to do so; knowing how to construct groups from other groups or other structures, in fruitful ways; recognizing that there are different kinds of groups (abelian, nilponent, solvable, finite vs. infinite, continuous vs. discrete) and being able and prone to make these distinctions; knowing various theorems about groups, and when and how to apply them; and so on.

You can check that this way of thinking about mathematical concepts jibes well with the observations in the previous paragraph. For example, concepts evolve as the ways and contexts we use them expand, and using a concept "implicitly" can mean employing an instance of a method associated to the concept without identifying it as such. To be sure, this makes mathematical concepts somewhat vague and open-ended. But the point is, our talk of mathematical concepts *does* treat them as vague and open-ended; and this analysis makes them vague and open-ended in just the right away.

It may be helpful to compare this to the more traditional view of mathematical concepts, which treats them as static and unchanging entities. On the Fregean view, the analysis of a concept amounts to fixing the proper definition, which, in a sense, determines everything there is to say about the concept. The stark difference between the two views is tempered by the fact that Frege was more specifically focused on the problem of justification of mathematical knowledge. When it comes to accounting for the normative rules for mathematical justification, Frege's account fares rather well. There are at least three senses, however, in which a Fregean analysis comes up short.

The first is foundational. In the early twentieth century, mathematical logic was able to reduce much of mathematical reasoning to a few basic concepts and axioms; Zermelo-Fraenkel set theory, for example, provides a foundation for mathematics based on a small list of assumptions about the universe of sets. But one is still left with the question as to what justi-

fies *those*. Most philosophers of mathematics take logicism to have failed, in the sense that doing mathematics requires commitments to entities and principles that cannot be accounted for by logic alone. But attempts to find compelling extralogical principles that can provide us with perfect knowledge of abstract mathematical objects have largely stalled. In other words, we have swept all our dust into one neat little pile and invested a good deal of effort in moving the pile of dust around, but it never really goes away. It seems likely that if one is aiming to justify one's axioms and basic concepts on broader grounds, one will ultimately have to attend to the roles they play in organizing our mathematical knowledge, and the role that knowledge plays in organizing and structuring our scientific experiences. (Maddy [20] urges something like this approach.) A more robust notion of concept can help in that regard, by giving us a richer vocabulary to explain why certain ways of organizing and structuring our knowledge are better than others.

A second sense in which the Fregean analysis falls short is that it fails to account for the kind of mathematical reasoning that often takes place in the absence of a clear foundational framework. For example, in the eighteenth century Euler employed a number of novel and striking arguments involving infinite sequences and series which were not made rigorous, according to the modern understanding of that term, until much later; sometimes not even until the twentieth century (see [33]). Mark Wilson's engaging and thorough exploration of concepts [36] offers a number of similar examples in applied mathematics and physics, as does Urquhart [32]. The work on Euclidean geometry described in Section 3.5 shows that the geometry in Euclid's *Elements* is also governed by precise norms, once again in the absence of a Fregean foundation. Once again, a more robust notion of concept may be able to help explain the way informal concepts guide our reasoning, even in the absence of precise definitions.

Finally, Frege's analysis was simply not designed to account for the kinds of normative evaluations discussed in Section 2. As Poincaré observed, telling us how we are *allowed* to use our mathematical concepts is a far cry from telling us how we *ought* to use them. Thus we can view the Fregean analysis as more specifically trying to provide an idealized account of the normative rules of justification in situations where our concepts can be clearly defined. When it comes to justification in the broader sense of using mathematical notions fruitfully and appropriately, once again, we should be prepared to look beyond the Fregean framework.

4.3 Mathematical ease and difficulty

There are other ways that the issues raised in Sections 2 and 3 push us to look beyond traditional foundational analysis. The problem is simply that foundational reduction washes out many important nuances. For example, from a set-theoretic standpoint, there is only one type of mathematical object (set); there is only one fundamental relationship between mathematical objects (the binary element-of relation); and one only needs one "method" to verify inferences, that is, systematic search for a proof from the axioms of set theory. From this standpoint, it is hard to recognize differences between algebraic and geometric methods; different styles of proof; or the value of a good definition.

What makes foundational reduction an oversimplification in such contexts are issues of complexity. Knowing that, in principle, definitions and theorems can be unpacked until we get to set-theoretic primitives does not help us reason about them pragmatically. Differences in the way we organize our mathematical knowledge and express our mathematical ideas matter precisely because we have limited time, energy, memory, and reasoning capacities. Part of understanding why we value certain mathematical developments involves understanding how the right concepts and methods simplify the mathematical tasks before us.

But how shall we measure complexity? As philosophers, we won't be the first to grapple with the issue; "complexity" is a term of art in a number of applied disciplines. In computer science, one measures the complexity of problems in terms of asymptotic bounds on the time and space needed to compute solutions in a fixed machine model. In logic, one can measure the complexity of definitions, say, by the number (and type) of quantifiers they employ; and one can measure the complexity of proofs by the number of symbols they contain. Psychologists measure the complexity of basic cognitive tasks in terms of the amount of time it takes to carry them out, or the stage of our development at which we are capable of doing so.

For example, the field of proof complexity [25] provides a number of interesting "speedup results," which show how expanding the language and conceptual resources of a mathematical theory, even conservatively, can serve to shorten the lengths of proofs dramatically. With some cleverness, one can find explicit combinatorial theorems that exemplify this behavior (see [7, 23]). This may seem to offer good explanations as to how a careful choice of language and concepts serves to reduce complexity. But such results do not tell the whole story. First, "length of proof" is a count of the number of symbols in proofs in a formal axiomatic system. Such a

measure depends very much on how the system is formulated; although such measures tend to be stable, up to a polynomial, across reasonable systems, that says nothing about the length of the proof of a *single* theorem. Second, ordinary textbook proofs are much higher-level objects than formal axiomatic derivations, and, as I argue elsewhere [2], the kinds of normative assessments that we are interested in here do not accord well with the low-level modeling. Third, the combinatorial examples are somewhat contrived, cooked up by logicians to serve as counterexamples, and the speedup vanishes when one replaces the systems in question with ones only slightly stronger. Indeed, ordinary mathematical theorems that skirt unprovability using ordinary mathematical methods are surprisingly hard to come by; most mathematics can be carried out in fairly weak theories, where the obvious reflection principles that can shorten a proof are uncontroversial (see [1]). Moreover, length of proof is a measure of the formal object, not the ease or difficulty we encounter in trying to find, remember, or reconstruct it; it is a measure of "syntactic complexity" rather than "difficulty." Finally, speedup results are overly dramatic. A definition that makes it possible to carry out our reasoning more cleanly and efficiently, and thereby reduce the length of a journal article by a third, is clearly a good definition. What we really care about are the subtle ways that good definitions streamline our ordinary mathematical experiences, not the dramatic and clever ways we can abuse a formal axiomatic system.

One can raise similar objections to other complexity measures on offer. We might try to classify the difficulty of common mathematical tasks in terms of their computational complexity, but this is an asymptotic model; what we often care about are the complexity of individual tasks, or, for example, a class of tasks where the parameter in question can reasonably be taken to be bounded by a constant (say a trillion). This objection isn't just pedantic; at small sizes, the particular choice of machine model can make a huge difference. Turing machines are not good models for the kinds of things that *we* find it easy to do, nor are they good models for the kinds of tasks that take place against richly structured background knowledge. Finally, computational complexity is best at modeling deterministic algorithms; in mathematics, we often employ heuristic methods that tend to do well in the kinds of situations that arise in ordinary mathematical practice. What is missing is an informative theoretical characterization of what it means to "do well in the kinds of situations that arise in ordinary mathematical practice." Computational complexity was simply not designed for this purpose.

Psychological measures of difficulty go too far to the other extreme. For

one thing, we only have clean experimental results for very basic cognitive tasks. Moreover, this shifts our focus to "incidental" features of our cognitive abilities, rather than the "essential" features of the mathematics we are trying to model. What we want is an account of how mathematics helps us take advantage of the distinctly mathematical features of a problem at hand and tame a combinatorial explosion of possibilities, one that is not overly sensitive to the number of digits we are capable of holding in our short-term memory.

But it is important to keep in mind that saying that the measures of complexity we have considered are not quite right does not mean that they are entirely wrong. Certainly the lengths of proofs and calculations and our cognitive limitations have a lot to do with what makes a piece of mathematics hard or easy. Conventional complexity measures therefore provide a good starting point. What we need now are ways of talking about complexity that are suitable for analyzing the features of the *mathematics* that extend the capacity and reach of our thought.

Once again, getting a better grip on mathematical ease and difficulty may have broader philosophical implications. I began this essay by describing recent efforts to come to terms with the various values that one comes across in ordinary mathematical discourse. Such explorations have given rise to methodological concerns. It is all well and good to make lists of theoretical virtues; but what, exactly, endows them with normative force? Mathematics is ultimately a matter of getting at the truth; isn't everything else incidental? Aren't all the other judgments merely subjective and pragmatic, falling outside the proper scope of philosophy? Tappenden [31] raises such concerns as follows:

> [J]udgements of "naturalness" and the like are *reasoned*. It is not just some brute aesthetic response or sudden, irrational "aha!" reaction that brings about the judgement that — for example — "the scheme is the more natural setting for many geometric arguments"... Quite the contrary: elaborate reasons can be and are given for and against these choices. One job facing the methodologist of mathematics is to better understand the variety of reasons that can be given, and how such reasons inform mathematical practice.
>
> The factual observation should be beyond controversy: reasoned judgements about the "proper" way to represent and prove a theorem inform mathematical practice. I have found that more contention is generated by the disciplinary classi-

> fication of the study of these judgements and the principles informing them: is this *philosophy,* or something else, like cognitive psychology?

At issue is not whether we can clarify and explain our mathematical assessments; instead, the question is whether we can find any "objective" sense in which they should be taken to be normative, rather than reflections of personal preference, historical accident, or incidental features of our biological makeup. In other words, what is lacking is a sense in which our mathematical values are mathematically valuable.

Such concerns are not limited to mathematics; they apply just as well to questions concerning the objectivity of ethical and aesthetic judgments. But when it comes to mathematics, a suitable theory of ease and difficulty may provide an informative way of addressing these concerns. Insofar as we can develop appropriate idealizations of our cognitive capacities and limitations, there is a sense in which we can determine some of our value judgments to be objective; that is, we can show how our various machinations and stratagems serve to extend the capacities for the use and discovery of mathematical knowledge in any beings with cognitive constraints roughly like ours. This may not put all our concerns about normativity and objectivity to rest, but it provides a sense in which clear philosophical progress can be made.

5 Conclusions

Encouraged by these musings, you may find yourself tempted to get right to work and start defining basic terms like "understanding," "ability," and "concept." Resist that temptation! Before we begin to construct an overarching theory, we have to start coming to terms with some of the basic data. It therefore makes sense to begin with more focused questions, ones for which satisfying answers are within reach. To that end, it is also helpful to look to domains of application, among those fields that explicitly or implicitly depend on notions related to mathematical understanding: fields such as formal verification and automated reasoning; mathematical pedagogy and cognitive science; history (and historiography) of mathematics; and mathematics itself. In other words, we should begin by trying to clarify specific *aspects* of mathematical understanding, and the roles that our conceptions of mathematical understanding play in particular scientific practices. Over time, the data we accumulate from such smaller and

more focused studies should come together to provide us with a coherent, comprehensive picture.

But what if they don't? It is conceivable that our disparate attempts to get at mathematical understanding will take us down divergent paths. We may decide, in the end, that notions of understanding that arise in automated reasoning have nothing to do with notions of understanding that arise in cognitive science, which, in turn, tell us nothing about the methods and goals of working mathematicians. What then?

Well, in that case, our work will have merely contributed to the conceptual foundations of automated reasoning, cognitive science, pedagogy, history of science, and so on; and taught us some interesting things about mathematics as well. Surely we could do a lot worse.

References

[1] Jeremy Avigad. Number theory and elementary arithmetic. *Philosophia Mathematica*, 11:257–284, 2003.

[2] Jeremy Avigad. Mathematical method and proof. *Synthese*, 153:105–159, 2006.

[3] Jeremy Avigad. Understanding proofs. In [21], pages 317–353.

[4] Jeremy Avigad, Edward Dean, and John Mumma. A formal system for Euclid's *Elements*. *Review of Symbolic Logic*, 2:700–768, 2009.

[5] Jeremy Avigad, Kevin Donnelly, David Gray, and Paul Raff. A formally verified proof of the prime number theorem. *ACM Transactions on Computational Logic*, 9:2, 2007.

[6] Solomon Feferman. Kreisel's "unwinding" program. In Piergiorgio Odifreddi, editor, *Kreiseliana: About and Around Georg Kreisel*, pages 247–273. A.K. Peters Ltd., Wellesley, MA, 1996.

[7] Harvey Friedman. *Boolean Relation Theory and Incompleteness*. ASL Lecture Notes in Logic, to appear.

[8] H. Furstenberg. Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *Journal d'Analyse Mathématique*, 31:204–256, 1977.

[9] Marcus Giaquinto. *Visual Thinking in Mathematics: An Epistemological Study*. Oxford University Press, Oxford, 2007.

[10] Georges Gonthier. Formal proof—the four-color theorem. *Notices of the American Mathematical Society*, 55:1382–1393, 2008.

[11] Georges Gonthier and Assia Mahboubi. A small scale reflection extension for the coq system. Technical Report INRIA-00258384, Microsoft Research and INRIA, 2008.

[12] Georges Gonthier, Assia Mahboubi, Laurence Rideau, Enrico Tassi, and Laurent Théry. A modular formalisation of finite group theory. In Klaus Schneider and Jens Brandt, editors, *Theorem Proving in Higher Order Logics 2007*, pages 86–101. Springer, Berlin, 2007.

[13] Thomas C. Hales. The Jordan curve theorem, formally and informally. *American Mathematical Monthly*, 114:882–894, 2007.

[14] Thomas C. Hales. Formal proof. *Notices of the American Mathematical Society*, 55:1370–1380, 2008.

[15] John Harrison. A formalized proof of Dirichlet's theorem on primes in arithmetic progression. *Journal of Formalized Reasoning*, 2:63–83, 2009.

[16] John Harrison. Formalizing an analytic proof of the prime number theorem. *Journal of Automated Reasoning*, 43:243–261, 2009.

[17] John Harrison. *Handbook of Practical Logic and Automated Reasoning*. Cambridge University Press, Cambridge, 2009.

[18] Steven Kieffer, Jeremy Avigad, and Harvey Friedman. A language for mathematical language management. *Studies in Logic, Grammar and Rhetoric*, 18:51–66, 2009.

[19] Ulrich Kohlenbach. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer, Berlin, 2008.

[20] Penelope Maddy. *Second Philosophy: A Naturalistic Method*. Clarendon Press, Oxford, 2007.

[21] Paolo Mancosu, editor. *The Philosophy of Mathematical Practice*. Oxford University Press, Oxford, 2008.

[22] Kenneth Manders. The Euclidean diagram. In [21], pages 80–133.

[23] Jeff Paris and Leo Harrington. A mathematical incompleteness in Peano arithmetic. In Jon Barwise, editor, *Handbook of Mathematical Logic*, pages 1133–1142. North-Holland, Amsterdam, 1977.

[24] Henri Poincaré. *Science et Méthode*. Flammarion, Paris, 1908. Translated by Francis Maitland as *Science and Method*, Dover Publications, New York, 1952.

[25] Pavel Pudlák. The lengths of proofs. In Samuel Buss, editor, *Handbook of Proof Theory*, pages 547–637. North-Holland, Amsterdam, 1998.

[26] John Alan Robinson and Andrei Voronkov, editors. *Handbook of Automated Reasoning (in 2 volumes)*. Elsevier and MIT Press, 2001.

[27] Piotr Rudnicki. An overview of the Mizar project. In *1992 Workshop on Types for Proofs and Programs*. Chalmers University of Technology, Bastad, 1992.

[28] Bertrand Russell. *The Problems of Philosophy*. Home University Library, 1912. Reprinted by Oxford University Press, Oxford, 1959.

[29] Endre Szemerédi. On sets of integers containing no k elements in arithmetic progression. *Acta Arithmetica*, 27:199–245, 1975.

[30] Terence Tao. A quantitative ergodic theory proof of Szemerédi's theorem. *Electronic Journal of Combinatorics*, 13(1):Research Paper 99, 2006.

[31] Jamie Tappenden. Proof style and understanding in mathematics I: visualization, unification, and axiom choice. In Paolo Mancosu, Klaus Frovin Jorgensen, and Stig Andur Pedersen, editors, *Visualization, Explanation and Reasoning Styles in Mathematics*, pages 147–214. Springer, Berlin, 2005.

[32] Alasdair Urquhart. Mathematics and physics: strategies of assimilation. In Mancosu [21], pages 417–440.

[33] V. S. Varadarajan. *Euler Through Time: A New Look at Old Themes*. American Mathematical Society, Providence, RI, 2006.

[34] Makarius Wenzel. Isabelle/Isar — a generic framework for human-readable proof documents. *Studies in Logic, Grammar, and Rhetoric*, 10(23):277-298, 2007.

[35] Freek Wiedijk. *The Seventeen Provers of the World*. Springer, Berlin, 2006.

[36] Mark Wilson. *Wandering Significance: An Essay on Conceptual Behavior*. Oxford University Press, Oxford, 2006.

[37] Ludwig Wittgenstein. *Remarks on the Foundations of Mathematics*. Blackwell, Oxford, 1956. Edited by G. H. von Wright, R. Rhees and G. E. M. Anscombe, Translated from the German by Anscombe. Revised edition, MIT Press, Cambridge, Mass., 1978.

Computability Theory

S. BARRY COOPER*

Nature was computing long before humans started. It is the algorithmic content of the universe makes it an environment we can survive in. On the other hand, computation has been basic to civilisation from the earliest times. But *computability*? Computability theory is computation with consciousness, and entails the huge step from *doing* computation to observing and analysing the activity, and understanding something about what we can and cannot compute. And then — using the knowledge acquired as a stepping stone to a better understanding of the world we live in, and to new and previously unexpected computational strategies.

It is relatively recently that computability graduated from being an essential element of our daily lives to being a concept one could talk about with precision. Computability as a *theory* originated with the work of Gödel, Turing, Church and others in the 1930s. The idea that reasoning might be essentially algorithmic goes back to Gottfried Leibniz — as he says in *The Art of Discovery* (1685), [24, p.51]:

> The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, to see who is right.

It was Gödel's reduction of the first-order *theory* of Peano arithmetic to *recursive* arithmetic which gave a first formal expression to this idea. Since then, we have seen what it means to be computable captured in diverse models, invariably equivalent to the standard Turing machine model from 1936 of Alan Turing.

Ironically, just when the mathematical models gave new clarity to basic questions about the nature of real-world computability, and the speculations of Leibniz and others, the mathematicians and philosophers went their different ways. And, growing out of the theoretical construct of the Universal Turing Machine, the 1940s and 50s saw the development of the

*University of Leeds, Leeds LS2 9JT, U.K.

stored-program computer and a third and dominant stream of research — *theoretical computer science*. While philosophers continued to think about the nature of computability, mathematical logicians became preoccupied with the esoteric technical questions of what came to be called *recursion theory*. And theoretical computer scientists mined the seemingly inexhaustible riches of the Turing model of computation — which would have been no surprise to Turing himself, who is frequently quoted: "Machines take me by surprise with great frequency".

Nowadays, computability as a field is still broadly defined by the work and research interests of Alan Turing. The reputation of Turing himself continues to grow, and he has been variously adopted by mathematicians, computer scientists, biologists, cryptologists and philosophers. Most of Turing's various research interests are still key ones today. And there is a new coming together of people working in different areas and approaching basic questions from different directions.

In mathematics there is a new awareness of how technical work related to incomputability has real-world consequences. In computer science there is a growing dissatisfaction with the constraints of the standard Turing model of computation, the development of computational paradigms derived from nature — biological, quantum and connectionist models — and a readiness to apply logic and mathematical structures. The continued importance of the Turing Test for those trying to build intelligent machines reminds us of how bereft we are of adequate models of human intelligence (see [3]). While philosophers variously call on sources from different areas of science and the humanities, examining the computational significance of such natural phenomena as quantum randomness and emergence, in an effort to pin down the extent to which nature computes, and what computability is in the real world.

1 The Origins of Computability Theory

So, for the beginnings of computability theory, we need to go back to the year 1936. The scientific developments of the decade 1927-36, such as the development of quantum mechanics, Gödel's incompleteness theorems, the discovery of the Universal Turing machine, had an effect that would not have been obvious to those reading the newspaper headlines of that year, concerned with such things as the civil war in Spain, economic recession, and the Berlin Olympics. The end of that decade saw the publication of a thirty-six page paper [21] by a young mathematician, Alan Turing, claim-

ing to solve a longstanding problem of the distinguished German mathematician David Hilbert — see Hodges [11] for biographical background. A byproduct of that solution was the first machine-based model of what it means for a number-theoretic function to be computable, and the description of what we now call a *Universal Turing Machine*. At a practical level, as Martin Davis describes in his 2000 book [5] *The Universal Computer: The Road from Leibniz to Turing*, the logic underlying such work became closely connected with the later development of real-life computers. The stored program computer on ones desk is a descendant of that first universal machine. What is less often remembered is Turing's theoretical contribution to the understanding of the *limitations* on what computers can do. There are quite easily described arithmetical functions which are not computable by *any* computer, however powerful. And — as shown by David Deutsch [6] in 1985 — the advent of quantum computers will not change this.

Before computers, computer programs used to be called *algorithms*. Algorithms were just a finite set of rules, expressed in everyday language, for performing some general task. What is special about an algorithm is that its rules can be applied in a potentially unlimited number instances of a particular situation. We talk about the *algorithmic content* of Nature when we recognise patterns in natural phenomena which appear to follow general rules. One of the main tasks of science, at least since the time of Isaac Newton, is to make mathematically explicit the algorithmic content of the world about us. A more recent task is to come to terms with, and analyse, the theoretical obstacles to the scientific approach. This is where the discovery of *incomputability*, and the theory which flows from that discovery, play such an important role.

But it was David Hilbert's famous address to the 1900 International Congress of Mathematicians in Paris, which set out a mathematical agenda which turned out to be of fundamental importance to the history of computability. Many of the twenty-three problems he posed have been solved. But a main theme running through them still preoccupies us, although in ways very different to what Hilbert would have expected. Essentially, Hilbert hoped to reduce a wide spectrum of familiar problems to computation. For instance, Hilbert's tenth problem, asked for an algorithm for locating solutions to Diophantine equations. Another, leading to Turing's seminal 1936 paper, was the question (Hilbert's 'Entscheidungsproblem') of whether there is an algorithm for deciding of a given sentence whether it is logically valid or not. More generally, he raised the question of whether there exist *unsolvable* problems in mathematics. Or whether there exist

computational tasks for which there is no valid program. Hilbert believed there were no unsolvable problems, famously declaring in Königsberg in September 1930:

> For the mathematician there is no Ignorabimus, and, in my opinion, not at all for natural science either. ... The true reason why [no one] has succeeded in finding an unsolvable problem is, in my opinion, that there *is no* unsolvable problem.

The capturing of the *notion* of computability via abstract mathematical models led to a very different view.

2 The Standard Model of Computability

Coincidentally, just the day before Hilbert's declaration quoted above, what became known as *Gödel's Incompleteness Theorem* was being quietly announced by the young Kurt Gödel at another meeting in the same city. An important technical feature of the proof of the theorem was the first formalisation of the notion of a computable function, enabling us to *talk about* computability from the outside. It was not long before there were a number of formalisations, or models, of computability, such as the recursive functions; the λ-computable functions; the Turing computable functions; Markov Algorithms; and unlimited register machines (URMs).

All of these frameworks enable one to effectively list *all possible* algorithms of that kind, and to use the list to devise a problem which cannot be solved by such an algorithm. This is essentially what Turing did in constructing a *universal* Turing machine, and hence finding a problem unsolvable by such a machine. By arguing convincingly that *any* algorithm could be performed by a suitable Turing machine, he was able to conclude that there existed problems that were *unsolvable* by any algorithm. Church, also in 1936, did something similar using λ-computability instead. For further details see, for example, [1] or [20].

An important fact is that however different these notions appear to be, they all lead to the same class of functions. Even more remarkable is that computability appears to exist independently of any language used to describe it. Any sufficiently general model gives the *same* class of functions — this assertion is captured in the *Church-Turing Thesis*, and has stood the test of time.

Of course, all these notions of computability deal with discrete data. This does reflect everyday practice, in that scientific measurements can only be

made to a given level of exactness, and this is reflected in the sort of data can computers work with. However, the *mathematics* of the real world does sometimes require us to say something about computations over continuous data, that is data described by real numbers, and this requires extended models, as we shall see below. Turing himself [21] introduced the concept of a *computable real number*.

Notice that it does not change our thesis to allow in our computations *non-deterministic* steps, wherein we are allowed free choice between a list of computational actions. For instance, any function computable by a non-determistic Turing machine can also be computed by a deterministic one. But the picture changes radically, as we see in the next section, if we compute *relative* to incomplete information. And the jury is still out on what happens when we work with time or space bounds on our computations. For instance, in the context of polynomial time bounds, the fundamental open question $P =?NP$ asks whether there are functions computable in polynomial time using non-deterministic Turing machines which are not so computable with deterministic ones.

In recent years, the reducibility of computation in real environments to the standard Turing model has been brought increasingly into question. At one time it would have been heretical to suggest that there is any computational model 'breaking the Turing barrier' of relevance outside mathematics. And the popular 'reverse mathematics' project has as one of its aims the reinstatement of a partial version of Hilbert's programme, by showing that all everyday mathematics — that is, mathematics done by real mathematicians, not logicians! — is derivable in quite weak axiomatic theories. But the attack on the standard model comes from many directions, and there is the distinct impression of an uncompleted Kuhnian paradigm change in progress.

3 The Role of Incomputability

From a modern perspective, incomputability is an *emergent* phenomenon. As such, it parallels phenomena in nature which challenge our computational capabilities, and points to something more than complexity at work in the real world (see [2]).

Turing machines can become data for other Turing machines to compute with. This observation enabled the design of a *Universal Turing Machine* which could be used to simulate computations of *any* other machine. The key technical idea here is to code information, such as machines or

their computations, as machine-friendly data, such as numbers or binary sequences. This is the theoretical basis for modern computers, which treat programs as data, to be stored and used as needed.

Having theoretically captured computers, it is a short step to finding quite natural problems which are beyond the grasp of any computer, however large or efficient. For instance, there are all sorts of general questions about *how* a given Turing machine computes which cannot be answered by any computer program. The best known of these is the so-called *Halting Problem*, which is that of determining of an arbitrary Turing machine whether it will successfully compute – that is, halt and give an output – for arbitrary input data.

So incomputability emerges at the edge of computability. Its origins are mathematically uncomplicated enough to produce this complex and intimate relationship. And just as the non-computable universe is woven from the algorithmic fabric of everyday life, so the structures on which we base its analysis are derived from appropriate computable relationships on information content — itself abstracted from the way science describes the material universe. The standard model of computationally complex environments first appeared in a difficult and still slightly mysterious paper [22] of Turing's from 1939. The *Turing universe* is a structuring of the binary real numbers based on the notion of an *oracle Turing machine*, which models what we can mathematically map of algorithmic interactions between data — coded as reals — which may or may not be algorithmic in origin.

Of course, binary reals are mathematically interchangeable with sets of natural numbers. If you are familiar with the standard definition of a Turing machine, where a Turing program consisting of a finite set of instructions for performing very basic computational actions, you can obtain the notion of an *oracle Turing machine* $\widehat{T}$ by allowing some instructions to be ones which require the machine to ask questions about membership in a set A, where A is an *oracle* which may be far from computable. An oracle machine $\widehat{T}$ computes in the usual way from input n, except that when it applies an oracle-asking instruction it reads off a number (say p) from its memory in some unambiguous way, and asks the **oracle** A "Is $p \in A$?" — where the program will dictate a different course for the subsequent computation according as the answer is "*yes*" or "*no*".

Then B is said to be *A-Turing computable* — or *Turing reducible to A* — if the characteristic function χ_B of B is A-Turing computable. We often write $B \leq_T A$ for "B is Turing computable from A".

Oracle Turing machines also allow us to compare the solvability of different mathematical problems, which is why Emil Post called the mathe-

matical structure *degrees of unsolvability* — today more usually called the *Turing degrees*, denoted by $\boldsymbol{\mathcal{D}}$. The Turing degrees consist of equivalence classes of binary reals under the equivalence relation of being Turing computable from each other. See [1] for a more detailed description of the Turing degrees and their properties.

This model can be refined in various ways to take account of constraints of time or space imposed on computations. Just as Turing machines provide a useful basis on which to base the study of *complexity* of computations, so oracle Turing machines provide degree structures relevant to complexity-theoretic questions.

Actually, there is a more general model based on *non-deterministic* Turing machines, which is applicable to computations relative to *partial* information, or equivalently, relative to data which is *enumerated* in real-time rather than being available as required.

What makes computation relative to partial information *different* is that we cannot computably tell whether our oracle is going to answer or not. In the new model, there is not just one computation dependent on guaranteed oracle responses to our queries. It is more like real life where our search for useful knowledge is only partially rewarded and new information emanates from our environment in a fairly surprising and unpredictable way.

This gives rise to one slightly different way of looking at things, based on looking what we *do* — how we *guess* and pursue different alternative computations. This leads to *nondeterministic Turing machines*, with corresponding reducibility $A \leq_{NT} B$.

Another viewpoint emphasises how knowledge is *delivered* to us by our source of outside information — in a sense *enumerated*, and according to a timetable not under our own control. This leads to a notion of *enumeration reducibility*. Formally, we define $B \leq_e A$ to mean we can computably enumerate the members of B from an enumeration of the members of A — where this enumeration of B does not depend on the *order* in which A is enumerated.

We have in mind here the real world where we make scientific calculations according to the available data, but where the eventual answers we get do not depend on the order of discovery of these data.

These two viewpoints appear to gives us *two* models, though it can be easily proved that both are equivalent. Another important fact is that for computations relative to total functions, the computational models based on deterministic or non-deterministic machines are the same, so there is a natural embedding of the structure $\boldsymbol{\mathcal{D}}$ of the Turing degrees within the extended *enumeration degree* structure $\boldsymbol{\mathcal{D}}_e$. There are a number of deep

and intractable open problems concerning the relationship between these two structures.

Of course, there are underlying philosophical questions here concerning incomputability in mathematics and real life: *How rich a variety of unsolvable problems is there? Does incomputability impinge on everyday life? And if so, can we find an informative theory of incomputability?*

There are various ways in which incomputability, and its corresponding mathematical structures, may impinge on the real world. And ways in which incomputability, even if present in everyday phenomena, may not be very relevant to our understanding of what we can compute.

For instance, it is well-known that there are what seem to be computationally insurmountable obstacles to the prediction of the weather, even a few days ahead. It might be useful to know the seriousness of those obstacles. But is it useful to know that weather is actually mathematically *incomputable*, as opposed to *in-principle* computable, but computationally *complex* enough for it to be practically incomputable? Presumably, the answer is "no", unless there is something different about the mathematics of the incomputable which distinguishes it from that of the computationally computable, but complex. Two important considerations here are: Firstly, that a scientific world in which causality is basic gives rise to the mathematics of *reducibilities*, with a very rich theory; and secondly, computability theory tells us that the mathematics of different reducibilities can be very different — and, in fact, different reducibilities (which are not clearly notational variants of each other) *invariably* give rise to mathematical structures with very different characteristics. So the mathematics of causality — which is what computability theory is at the level of reality we comfortably inhabit — gives rise to an underlying structure which may (or may not) be very important to an understanding of the world we live in.

Anyway, the above questions can be seen to underlie many bitter debates in science and mathematics, and prominent figures can be found ranged on both sides of this controversy. It has to be said that there are as yet no generally agreed answers to these questions, but quite a lot of pointers to positive ones. But there exist fascinating discussions concerning extensions of the Church–Turing Thesis to the material Universe (see Section I.8 of Volume I of Odifreddi's book [15] on *Classical Recursion Theory*) and of incomputability in Nature (see, for example, Roger Penrose's *The Emperor's New Mind*, [17]).

Even more divisive is the debate as to how the human mind relates to practical incomputability. The unavoidable limitations on computers suggest that mathematics — and life in general — may be an essentially *cre-*

ative activity which transcends what computers do.

The basic inspiration for Alan Turing's computing machines was, of course, the human mind, with things like "states of mind" feeding into the way he described the way his machines worked. Turing made clear in a number of places which side of the argument he was on. On the other, we feel subjectively that our mental processes are not entirely mechanical, in the sense that a Turing machine is. And various people have explicated these feelings to a point where it can be argued convincingly that these feelings have more than purely subjective content. For instance, there is the famous and influential book [10] of Jacques Hadamard on *The Psychology of Invention in the Mathematical Field*, or the philosophically remarkable *Proofs and Refutations: The Logic of Mathematical Discovery* [13] by Imre Lakatos. In science, Karl Popper effectively demolished the inductive model of scientific discovery — as was accomplished, more debatably, by Thomas Kuhn [12] at the social level. This raises the question of how to model the way theories are hypothesised, via a process which seems neither random nor simply mechanical.

A purely mathematical answer to the question is very difficult. Roger Penrose (in his *Shadows of the Mind*, [18]) has argued (unsuccessfully it seems) that the overview we have of Gödel's Incompleteness Theorem for axiomatic arithmetic shows that the human mind is not constrained by that theorem. But it is hard to be clear what it is that the human mind may be doing that Turing machines are incapable of. Obviously it will help to know more about both the physical and the logical structures involved. What is really needed is an *alternative* mathematical model to that of the Turing machine, and providing this must be one of the main aims of computability theory. Some speculations in this direction are provided in the 2003 paper [4] by myself and George Odifreddi on *Incomputability in Nature*.

A large part of the scientific enterprise is bringing plausible descriptions of reality within practical computational frameworks. As we have seen, it is easy to describe classically incomputable objects from algorithmic ingredients, raising the question of to what extent this is mirrored in real-world. It is also the case that in mathematics, the links between computability and descriptions in natural languages is an intimate one which has been extensively mapped out.

Descriptions in natural languages give rise to various hierarchies. At a very basic level, one can build the *arithmetical hierarchy* by starting with computable relations on numbers — essentially, general statements about numbers which do not involve quantifiers — and adding existential and universal quantifiers. There are other hierarchies which mix language and

computational elements. Degree structures and hierarchies are two complementary ways of looking more closely at the universe of incomputable — or computable — objects.

We are all familiar with the hierarchical structure of science itself. Within the life sciences, say, we have the fragmented focus on the quantum level, on atoms, on molecules, on cells, on multicellular organisms, on social structures. Within this descriptive framework the dynamic relationships at each level have to be investigated within the local constraints operating at each level. From a basic mathematical perspective, we find different levels of the arithmetical hierarchy reveal an analogous dynamic infrastructure — its analysis based on a detailed examination of algorithmic relationships.

4 The Turing Universe

Degree structures and hierarchies provide two complementary ways of looking more closely at the universe of incomputable — or computable — objects.

The former is useful in that is built upon and models the basic causal structure of the the natural world via *reducibilities*; while the second is important in that it can capture higher levels of algorithmic content, that with a non-local dimension, typically associated with descriptions in natural language rather than with purely algorithmic relationships. The most important reducibility, the one which is sufficient to model basic scientific relationships, such as those underlying Newtonian mechanics, is Turing reducibility, giving us the structure of the Turing universe over the reals.

As described above, Post's mathematically useful first step was to gather together binary reals which are computationally indistinguishable from each other, in the sense that they are mutually Turing computable from each other. This delivered the familiar upper semi-lattice of the *Turing degrees*. This provides us with a a mathematical framework for the causal structures arising from the natural world. And an investigation of the properties of this structure enables us to achieve a mathematically informative view of global aspects of our natural environment, one which is only hinted at by standard physical theories, with all their arbitrary and ad hoc ingredients. When we look at the mysterious emergence of structure in nature, either subatomic laws, or the richness of life forms, or large-scale galactic or super-galactic structures, we are not just looking at information, but at expressions of patterns of a universal nature. And patterns the origins of which science is as yet unable to explain.

When we inspect the intricacies of the Cat's Eye Nebula, say, as revealed by the Hubble Space Telescope, we feel we should be able to explain the remarkable complexity observed on the basis of our understanding of the local physics. The intuition is that it should be possible to describe global relations in terms of local structure, so capturing the emergence of large-scale structure. The mathematics pertaining to any particular example should be framed in terms of the specific interactive structure on which it is based. But if one wants to reveal general characteristics, and approach deep problems around the emergence of physical laws and constants, which current theory fails to do, one needs something more fundamental. This is where the computability theorist can contribute a basic understanding, in the same way that Turing gave the early developers of the stored-program computer a consciousness of what they were doing via the concept of the universal Turing machine.

We describe some basic properties of the Turing universe as structured by Post. Within the Turing degrees, one has the degree $\mathbf{0}$ of the computable reals, and then the degree $\mathbf{0}'$ of familiar unsolvable problems such as the Halting Problem. We call $\mathbf{0}'$ the *Turing jump* of $\mathbf{0}$. In fact, given a Turing degree $\mathbf{a}$ and a set $A \in \mathbf{a}$, one can *relativise* the halting problem by considering it for a universal Turing machine U with oracle A. This will give us a *halting set* A' of inputs for which U computes, its Turing degree $\mathbf{a}'$ being the *Turing jump* of $\mathbf{a}$. Of course, one can apply this jump operation to $\mathbf{0}'$ and get $\mathbf{0}''$, $\mathbf{0}'''$, $\ldots$, $\mathbf{0}^{(\mathbf{n})}$, $\ldots$ etc.

The nice thing is it turns out that the Turing jump is closely related to how one uses the language of simple high-school arithmetic, codified within the arithmetical hierarchy. This is the basis of *Post's Theorem*, which relates statements in first-order arithmetic to iterations of the Turing jump. For instance, if one wanted to test arbitrary Turing machines to see if they were defined on every input, one would have to decide a statement which involved a universal quantifier followed by an existential quantifier, and this would require an oracle at the $\mathbf{0}''$ level.

Particular interest attaches to sets of numbers which can be *computably enumerated*. This is because many naturally occurring mathematical problems seem to involve such sets, including the Halting Problem. And in the context of the arithmetical hierarchy, they are the sets definable from a computable relation using just one existential quantifier, and are called Σ^0_1 sets — where the subscript tells us there is just one quantifier, and the superscript tells us it is just number variables we are quantifying over. There are many unsolved problems relating to its degree structure $\boldsymbol{\mathcal{E}}$.

Many naturally occurring incomputable sets turn out to be computably

enumerable. A basic question motivating research since the earliest days is: *Just how rich is the Turing structure of the computably enumerable sets?*

There are very different ways of looking at this question, each with its own strengths and technical beauties. An intuitively satisfying approach — first tried by Emil Post — is to look for links between natural information content and relations on $\boldsymbol{\mathcal{D}}$. Another is to delve into the intricacies of $\boldsymbol{\mathcal{D}}$ by directly constructing interesting features of the Turing universe. It is the relationship between these approaches which seems to have a special potential for modelling aspects of the material Universe. This is an area in which the techniques are quite hard to handle even at the classical level — and it is not surprising that their wider potential is largely unrealised.

One approach involves the search for richness of information corresponding to local degree theoretic structure. Ideally we would like something corresponding to the arithmetical hierarchy below $\emptyset'$. One can use *jump inversion* to bring aspects of that very natural hierarchy down to the local level. The resulting *high/low hierarchy* provides an invaluable frame of reference at the local level. But it is hard to characterise in terms of *natural* information content, or to describe in the local structure of $\boldsymbol{\mathcal{D}}$.

(1) The *high/low hierarchy* is defined by

$$\mathbf{High}_n = \{\mathbf{a} \leq \mathbf{0}' \mid \mathbf{a}^{(n)} = \mathbf{0}^{(n+1)}\}, \quad \mathbf{Low}_n = \{\mathbf{a} \leq \mathbf{0}' \mid \mathbf{a}^{(n)} = \mathbf{0}^{(n)}\},$$

for each $n \geq 1$.

(2) If $\deg(A) \in \mathbf{High}_n$ we say A and $\deg(A)$ are $high_n$. We similarly define the low_n sets and degrees. For $n = 1$ we often drop the subscript — A and $\deg(A)$ are *low* if $A' \in \mathbf{0}'$, and *high* if $A' \in \mathbf{0}''$.

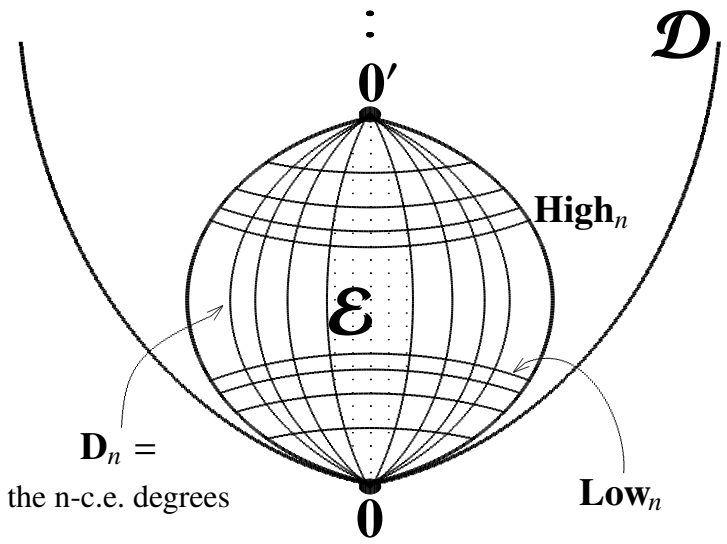
Intuitively, $\mathbf{a}$ is high_n or low_n according as $\mathbf{a}^{(n)}$ takes its greatest or least possible value.

If the high/low hierarchy is thought of as defining a *horizontal* stratification of $\boldsymbol{\mathcal{D}}$ below $\mathbf{0}'$, there is another very important hierarchy whose effect is *vertical*. The *n-c.e. hierarchy* — independently devised by Putnam and Gold around 1965 — inductively builds on the way we can form new sets using boolean combinations of c.e. sets. We get d.c.e. sets — or equivalently 2-c.e. sets — from the c.e. sets by forming *differences* $A - B$ of c.e. sets. More generally, an n-c.e. set is got via boolean operations on c.e. sets allowing up to n differences. Another way of looking at this generalisation of the notion of computably enumerable is that the n provides a bound on the number of mistakes one is allowed to make in computing the status of

potential members of the set. So for a c.e. set A we can make just one mistake, deciding on at most one occasion to change the status of a number from being not in A to being in A.

A degree **a** is *d.c.e.* if it contains a d.c.e. set. And for each $n \geq 2$ we write $\mathbf{D}_n$ for the n$^{\text{th}}$ level of the corresponding *n-c.e. hierarchy* of n-c.e. degrees.

This is the local framework we get from these two basic hierarchies:



One should not be misled by the diagram into thinking that either of the hierarchies eventually includes all of the Turing universe below $\mathbf{0}'$. Far from it. This can only be achieved by extending the levels of the hierarchies into the transfinite, as was done by Yuri Ershov [8], using the constructive ordinals to notate the different levels, in a sequence of three papers in the period 1968–70.

There are other hierarchies based on notions of randomness, forcing, etc., though the local significance of these is limited or unclear. Randomness-related notions have led to some surprising refinements of the low and high degrees. Further information on these can be found in two comprehensive books on computability and randomness, one by Rodney Downey and Denis Hirschfeld [7], and another by André Nies [14].

5 Post's Programme

As we have seen, basic to the relevance of computability theory is the investigation of the extent to which it explains and structures naturally arising information, initially within the mathematical context. We now have some fine hierarchies which provide invaluable landmarks in this exploration.

The programme of mapping out the various links between information and computability-theoretic structure can be traced back to the seminal 1944 Bulletin of the American Mathematical Society paper of Emil Post (in [19]). It was Post — before anyone had discovered the local hierarchies — who pointed the way. Post's work is still important to us, and his approach is relevant to basic science. And certain difficult technical problems in computability theory promise to have far-reaching implications.

Primary ingredients of science include firstly *observation* — that is, our experience of interacting with the Universe. And then mathematical *descriptions*, or information content, pinning down plausible relationships on the Universe in a widely communicable form. Computability is intrinsic to both, and at the same time stands outside, the theory providing a level of meta-science.

Process, causality, algorithmic content — all basic aspects – perhaps *the most* basic aspect of the real world of observation. And it is computability theory — suitably fleshing out and qualifying the Church–Turing Thesis — which mathematically models this. But this is not the only such modelling process. Science routinely builds much more specific mathematical models of natural phenomena, codifying all sorts of observed data into general laws. What is different about computability is that it also has something to say about this extraction of information content as an aspect of the real world. It has the potential to explain how this *information* content — natural laws — relates to the basic *algorithmic* content of the Universe.

The technical expression of this relationship is the notion of *Turing definability*. It is basic to understanding how the beautiful descriptions science gives us of the real world actually derive their material substance.

Definability in a structure is a key mathematical concept, and not widely understood. It is easy to give an intuitive idea of what definability is and how it relates to another useful notion, that of *invariance*. This is not necessarily because the notions are very simple ones, but because they do correspond to phenomena in the real world which we already, at some level, are very familiar with.

As one observes a rushing stream, one is aware that the dynamics of the individual units of flow are well understood. But the relationship between

this and the continually evolving forms manifest in the streams surface is not just too *complex* to analyse — it seems to depend on globally emerging relationships not derivable from the local analysis. The form of the changing surface of the stream appears to constrain the movements of the molecules of water, while at the same time being traceable back to those same movements. The mathematical counterpart is the relationship between familiar operations and relations on structures, and globally arising new properties based on those locally encountered ones. The emergence of form from chaos, of global relations within turbulent environments, is a particularly vivid metaphor for the assertion of definability, or invariance. Let us take a simple mathematical example from arithmetic.

Given the usual operation + of addition on the set $\mathbb{Z}$ of integers, it is easy to see that the set Ev of even integers is describable from + within $\mathbb{Z}$ via the formula

$$x \in Ev \iff (\exists y)(y + y = x).$$

So all we mean by a relation being *definable* from some other relations and/or functions on a given domain is that it can be *described* in terms of those relations and/or functions in some agreed standard language. Of course, there are languages of varying power we can decide on. In the above example, we have used very basic first order language, with finitary quantification over individual elements — we say that Ev is *first order definable* from + over $\mathbb{Z}$. What has happened is that we started off with just an arithmetical operation on $\mathbb{Z}$ but have found it distinguishes certain subsets of $\mathbb{Z}$ from all its other subsets. Intuitively, we first focused on a dynamic flow within the structure given locally by applications of the form $n + m$ to arbitrary integers m, n. But then, standing back from the structure, we observed something global — $\mathbb{Z}$ seemed to fall into two distinct parts, with flow relative to even integers constrained entirely within Ev, and flow from outside Ev being directed into Ev — with Ev being a maximal such subset of $\mathbb{Z}$. From within the structure, + is observable and can be algorithmically captured. Further than that, we are dealing with "laws" which cannot be related to the local without some higher analysis. This feature of the integers is not, of course, a deep one, but it does act as a basic metaphor for other ways in which more or less unexpected global characteristics of structures emerge quite deterministically from local infrastructure.

For the following definition, the *first order language* for $\boldsymbol{\mathcal{D}}$ is one with just the basic variables, brackets, quantifiers and logical connectives, and one 2-place symbol for the ordering $\leq$.

(1) Let $R(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ be a relation on $\boldsymbol{\mathcal{D}}$.

We say that R is *Turing definable* — or *definable in* $\boldsymbol{\mathcal{D}}$ — if there is some first order formula $\varphi(x_1, \ldots, x_k)$ in the language for $\boldsymbol{\mathcal{D}}$ such that for all $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \boldsymbol{\mathcal{D}}$ $R(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ holds if and only if the formula $\varphi(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ holds in $\boldsymbol{\mathcal{D}}$.

(2) If $\mathcal{A}$ is some family of sets, we say $\mathcal{A}$ is *Turing definable* if the set of all Turing degrees of members of $\mathcal{A}$ is Turing definable.

As a simple example, notice that $\mathbf{0}$ is Turing in $\boldsymbol{\mathcal{D}}$ via the formula

$$\varphi(\mathbf{x}) \iff_{\text{defn}} (\forall \mathbf{y})\,[\,\mathbf{x} \leq \mathbf{y}\,].$$

One can, of course, talk about definability in other structures. For instance, $\mathbf{0}_e$ is definable in $\boldsymbol{\mathcal{D}}_e$, and $\mathbf{0}'$ is definable in $\boldsymbol{\mathcal{E}}$.

The notion of *invariance* gives a useful, if slightly more abstract, way of looking at definability. Being able to uniquely *describe* a feature of a structure is a measure of its uniqueness. But some feature of a structure may be quite unique, without one being able to describe that uniqueness in everyday language. Mathematically, we use the notion of *automorphism* to capture the idea of a *reorganisation* of a structure which does not change any of its properties. A feature of that structure is *invariant* if it is left fixed by any automorphism of the structure. Obviously if one can uniquely describe such a feature, it must be invariant, but not necessarily conversely.

(1) Let $R(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ be a relation on $\boldsymbol{\mathcal{D}}$.

Then R is *Turing invariant* if for every automorphism $\psi : \boldsymbol{\mathcal{D}} \to \boldsymbol{\mathcal{D}}$ and every $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \boldsymbol{\mathcal{D}}$ we have $R(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ holds if and only if we have $R(\psi(\mathbf{a}_1), \ldots, \psi(\mathbf{a}_k))$ holds in $\boldsymbol{\mathcal{D}}$.

(2) If $\mathcal{A}$ is some family of sets, we say $\mathcal{A}$ is *Turing invariant* if the set of all Turing degrees of members of $\mathcal{A}$ is Turing invariant.

Only in recent years have we become aware that much of the past fifty years' research into computability has actually been about Turing definability and invariance. Current research focuses on getting optimal Turing definitions of the various levels of the n-c.e. hierarchies we have been looking at. There are also deep questions concerning the nature of the automorphism groups pertaining to different degree structures.

For more background information the joint article [4] by myself and George Odifreddi is an approachable source. For more technical material try volume II of Odifreddi's mammoth *Classical Recursion Theory*, [15].

6 Definability, Invariance and the Real World

What is far from clear is the extent to which the hierarchies and fine-structure theories built from simple computational ingredients are mirrored in the material world. On the other hand, the recognition of the real-world computational content of definability does seem to offer a widely applicable explanatory potential.

It is not surprising that attention has turned to Turing's universe of computably related reals as providing a model for scientific descriptions of a computationally complex real universe.

Let us now return to what the Turing model can do. Let us try to be more clear about how, from very simple beginnings, we can get from the basic fact of existence to what is for us an even greater puzzle — because we have to take what is happening under the umbrella of sufficient reason — the quite amazing emergence of individual entities. From this point of view, it is not quantum ambiguity which is surprising, but the existence of the well-defined world of our everyday experience.

More generally, we have the problem that even though we have natural laws to help us understand much of what happens in the universe, we have no idea where those laws themselves come from. Their apparent arbitrariness lies at the root of the present day confusion of speculative science, verging on the metaphysical.

For Alan Guth [9] in 1997, the problem is:

> *" If the creation of the universe can be described as a quantum process, we would be left with one deep mystery of existence: What is it that determined the laws of physics? "*

While in 1987 Roger Penrose [16] asks for a *strong determinism*, according to which:

> *" . . . all the complication, variety and apparent randomness that we see all about us, as well as the precise physical laws, are all exact and unambiguous consequences of one single coherent mathematical structure."*

The match between mathematics and experience has become much more all-embracing, with string theory perhaps the most ambitious of the attempts to unify the two. The Turing model may be as yet very far from clarifying the specific details of relativity or quantum theory, but it does promise a release from the arbitrariness to which all less basic theories — superstring theory, M-theory, inflation, decoherence, the pilot wave, gauge theory, etc. — are subject, and is based almost entirely upon experience.

What the Turing model primarily tells us about is not an emergence of particular events from events, but of natural laws from the structure of information content.

What does the Turing model suggest regarding the basic structure of matter and the laws governing it?

What we know of the Turing universe is consistent with the possibility that the information content or level of interactivity of a given entity may be insufficient to guarantee it a unique relationship to the global structure. This is what one might expect to apply at an early stage in the development of the universe, or at levels where there is not a sufficiently density of interactions to give information a global role. A number of classic experiments on subatomic particles confirm such a prediction. On the other hand, mathematically entangling such low level information content, perhaps with content at levels of the Turing universe at which rigidity sets in, will inevitably produce new content corresponding to a Turing invariant real. The prediction is that there is a level of material existence which does not display such ambiguity as seen at the quantum level, and whose interactions with the quantum level have the effect of removing such ambiguity — confirmed by our everyday experience of a classical level of reality, and by the familiar 'collapse of the wave function' associated with observation of quantum phenomena. Since there is no obvious mathematical reason why quantum ambiguity should remain locally constrained, there may be an apparent non-locality attached to the collapse. Such a non-locality was first suggested by the well-known Einstein-Podolsky-Rosen thought experiment, and, again, has been confirmed by observation. The way in which definability asserts itself in the Turing universe is not known to be computable, which would explain the difficulties in predicting exactly how such a collapse might materialise in practice, and the apparent randomness involved.

As we have already mentioned, the Turing model may have implications for how the laws of nature immanently arise. And also how they collapse near the big bang 'singularity', and the occurrence or otherwise of such a singularity. What we have in the Turing universe are not just invariant individuals, but a rich infrastructure of more general Turing definable relations. These relations grow out of the structure, and constrain it, in much the same sort of organic way observable in familiar emergent contexts. These relations operate at a universal level. The prediction is that a Universe *with sufficiently developed information content* to replicate the defining content of the Turing universe will manifest corresponding material relations. The existence of such relations one would expect to be

susceptible to observation, these observations in turn suggesting regularities capable of mathematical description. And this is what the history of science confirms. The conjecture is that there is a corresponding parallel between natural laws and relations which are definable in an appropriate fragment of the Turing universe.

The early Universe one would not expect to replicate such a fragment. The homogenisation and randomisation of information content consequent on the extreme interconnectivity of matter would militate against higher order structure. The manifest fragment of the Turing universe, based on random reals, might still contain high information content, but content dispersed and made largely inaccessible to the sort of Turing definitions predicted by the theory. Projected singularities, such as within black holes or associated with boundary states of the Universe, depend on a constancy of the known laws of physics. But immanently originating laws must be of global extraction. This means that their detailed manifestations may vary with global change, and disappear even.

Notice the difference here between what we are saying, and what the upholders of the various versions of Everett's many worlds scenario are. On the one hand, we have an application of the principle of sufficient reason to the world as we know it, which gives a plausible explanation of quantum ambiguity, the dichotomy between quantum and classical reality, and promises some sort of reconciliation between science, the humanities, and our post-modern everyday world. On the other we have something more like metaphysics.

The Turing model, and its connections with emergence, also lead us to expect the familiar fragmentation of science, and human knowledge in general. As we know from computability theory, a Turing definition of a given relation does not necessarily yield a computable relationship with the defining information content. But working within the relations at a given level, there may well be computable relationships emerging, which may become the basis for a new area of scientific investigation. For instance research concerning the cells of a living organism may not be usefully reduced to atomic physics, but deals with a higher level of directly observed regularities. Sociologically, one studies the interactions governing groups of people with only an indirect reference to psychological or biological factors. Entire relations upon cells (humans) defined in some imperfectly understood way by the evolutionary process provide the raw material underlying the new discipline, which seeks to identify a still further additional level of algorithmic content.

There are questions about the range of possibilities embodied in such things as quantum ambiguity: Going from the uniqueness of a defined phenomenon to — what? Are there any overall constraints apart from those imposed by the mathematics specific to the emergent structures? There seems to be one unavoidable rule — obvious when it is pointed out — which is that each superimposed alternative must be viable by itself. Which, in addition to the specifics, demands that the information content develops within the rules experience and the computability theory lead us to expect. In particular, there can be at most countably many such alternatives. The existence of at most coumtably many Turing automorphisms is already known.

For a more detailed review of the deep and fundamental computability-theoretic problems reviewed here see [4].

References

[1] S. B. Cooper. *Computability Theory.* Chapman & Hall/CRC, Boca Raton, London, New York, Washington, D.C., 2004.

[2] S. B. Cooper. Incomputability, Emergence and the Turing Universe. In *Causality, Meaningful Complexity and Knowledge Construction* (A. Carsetti, Ed.), Springer, 2009.

[3] S. B. Cooper. From Descartes to Turing: The Computational Content of Supervenience. To appear in *Information and Computation* (eds. M. Burgin and G. Dodig-Crnkovic), World Scientific Publishing Co.

[4] S. B. Cooper and P. Odifreddi. Incomputability in Nature. In S.B. Cooper and S.S. Goncharov, *Computability and Models*. Kluwer Academic/Plenum, New York, Boston, Dordrecht, London, Moscow, 2003, pages 137–160.

[5] M. Davis. *The Universal Computer: The Road from Leibniz to Turing.* W.W. Norton, New York, London, 2000.

[6] D. Deutsch. Quantum theory, the Church Turing principle, and the universal quantum computer. Proc. Roy. Soc., A 400 (1985), 97–117.

[7] R. Downey and D. Hirschfeldt, *Algorithmic Randomness and Complexity*, to appear.

[8] Y. L. Ershov, A certain hierarchy of sets, i, ii, iii. (Russian), *Algebra i Logika*, **7** (1968), 47–73; **7** (1968), 15–47; **9** (1970), 34–51.

[9] Guth A. H. (1997) *The Inflationary Universe – The Quest for a New Theory of Cosmic Origins*. Addison-Wesley, New York, Harlow, England, Tokyo, Paris, Milan.

[10] J. Hadamard. *The Psychology of Invention in the Mathematical Field*. Princeton Univ. Press, Princeton, 1945.

[11] A. Hodges. *Alan Turing: The Enigma.* Vintage, London, Melbourne, Johannesburg, 1992.

[12] T. S. Kuhn, *The Structure of Scientific Revolutions*, Third edition 1996, University of Chicago Press, Chicago, London.

[13] I. Lakatos. *Proofs and Refutations.* Cambridge University Press, 1976.

[14] A. Nies, *Computability and Randomness*, Oxford University Press, 2010.

[15] P. Odifreddi. *Classical Recursion Theory*, Volumes I and II. North-Holland/Elsevier, Amsterdam, New York, Oxford, Tokyo, 1989 and 1999.

[16] R. Penrose, Quantum physics and conscious thought, in *Quantum Implications: Essays in honour of David Bohm* (BJ̇. Hiley and F.D. Peat, eds.), Routledge & Kegan Paul, London, New York, 1987, pp. 105–120.

[17] R. Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* Oxford University Press, Oxford, New York, Melbourne, 2002.

[18] Penrose R. (1994) *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press, Oxford.

[19] E. L. Post. *Solvability, Provability, Definability: The Collected Works of Emil L. Post* (Martin Davis, Editor). Birkhäuser, Boston, Berlin, 1994.

[20] R. I. Soare. *Recursively Enumerable Sets and Degrees.* Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1987.

[21] A. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Vol. 42, 1936–37, pages 230–265; reprinted in A. M. Turing, *Collected Works: Mathematical Logic*, pages 18–53.

[22] A. Turing, Systems of logic based on ordinals, *Proceedings of the London Mathematical Society*, Vol. 45, 1939, pages 161–228; reprinted in A.M. Turing, *Collected Works: Mathematical Logic*, pages 81–148.

[23] A. M. Turing. *Collected Works: Mathematical Logic* (R.O. Gandy and C.E.M. Yates, Editors). Elsevier, Amsterdam, New York, Oxford, Tokyo, 2001.

[24] P. P. Wiener (Editor). *Leibniz: Selections*, Charles Scribner's Sons, New York, 1951.

Algebraic Logic

HIROAKIRA ONO*

1 Introduction

Algebraic methods have been important tools in the study of nonclassical logics, in particular propositional logics. In the present paper, we will explain what algebraic approaches are, where their special features lie, and how effectively they are applied, touching on basic concepts in algebraic logic and universal algebra. In the last section we mention briefly a recent development of substructural logics, where algebraic methods have made a remarkable success. Topics taken up in the following are roughly ordered in a chronological way.[1]

2 Algebras as models of logic

In this section, we will explain what is an algebraic approach to logic at all, by considering three important logics, i.e., classical, intuitionistic and Łukasiewicz's many-valued logics.

2.1 Boolean algebras for classical logic

A standard way of introducing a logic is to give it through a formal system (or a calculus) for it. A formal system consists usually of *axioms* and *rules of inference*, and the *provability* of formulas in the system is determined by them. There are many different types of formal systems, e.g. Hilbert-style formal systems, natural deductions, and sequent systems. Classical logic can be introduced not only by these syntactical ways but also by semantical ways. The most popular way among them is to understand it as the set of all *tautologies*. Here, a tautology is a propositional formula which always takes the value 1 (which denotes the *truth*) always for every

*Research Center for Integrated Science, Japan Advanced Inst. of Sc. and Technology

[1]The author would like to thank Johan van Benthem for his valuable suggestions, and also Leonardo Cabrer, José Gil-Férez and Norbert Preining for their helpful comments.

valuation (or truth assignment) of propositional variables appearing in it. Here, a valuation is a mapping from the set of propositional variables to the set $\{0, 1\}$ of truth values, where 0 denotes the *falsehood*. The truth table of each logical connective, which is given below, determines how to extend a valuation to a mapping from the set of formulas to the set $\{0, 1\}$.

$$a \wedge b = \min\{a, b\},\ a \vee b = \max\{a, b\},\ \neg a = 1 - a,$$

$a \to b = \max\{1 - a, b\}$ (= 1 if $a = 0$ or $b = 1$, and = 0 otherwise).

For instance, the formula $p \vee \neg p$ is a tautology, since for every valuation f the value $f(p) \vee \neg f(p)$ is equal to $\max\{f(p), 1 - f(p)\}$, which is always 1, while the formula $(p \vee q) \to p$ is not a tautology, as it takes th value 0 by such a valuation g that $g(p) = 0$ and $g(q) = 1$. The following completeness says that a given formal system of classical logic can capture exactly tautologies by the provability in it.

Theorem 2.1. *The following are mutually equivalent for each propositional formula α.*

1. *α is provable in classical logic.*
2. *α is a tautology.*

We call the above structure consisting of the set $\{0, 1\}$ with the truth table, $\mathbf{B}_2$, which gives us a miniature of algebraic models for classical logic. Algebraic models for classical logic in general are called *Boolean algebras*. In particular, $\mathbf{B}_2$ is called the *2-valued Boolean algebra*. Formally speaking, a Boolean algebra is a structure $\mathbf{A} = \langle A, \wedge, \vee, \neg, 0, 1\rangle$ such that $\langle A, \wedge, \vee\rangle$ is a distributive lattice satisfying that for any $a \in A$

$$a \wedge \neg a = 0,\ a \vee \neg a = 1,\ a \wedge 0 = 0 \text{ and } a \vee 1 = 1.$$

In each Boolean algebra, 0 and 1 are shown to be the least and the greatest elements, respectively, with respect to the partial order $\leq$ induced by the lattice operations. We define $a \to b$ by $\neg a \vee b$. By using the distributivity we can show the following *law of residuation*;

$a \wedge b \leq c$ if and only if $a \leq b \to c$.

It is easy to see that $\neg a = a \to 0$, $\neg 0 = 1$ and $\neg\neg a = a$. For more information on lattices and Boolean algebras, see [5].

For a given Boolean algebra $\mathbf{A}$, we can generalize the notion of valuations over it as follows. A valuation over $\mathbf{A}$ is a mapping from the set of

all propositional variables to A. Then each valuation can be extended to a mapping from the set of all formulas to A by interpreting each logical connective by the corresponding algebraic operation. We say that a formula α is valid in a Boolean algebra when $f(\alpha) = 1$ for every valuation f over it. Clearly, a formula is a tautology if and only if it is valid in the Boolean algebra $\mathbf{B}_2$. Now Theorem 2.1 can be strengthened into the following.

Theorem 2.2. *The following are mutually equivalent for each propositional formula α.*

1. *α is provable in classical logic.*
2. *α is valid in every Boolean algebra.*
3. *α is a tautology.*

By the definition, every Boolean algebra contains 0 and 1 as its elements. Therefore, the Boolean algebra $\mathbf{B}_2$ is the simplest one among Boolean algebras in which $0 \neq 1$ holds. In algebraic terms we say that $\mathbf{B}_2$ is a *subalgebra* of every Boolean algebra in which $0 \neq 1$ holds. It is interesting to see that there are many Boolean algebras, and thus there are many algebraic models of classical logic. For instance, the powerset $\wp(X)$ of any given set X with set-theoretic operations forms a Boolean algebra, called a *field of sets*. (When X is a singleton set the field of sets becomes isomorphic to $\mathbf{B}_2$.) The set of all finite subsets together with all cofinite subsets (i.e., subsets whose complement is finite) of a set X forms another Boolean algebra, called a finite-cofinite algebra. Note that the cardinality of any field of sets is either finite or uncountable, while a finite-cofinite algebra of a countable set X is countable. In fact, a finite-cofinite algebra of an infinite set X has the same cardinality as X. Thus, there exists a Boolean algebra with an arbitrary infinite cardinality.

But one may ask why we need to consider such big Boolean algebras, since Theorem 2.2 tells us that we can get nothing new by considering various Boolean algebras as concerns the validity of formulas. As a matter of fact, once we consider extensions of classical propositional logic, e.g. algebraic models of modal logics on classical logic, these Boolean algebras of different types will play an essential role. Also, the way of introducing algebraic models of classical logic gives us a prototype when we consider algebraic models of other logics. As shown in the next subsection, the argument goes similarly but there exists a clear difference in the case of intuitionistic logic. (Compare Theorem 2.3 with Theorem 2.2.)

2.2 Heyting algebras for intuitionistic logic

Algebraic considerations made in the previous subsection can be applied to other logics. We will discuss them here for intuitionistic logic, which is the logic of constructive reasoning by L.E.J. Brouwer and whose formal system was introduced around 1930. It is a sublogic of classical logic. According to the constructive reasoning, the truth of formulas must be established by a proof. For instance, a proof of the disjunctive formula $\alpha \vee \beta$ is obtained by giving a proof of either α or β and showing which one holds. Thus, the axiom of excluded middle $\alpha \vee \neg\alpha$ of classical logic is not provable in intuitionistic logic. Similarly, the axiom of the double negation $\neg\neg\alpha \to \alpha$ is rejected. For information on intuitionistic logic and its formal systems, see [4].

The constructive feature of intuitionistic logic mentioned above is reflected as the *disjunction property* of the logic. That is, if a formula $\alpha \vee \beta$ is provable in intuitionistic logic then either α or β is provable in it. Note that classical logic does not have it, since $p \vee \neg p$ is provable in classical logic while neither p nor $\neg p$ is provable in it for any variable p.

Algebraic models for intuitionistic logic are *Heyting algebras*. A Heyting algebra is a structure $\mathbf{A} = \langle A, \wedge, \vee, \to, 0, 1\rangle$ such that $\langle A, \wedge, \vee\rangle$ is a lattice satisfying that

1) $a \wedge 0 = 0$ and $a \vee 1 = 1$, for all $a \in A$
2) $a \wedge b \leq c$ if and only if $a \leq b \to c$, for all $a, b, c \in A$

As a matter of fact, when $\mathbf{A}$ is a Heyting algebra, $\langle A, \wedge, \vee\rangle$ is a distributive lattice. This is shown by using the condition 2). We define $\neg a$ by $a \to 0$. Then $a \wedge \neg a = 0$ holds but $a \vee \neg a = 1$ does not hold in general. To see this, let us consider the partially ordered set $\{0, c, 1\}$ satisfying $0 < c < 1$. We define $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$, and $x \to y = 1$ if $x \leq y$ and $= y$ otherwise. Then, we can show that this structure, called $\mathbf{H}_3$, is a Heyting algebra. Since $\neg c = 0$, $c \vee \neg c = c \neq 1$. We can see that Boolean algebras are Heyting algebras that satisfy $a \vee \neg a = 1$. The validity of formulas in Heyting algebras is defined in the same way as the validity in Boolean algebras.

Let A be any linearly ordered set with the least element 0 and the greatest element 1. Then it forms a Heyting algebra by defining $\wedge, \vee$ and $\to$ in the same way as those in $\mathbf{H}_3$. It is easy to see that the formula $(p \to q) \vee (q \to p)$ is valid in every such linearly ordered Heyting algebra, since either $f(p) \leq f(q)$ or $f(q) \leq f(p)$ holds always. On the other hand there

exist Heyting algebras in which $(p \to q) \vee (q \to p)$ is not valid. Similarly to Theorem 2.2, we can show the following.

Theorem 2.3. *The following are mutually equivalent for each propositional formula α.*

1. *α is provable in intuitionistic logic.*
2. *α is valid in every Heyting algebra.*
3. *α is valid in every finite Heyting algebra.*

Condition 1 implies 2, which implies 3. But the converse directions are not so trivial. The proof that Condition 1 follows from 2, which is called completeness of intuitionistic logic with respect to the class of Heyting algebras, is given below.

Lindenbaum-Tarski algebra – a universal way of showing completeness

There is a standard technique of showing algebraic completeness of a given logic which uses Lindenbaum-Tarski algebra. We will give a brief outline of it for the case of intuitionistic logic. Let Φ be the set of all formulas. We define a binary relation $\equiv$ on Φ, putting $\alpha \equiv \beta$ if the formulas $\alpha \to \beta$ and $\beta \to \alpha$ are both provable in intuitionistic logic. When $\alpha \equiv \beta$ holds, we say that α and β are logically equivalent (in intuitionistic logic). We can show that the relation $\equiv$ is an equivalence relation on Φ. In fact, it is a congruence relation, i.e., an equivalence relation which is compatible with all logical connectives, and moreover it has the following property: if $\alpha \equiv \beta$ then $\sigma(\alpha) \equiv \sigma(\beta)$ for any substitution σ. Roughly speaking, logical equivalence of two formulas α and β means that they are indistinguishable in intuitionistic logic, and hence any replacement of one by the other is harmless.

These facts enable us to get an algebra $\mathbf{C}$ whose underlying set is the set of all equivalence classes with respect to $\equiv$, on which algebraic operations can be defined in a consistent way, since $\equiv$ is a congruence relation. (That is, $\mathbf{C}$ is obtained from Φ by identifying indistinguishable formulas.) An important point here is that the element 1 of the Heyting algebra $\mathbf{C}$ is the equivalence class whose members are exactly formulas provable in intuitionistic logic. Now, to show that Condition 2 implies 1 by taking the contraposition, we assume that a formula α is not provable in intuitionistic logic. Define a valuation g over $\mathbf{C}$ by $g(p)$ = "the equivalence class to which p belongs." By induction, we can show that for any formula φ,

$g(\varphi)$ = "the equivalence class to which φ belongs." Then, in particular $g(\alpha)$ cannot be the equivalence class 1. Therefore, α is not valid in the Heyting algebra **C**.

The Heyting algebra **C** is known as the *Lindenbaum-Tarski algebra* of intuitionistic logic. Though the above argument looks complicated at first sight, it can be applied to a wide class of logics. It suffices to replace the logical equivalence in intuitionistic logic by the logical equivalence in a given logic. In the case of classical logic, its Lindenbaum-Tarski algebra becomes a Boolean algebra. In this way, once we know how to use Lindenbaum-Tarski algebra, showing the algebraic completeness of a propositional logic becomes a routine work in a most case.

From the construction of Lindenbaum-Tarski algebras, one may think that they are simply shadows of the syntax on algebras. But this is not the case. For example, it is known that one variable fragment of the Lindenbaum-Tarski algebra of intuitionistic logic has a beautiful algebraic structure.

Finite countermodels for unprovable formulas

The equivalence of Condition 3 to Condition 1 in Theorem 2.3 means essentially that if a formula α is not provable in intuitionistic logic then there exists a *finite* Heyting algebra **A** which is a countermodel of α. A given logic **L** is said to have the *finite model property* (FMP), when for any formula α, if α is not provable in **L** then there exists a finite algebra **A** such that α is not valid in it while all provable formulas in this logic are valid. So, Theorem 2.3 says that intuitionistic logic has the FMP. A direct proof of the equivalence of Conditions 1 and 3 is obtained by using the *finite embeddability property* of the class of Heyting algebras.

The FMP of a logic **L** is used often to show the decidability of **L**. Here we say that a logic **L** is decidable if there is an effective procedure of deciding whether or not any given formula is provable in **L**. In fact, Harrop's Lemma says that a logic **L** is decidable if it is finitely axiomatizable and has the FMP. Since intuitionistic logic is finitely axiomatizable, Theorem 2.3 implies the decidability of intuitionistic logic.

Different from the case for classical logic where every non-provable formula can be falsified in the finite Boolean algebra $\mathbf{B}_2$, it can be shown that no single finite Heyting algebra can falsify all non-provable formulas in intuitionistic logic.

2.3 Łukasiewicz's many-valued logics – from algebra to logic

Classical logic can be introduced syntactically by using a formal system, and at the same time can be defined semantically as the set of tautologies. Boolean algebras are algebraic structures which are obtained by generalizing the truth tables of the truth values 0 and 1. Intuitionistic logic was introduced firstly as a formal system but then Heyting algebras were introduced as its algebraic models, similarly to Boolean algebras.

On the other hand, sometimes a logic is introduced semantically by using special algebraic structures. Typical examples are Łukasiewicz's many-valued logics. By extending 2-valued truth definition of classical logic, in the1920s J. Łukasiewicz introduced $n+1$-valued logics ($n > 0$) with the set of truth values $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$, and also the infinite-valued logic with the unit interval $[0, 1]$ as the set of truth values. The truth table of each logical connective is defined as follows:

$$a \wedge b = \min\{a, b\},\ \ a \vee b = \max\{a, b\},\ \neg a = 1 - a,$$

$a \to b = \min\{1, 1 - a + b\}$ ($= 1$ if $a \leq b$, and $= 1 - a + b$ otherwise).

In fact, when $n = 1$, these truth tables are equal to the truth tables of classical logic and hence 2-valued Łukasiewicz logic is nothing but classical logic. Later, axiom systems of these many-valued logics were discovered. Algebraic models of these logics, containing the above algebras as special cases, are called *MV-algebras* (many-valued algebras), though we omit the detailed definition.

2.4 Towards a general study of logics

What we have discussed can be summerized briefly as follows. Usually a logic is introduced as a formal system, in which the notion of provability of formulas in that logic is defined syntactically. On the other hand, from a semantical point of view the notion of validity of formulas in a given algebra or algebraic structure is introduced, and then a suitable formal system is devised for *axiomatizing* it, so that the set of all formulas provable in the system is equal to the set of all valid formulas. This is the case for Łukasiewicz's many-valued logics, for instance.

This shows that we are mainly concerned with sets of formulas satisfying certain conditions. In fact, for a given logic **L** algebraic models and algebraic completeness can be defined as follows. An algebra **A** is an *algebraic model* of a logic **L**, if all formulas provable in a logic **L** are valid in it. A

logic **L** is *algebraically complete* if and only if the set of formulas valid in all algebraic models of **L** is equal to the set of formulas provable in **L**.

These considerations lead us a general notion of logics. We define a logic to be a set of formulas satisfying some suitable conditions. As an example, we will introduce *superintuitionistic logics*, or logics over intuitionistic logic, as follows. A set **L** of formulas is a superintuitionistic logic, if **L** satisfies the following:

- **L** contains all formulas which are provable in intuitionistic logic,
- **L** is closed under *modus ponens*, i.e., if both α and $\alpha \to \beta$ belong to **L** then β belongs to **L**,
- **L** is closed under substitution, i.e., if α belongs to **L** then $\sigma(\alpha)$ belongs to **L** for every substitution σ.

Clearly, both classical logic and intuitionistic logic are superintuitionistic logic. Also, for each Heyting algebra **A**, the set **L**(**A**) of all valid formulas in **A** is a superintuitionistic logic. On the other hand, we can show that for each superintuitionistic logic **L** there exists a Heyting algebra **A** such that **L** = **L**(**A**). In fact, we can take the Lindenbaum-Tarski algebra of **L** for this **A**.

In the same way, we can define logics over any given logic, like logics over Łukasiewicz's infinite-valued logic, by assuming some necessary closure conditions in addition to the closure under modus ponens and substitution. Thus we have now many logics. In fact, there exist uncountably many superintuitionistic logics. This implies that many superintuitionistic logics cannot be finitely axiomatized.

Why so many logics?

A natural question may be posed here. Why do we consider so many *logics*? Our initial motivation of studying logic must be to understand several special logics like classical logic and intuitionistic logic, introduced by clear and reasonable motivations, not to know such incomprehensible and philosophically unmotivated logics.

Here I will try to defend such a study, comparing it with the situation in algebra. Historically, arithmetic is one of main subjects in mathematics. We want to know natural numbers, integers, are real numbers and so on, and to understand mathematical structures with arithmetic operations as a whole like the set of integers $\mathbb{Z}$ and the set of real numbers $\mathbb{R}$. But then

in order to get a deeper understanding of these particular structures, it becomes essential to introduce general notions of algebraic structures, which are obtained from these structures by extracting their essence, and discuss these particular structures in relation to them. In this way, groups, rings, fields, real closed fields and algebraically closed fields have been introduced. As we know, this generalization turned out to be extremely useful and successful.

Our logics in a general sense, sometimes collectively called *nonclassical logics*, are introduced by the same way of thinking. We want to get an understanding of logics, say, of superintuitionistic logics as a whole. For example, we know already that the disjunction property holds in intuitionistic logic but not in classical logic. When does a superintuitionistic logic has the disjunction property. Is the disjunction property related to some other properties? In general, what logical properties do they share in common, and how is one property related to another? Do these two logics occupy a special position among superintuitionisitic logics? These are typical questions which we pose for our general and comparative study of logics. Algebraic methods have turned out to be quite helpful in the study, as we show later.

It is easy to see that the intersection of two superintuitionistic logics (as sets of formulas) is also a superintuitionistic logic. In fact, this holds for the intersection of arbitrary number of superintuitionistic logics. Therefore the set of all superintuitionistic logics forms a complete lattice, a lattice in which any subset (of superintuitionistic logics) has the meet and the join. Then we want to understand what lattice structure it forms. Note that the join of two superintuitionistic logics is not always equal to the union of them (as sets of formulas). The greatest element of the lattice is the *inconsistent* logic, which is the set of all formulas. The second greatest is classical logic, and the least one is intuitionistic logic.

2.5 Historical notes

The notion of Boolean algebras is introduced and studied by G. Boole and A. de Morgan in the middle of 19th century. In his Book "*The Law of Thought*," Boole wrote as:

> *the laws of logic are in mathematical form and the logical method should be analoguos to the method of mathematics.*

In modern terms, by *mathematics* he meant *algebra*, and his laws of logic consist of some *true* basic equations, which can be regarded as axioms of Boolean algebras in an equational theory.

From the beginning of the 1920s, algebraic study of logic has been developed mainly by Polish logicians, including A. Tarski, J. Łukasiewicz and A. Lindenbaum. On the ground of this tradition, H. Rasiowa and R. Sikorski published a seminal book "*The Mathematics of Metamathematics*" on algebraic logic in 1963 [12]. See also [11]. The book [10] will offer an elementary guide to the topic.

Nonclassical logics contain many other kinds of logics. Some are developed in the extended language other than the standard logical connectives. One of important branches of them is modal logic, which usually contains unary connectives $\Box$ and $\Diamond$ in the language. Temporal logic and epistemic logic are extensions of modal logic, either of which is getting a quite active, interdisciplinary research field with philosophy, computer science and cognitive science. An accessible introduction to various nonclassical logics from philosophical point of view is given in several chapters of [8].

3 Universal algebraic view of logic

As we have seen, for a given superintuitionistic logic **L**, there are many algebraic models of **L**. To grasp the feature of the class of these algebraic models and to attain high generality, we will introduce some of basic algebraic notions and results on universal algebra. Here we discuss algebras in general, but one may consider them as Heyting algebras, for instance, to facilitate the comprehension. For further information on universal algebra, see [3].

3.1 Basics of universal algebra

An *algebra* is a set equipped with finitary operations. Usually, an algebra $\mathbf{A}$ is expressed as $\mathbf{A} = \langle A, f_1, \ldots, f_n \rangle$ where A is a set called the *universe* and f_1 through f_n are *basic operations*. The sequence $\langle k_1, \ldots, k_n \rangle$ where k_i is the arity of the operation f_i is called the *type* of $\mathbf{A}$. With a given algebraic type we associate a first-order language having the identity as its single relation symbol and also function symbols which correspond to operations of the type. We consider only atomic statements, called *equations*, or *identities*, which are of the form $s = t$, where s and t are terms. A set of equations Σ (in a given algebraic language) determines a class of algebras $\mathrm{Mod}(\Sigma)$,

which consists of all *models* of Σ. A class $\mathcal{K}$ of algebras is called an *equational class* if there exists a set of equations Σ such that $\mathcal{K} = \mathrm{Mod}(\Sigma)$, i.e., $\mathcal{K}$ is exactly the class of all algebras which are models of a set of equations Σ. It is known that monoids, lattices, Boolean algebras and Heyting algebras are defined by some sets of equations. Therefore, each of these classes forms an equational class. On the other hand, the class of fields is not an equational class.

Recall that a mapping $h : A \longrightarrow B$ between universes of algebras of the same type, is a *homomorphism* if it preserves operations, i.e., if for each basic operation f, $h(f^{\mathbf{A}}(a_1, \ldots, a_n)) = f^{\mathbf{B}}(h(a_1), \ldots, h(a_n))$, where $f^{\mathbf{A}}$ and $f^{\mathbf{B}}$ mean the interpretation of f respectively in $\mathbf{A}$ and $\mathbf{B}$. (We drop the superscripts, when no confusions may occur.) An injective homomorphism is called an *embedding*, and a bijective one an *isomorphism*.

We give three basic ways of obtaining new algebras from given one in the following.

- Homomorphic images: An algebra $\mathbf{A}$ is a *homomorphic image* of $\mathbf{B}$, if there is a homomorphism from $\mathbf{B}$ *onto* $\mathbf{A}$.
- Subalgebras: An algebra $\mathbf{A}$ is a *subalgebra* of $\mathbf{B}$, which is denoted by $\mathbf{A} \subseteq \mathbf{B}$, if $A \subseteq B$ and the identity mapping on A is an embedding.
- Direct products: An algebra $\mathbf{A}$ is a *direct product* of the *indexed system* $(\mathbf{A}_i : i \in I)$ of algebras, written $\prod_{i\in I} \mathbf{A}_i$, if the universe A is the direct product $\prod_{i\in I} A_i$ and the operations on A are defined coordinatewise.

3.2 Varieties and equational classes

In the following, when we take a class of algebras, we always assume that every algebra in $\mathcal{K}$ is of the same type. Now for a class $\mathcal{K}$ of algebras, we define $H(\mathcal{K})$, $S(\mathcal{K})$ and $P(\mathcal{K})$ to be the class of all homomorphic images of algebras in $\mathcal{K}$, the class of all subalgebras of algebras in $\mathcal{K}$ and the class of all direct products of algebras in $\mathcal{K}$, respectively. Since H, S and P operate on classes of algebras they are called *class operators*. We can show that each of the class operators H, S and P preserve equations, i.e., every equation valid in all algebras in a class $\mathcal{K}$ is also valid in all algebras in any of $H(\mathcal{K})$, $S(\mathcal{K})$ and $P(\mathcal{K})$.

A class $\mathcal{K}$ of algebras is a *variety* if and only if it is closed under H, S and P. Clearly, the intersection of varieties is also a variety. Therefore, there

exists the smallest variety $V(\mathcal{K})$ containing a given class $\mathcal{K}$ of algebras. To get $V(\mathcal{K})$, it may be necessary to apply these class operators successively to $\mathcal{K}$. But as shown in the following *Tarski's Theorem*, it is enough just to take HSP, i.e., to apply P, S, and H to $\mathcal{K}$ only once in this order.

Theorem 3.1. *For any class $\mathcal{K}$ of algebras, $V(\mathcal{K}) = HSP(\mathcal{K})$.*

This $V(\mathcal{K})$ is called the variety *generated by* $\mathcal{K}$. Now let $\mathcal{K}$ be an equational class, i.e., a class determined by a set of equations. By the fact that H, S and P preserve equations, $V(\mathcal{K}) = \mathcal{K}$ holds and hence it is a variety. Moreover, the converse can be shown. Thus we have the following *Birkhoff's Theorem*.

Theorem 3.2. *A class of algebras is a variety if and only if it is an equational class.*

This theorem shows that the syntactic notion of "equational classes" can be characterized completely by a purely algebraic notion on closure under natural operations. A consequence of this theorem is that if a class $\mathcal{K}$ of algebras is not closed under one of H, S and P, $\mathcal{K}$ is never defined by a set of equations. For instance, it is shown that the class of fields is not an equational class. We know that both the class of all Boolean algebras $\mathcal{B}$ and the class of all Heyting algebras $\mathcal{H}$ are equational classes and thus are varieties. In this case the former is included in the latter. So we can say that the variety $\mathcal{B}$ is a *subvariety* of the variety $\mathcal{H}$. As a matter of fact, all subvarieties of the variety $\mathcal{H}$ forms a complete lattice.

It is well-known that every natural number greater than 1 can be represented by the product of prime numbers. Similarly, every algebra can be represented by the subdirect product of *subdirectly irreducible* algebras. Here we say that $\mathbf{A}$ is a subdirect product of an indexed system $(\mathbf{A}_i : i \in I)$, if $\mathbf{A} \subseteq \prod_{i\in I} \mathbf{A}_i$ and all the coordinate projections are onto (in other words, each $\mathbf{A}_i$ is a homomorphic image of $\mathbf{A}$). In this case, we say that $(\mathbf{A}_i : i \in I)$ is a *subdirect representation* of $\mathbf{A}$. This representation is not unique, since $\mathbf{A}$ itself is always its own subdirect representation. But in some case, $\mathbf{A}$ has only this trivial representation. Precisely speaking, $\mathbf{A}$ is *subdirectly irreducible* if every subdirect representation $(\mathbf{A}_i : i \in I)$ of $\mathbf{A}$ contains (an isomorphic copy of) $\mathbf{A}$ as a factor. The following result was obtained by G. Birkhoff.

Theorem 3.3. *Every algebra has a subdirect representation with subdirectly irreducible factors.*

From the definition, we can see that $\mathbf{A}$ is a member of a variety $\mathcal{V}$ if and only if every subdirectly irreducible factor of $\mathbf{A}$ is also a member of $\mathcal{V}$. Thus every variety is generated by its subdirectly irreducible members.

3.3 Logics and varieties

We have mentioned two notions of validity, i.e., the validity of equations of the form $s = t$ with terms s and t, and the validity of formulas. The first one is defined formally as follows. An equation $s = t$ is valid in an algebra $\mathbf{A}$ if and only if $f(s) = f(t)$ holds for each valuation f on A. But apparently, they are closely related. For instance, let us consider the distributive law in lattices. It is represented by $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$, or equivalently by $a \vee (b \wedge c) \leftrightarrow (a \vee b) \wedge (a \vee c) = 1$ (in e.g. Heyting algebras) as an equation, while it can be respresented by $\alpha \vee (\beta \wedge \gamma) \leftrightarrow (\alpha \vee \beta) \wedge (\alpha \vee \gamma)$ as a formula. Here, $\varphi \leftrightarrow \psi$ is an abbreviation of $(\varphi \to \psi) \wedge (\psi \to \varphi)$.

By *identifying terms and formulas* we can see the following in general in case of Heyting algebras. The validity of an equation $s = t$ is equal to that of the formula $s \leftrightarrow t$ and the validity of a formula α is equal to that of the equation $\alpha = 1$. Based on this fact, in the rest of the paper we identify terms with formulas, which may not cause any confusion as we consider only propositional formulas. Letters s, t, α, β etc. are used to denote both terms and formulas.

We consider the class of all algebraic models of a given superintuitionistic logic $\mathbf{L}$. It is clear that the class of all algebraic models of intuitionistic logic is the class of all Heyting algebras $\mathcal{H}$. Let $\mathcal{V}_{\mathbf{L}}$ be the class of Heyting algebras which are algebraic models of $\mathbf{L}$, i.e., in which all formulas in $\mathbf{L}$ are valid. Then, by Theorem 3.2,

> for any superintuitionistic logic $\mathbf{L}$, $\mathcal{V}_{\mathbf{L}}$ is a subvariety of $\mathcal{H}$.

Conversely, for a given variety $\mathcal{V}$ of Heyting algebras let $\mathbf{L}_{\mathcal{V}}$ be the set of all formulas φ such that φ is valid in every algebra in $\mathcal{V}$. Then,

> for any variety $\mathcal{V}$ of Heyting algebras, $\mathbf{L}_{\mathcal{V}}$ is a superintuitionistic logic.

The mappings $\mathsf{V} : \mathbf{L} \longmapsto \mathcal{V}_{\mathbf{L}}$ and $\mathsf{L} : \mathcal{V} \longmapsto \mathbf{L}_{\mathcal{V}}$ are shown to be dual lattice isomorphisms between the lattice of all superintuitionistic logics and the lattice of all subvarieties of $\mathcal{H}$.

3.4 What follows from algebraic methods

Universal algebra takes a quite general perspective of algebras, and it can shed a new light on logics. At the same time, surprisingly deep results are sometimes obtained from it. In the following, we will give some logical consequences obtained by these algebraic notions and tools.

From a logical point of view, we get the following from the above considerations, by applying them to the case of Heyting algebras.

Lemma 3.4. *If* $\mathbf{B}$ *is a homomorphic image (a subalgebra, respectively) of* $\mathbf{A}$, $\mathbf{L}(\mathbf{A}) \subseteq \mathbf{L}(\mathbf{B})$ *holds between logics* $\mathbf{L}(\mathbf{A})$ *and* $\mathbf{L}(\mathbf{B})$. *If* $\mathbf{A}$ *is a direct product of the* $(\mathbf{A}_i : i \in I)$ *then* $\mathbf{L}(\mathbf{A}) = \bigcap_i \mathbf{L}(\mathbf{A_i})$. *Also, if* $(\mathbf{A}_i : i \in I)$ *is a* subdirect representation *of* $\mathbf{A}$ *then* $\mathbf{L}(\mathbf{A}) = \bigcap_i \mathbf{L}(\mathbf{A_i})$.

Subdirect irreducibility in Heyting algebras

In case of Heyting algebras, it can be shown that a Heyting algebra $\mathbf{A}$ is subdirectly irreducible if and only if it has the second greatest element, i.e., an element a such that $a < 1$ and moreover that $b \leq a$ for any $b < 1$. For example, a Heyting algebra $\mathbf{H}_3$ discussed in Section 2.2 is subdirectly irreducible. Moreover, it is easy to see that every subdirectly irreducible Heyting algebra except the 2 valued Boolean algebra $\mathbf{B}_2$ has a subalgebra isomorphic to $\mathbf{H}_3$. From these observations, it follows:

Theorem 3.5. *The logic* $\mathbf{L}(\mathbf{H}_3)$ *is shown to be the greatest among super-intuitionistic logics smaller than classical logic.*

From the above fact, it follows that the single subdirectly irreducible Boolean algebra is $\mathbf{B}_2$. This gives us a proof of the equivalence of validity in all Boolean algebras and tautolgy, in Theorem 2.2. For this it is enough to show that $\mathbf{L}(\mathbf{A}) = \mathbf{L}(\mathbf{B}_2)$ for any Boolean algebra $\mathbf{A}$ such that $0 \neq 1$. But this follows from Lemma 3.4 and the fact that every subdirectly irreducible Boolean algebra is isomorphic to $\mathbf{B}_2$.

Craig interpolation property

A logic $\mathbf{L}$ is said to have the *Craig interpolation property* (CIP) if and only if the following holds:

> for all formulas α and β, if $\alpha \to \beta$ is in $\mathbf{L}$ then there exisits a formula γ which contains only propositional variables appearing common in α and β such that both $\alpha \to \gamma$ and $\gamma \to \beta$ are in $\mathbf{L}$.

It can be shown (by a syntactical way, for example) that both classical logic and intuitionistic logic have the CIP. A question is whether having the CIP is a commonplace event or not. But when a given logic **L** does not have the CIP, how can we show this?

There is a nice result on the relation between a superintuitionistic logic **L** and a corresponding variety $\mathcal{V}_{\mathbf{L}}$ of Heyting algebras. That is, **L** has the CIP if and only if $\mathcal{V}_{\mathbf{L}}$ has a property, called the *amalgamation property*. By using this, L.L. Maksimova in 1977 showed the following surprising result, which says that a superintuitionistic logic rarely has the CIP.

Theorem 3.6. *Among uncountably many superintuitionistic logics, only 8 logics have the CIP.*

Halldén Completeness

A superintuitionistic logic **L** is *Halldén complete* (HC), if for all formulas α and β which have no variables in common, if $\alpha \vee \beta$ is in **L** then either α or β is in **L**. Obviously the disjunction property implies the Halldén completeness, while it is shown that the converse does not hold for uncountably many superintuitionistic logics.

A logic **L** is *meet irreducible* (in the lattice of all superintuitionistic logics) if it is not an intersection of two incomparable logics, or equivalently, if it is not a meet of strictly bigger logics. The following result is obtained by combining results by E.J. Lemmon and A. Wroński.

Theorem 3.7. *The following conditions are equivalent for every superintuitionistic logic* **L***.*

1. **L** *is Halldén complete,*
2. **L** = **L**(**A**) *for a subdirectly irreducible Heyting algebra* **A***,*
3. **L** *is meet irreducible.*

The importance of this theorem is that it states the equivalence of three different faces of logics. That is, Halldén completeness is a syntactic, local property of a logic, the second condition is on algebraic characterization of a logic, and the meet irreducibility is a global feature of a logic within the lattice of all superintuitionistic logics. In this way, algebraic approach will enable us to consider logics from a wide range of viewpoints.

3.5 Historical notes

Universal algebra is a general study of properties with which many algebraic structures share. The word “universal" means that it is not a study

of particular algebraic structures, like group theory, ring theory and lattice theory, but a study of properties common to them. In the 1930s, G. Birkhoff published papers in this direction, and later A. Tarski has developed it in the 1940s and 1950s. Around the same time, A.I. Mal'cev has made important contributions to the topic.

One can see a strong similarity between the viewpoint of universal algebra mentioned in the above and our attitude to a general study of nonclassical logics discussed in Section 2.4. This will explain the reason why tools of universal algebra play an essential role in the study of nonclassical logics.

4 Interlude – Rise of Kripke semantics

Till the beginning of the 1960s, semantics for nonclassical logics are mostly limited to algebraic ones, including matrices that consist of algebras with its designated subsets. But, after the introduction of relational semantics by S. Kripke called *Kripke semantics*, relational semantics became the main stream of semantics for nonclassical logics. They work particulary well for modal logics and superintuitionistic logics. An apparent merit of Kripke semantics lies in the facts that it is much more intuitively understandable and philosophically persuasive than algebraic semantics, and also that it is mathematically tractable since each Kripke frame consists of a quite simple structure, i.e., a set (of possible worlds) with binary relations, called accessibility relations. This became also a key for the driving force of a rapid development of study of extensions and applications of modal logic, like temporal logic and epistemic logic. While model theoretic approach based on Kripke semantics has made important and quite successful progresses in the latter half of the last century, algebraic approach had not drawn much attention for a while. As for history of study of modal logic, see [9], and for various approaches to modal logic, see e.g. [2].

We give a brief explanation on a connection between algebras and Kripke frames, considering their semantics for intuitionistic logic, and also superintuitionistic logics in general. As we mentioned before, algebraic models of these logics are given by Heyting algebras. On the other hand, Kripke frames for them are given by partially ordered sets. Suppose that a partially ordered set $\langle S, \leq \rangle$ is given. Take the set U_S of all upward-closed subsets of S. Then, the structure $\mathbf{U}_S = \langle U_S, \cap, \cup, \rightarrow, \emptyset, S \rangle$ forms a Heyting algebra, where $\cap$ and $\cup$ are set-theoretic intersection and union, respectively, and $X \rightarrow Y$ is defined by $\{u \in S$: for all $v \geq u, v \in X$ implies $v \in Y\}$ for all

$X, Y \in U_S$. Conversely, for a given Heyting algebra $\mathbf{A} = \langle A, \wedge, \vee, \rightarrow, 0, 1\rangle$, the set of all *prime filters* of $\mathbf{A}$ forms a partially ordered set $\langle P_A, \subseteq\rangle$. Here, a non-empty subset F of A is a *filter* if and only if F is upward-closed and $a \wedge b \in F$ whenever $a, b \in F$, and a filter F is *prime* if and only if for all $a, b \in A$, $a \vee b \in F$ implies either $a \in F$ or $b \in F$. Moreover, the mapping h from $\mathbf{A}$ to $\mathbf{U}_{P_A}$ defined by $h(a) = \{F \in P_A : a \in F\}$ is an embedding. The algebra $\mathbf{U}_{P_A}$ is called the *canonical extension* of $\mathbf{A}$. This argument is based on Stone-type representation theorem for Heyting algebras.

When $\mathbf{A}$ is finite, its canonical extension is isomorphic to $\mathbf{A}$. But, in general, it is not always the case that the canonical extension does not belong to the variety generated by the singleton set $\{\mathbf{A}\}$. In logical terms, it may happen that the logic $\mathbf{L}(\mathbf{U}_{P_A})$ is properly included by the logic $\mathbf{L}(\mathbf{A})$. For a given logic $\mathbf{L}$, if the corresponding variety $\mathcal{V}_{\mathbf{L}}$ is closed under taking canonical extensions, $\mathbf{L}$ is called *canonical*. Clearly if $\mathbf{L}$ is canonical then it is complete with respect to a class of Kripke frames (*Kripke-completeness*). It is known that there exist uncountably many Kripke-incomplete superintuitionistic logics. This shows a limitation of Kripke semantics.

Kripke semantics works quite well as long as a nice Stone-type representation theorem holds, which is usually followed from the distributive law. But Kripke semantics does not always work smoothly for substructural logics. This is one of the reasons why algebraic approaches to logics have revived and have attracted a great deal of attention again from the beginning of the 1990s, which is the moment when algebraic study of substractural logics has started. It is fair to say that algebraic semantics and Kripke semantics play complementary roles to each other.

5 Algebraization

We have mentioned a close relation between logic and algebra in Section 3. But this relationship can be extended to a higher level, the one between *deducibility* and *equational consequence*. This fact is called *algebraization* (in the sense of Blok-Pigozzi). The book [1] published in 1989 by Blok and Pigozzi became a starting point of the area known as *abstract algebraic logic*. Abstract algebraic logic is a study which focuses on relations between logic and algebra, in fact connections between deductive systems and classes of algebraic structures. It discusses in a general setting what is the algebraic counterpart of a given deductive system, how to get it and moreover how logical properties are related to algebraic properties in this connection. See [6] for the details. Abstract algebraic logic and universal

algebra compose the core of algebraic logic today. This section is a short introduction to abstract algebraic logic.

5.1 Deducibility and provability

We introduce a consequence relation for intuitionistic logic, which fits also for all superintuitionistic logics. Suppose that a formal system $\mathcal{S}$ for intuitionistic logic is given. It is irrelevant here whether $\mathcal{S}$ is a Hilbert-style system or a sequent system, or has some other formal system. For a set of formulas Σ and a formula ϕ, we write $\Sigma \vdash_{\mathbf{Int}} \phi$ whenever a formula ϕ is provable in the system which is obtained from $\mathcal{S}$ by adding all formulas σ, for $\sigma \in \Sigma$, as new axioms. We call this $\vdash_{\mathbf{Int}}$ the *deducibility relation* associated with intuitionistic logic. This notion of deducibility can be naturally extended to any superintuitionistic logic $\mathbf{L}$. In this case, we write the deducibility in $\mathbf{L}$ as $\vdash_{\mathbf{L}}$.

Can the deducibility relation be reduced to the provability? For intuitionistic logic and in fact for any superintuitionistic logic, the answer is yes, since the deducibility is finitary and the following *deduction theorem* holds.

Theorem 5.1. *For any set of formulas* Σ *and any formula* ϕ, $\Sigma \cup \{\alpha\} \vdash \beta$ *if and only if* $\Sigma \vdash (\alpha \to \beta)$.

Here is an outline of the proof (in case of intuitionistic logic). It is trivial that the right-hand side implies the left-hand one. Suppose that the left-hand side holds. We assume that the system $\mathcal{S}$ is formalized in a (standard) Hilbert-style system. Take a proof of β from $\Sigma \cup \{\alpha\}$, which consists of a sequence of formulas $\delta_1, \ldots, \delta_m$ such that (1) δ_m is equal to β and (2) for each $j \leq m$, δ_j is either an axiom of $\mathcal{S}$, or a member of $\Sigma \cup \{\alpha\}$, or a consequence of modus ponens of two formulas δ_i and δ_k with $i, k < j$ (that is, δ_k is of the form $\delta_i \to \delta_j$). Let γ_i be $\alpha \to \delta_i$. Then the sequence of formulas $\gamma_1, \ldots, \gamma_m$ (by inserting some necessary formulas between them) gives a proof of $\alpha \to \beta$ from Σ. This is assured by using the fact that all formulas of the form $\alpha \to \alpha$, $\delta \to (\alpha \to \delta)$ and $(\alpha \to (\delta \to \theta)) \to ((\alpha \to \delta) \to (\alpha \to \theta))$ are provable in intuitionistic logic.

On the other hand, as shown later, if one of these formulas is not provable then the above argument may collapse and therefore the deduction theorem may not hold in such a case.

5.2 Equational consequence

A consequence relation among equations can be introduced by using equational calculus in the following way. Let $\mathcal{K}$ be an arbitrary class of algebras. For any set of equations $\{u_i = v_i; i \in I\} \cup \{s = t\}$, the *equational consequence* $\models_{\mathcal{K}}$ of $\mathcal{K}$ is defined as follows: $\{u_i = v_i; i \in I\} \models_{\mathcal{K}} s = t$ if and only if

> for each algebra $\mathbf{A}$ in $\mathcal{K}$ and each valuation f on A, $f(s) = f(t)$ holds whenever $f(u_i) = f(v_i)$ holds for all $i \in I$.

In particular when I is a finite set $\{i : 1 \leq i \leq m\}$, $\{u_i = v_i; 1 \leq i \leq m\} \models_{\mathcal{K}} s = t$ becomes equivalent to the validity of the following *quasi-equation* in every $\mathbf{A}$ in of $\mathcal{K}$.

> $(u_1 = v_1$ and ... and $u_m = v_m)$ implies $s = t$.

5.3 Algebraization a la Blok-Pigozzi

The close relation mentioned in Section 3.3 between logics and varieties can be extended to the one between deducibility and equational consequence, as shown in the following.

Lemma 5.2. *Let $\mathcal{V}$ be any subvariety of the variety of Heyting algebras. For any set of equations E, $E \models_{\mathcal{V}} s = t$ if and only if $\{u \leftrightarrow v : u = v \in E\} \vdash_{\mathbf{L}_{\mathcal{V}}} s \leftrightarrow t$.*

This lemma gives us a way of translating equations into formulas. It turns out that the reverse translation is also possible.

Lemma 5.3. *Let $\mathbf{L}$ be any superintuitionistic logic. For any set Γ of formulas, $\Gamma \vdash_{\mathbf{L}} \delta$ if and only if $\{\gamma = 1 : \gamma \in \Gamma\} \models_{\mathcal{V}_{\mathbf{L}}} \delta = 1$.*

What is more, these translations are mutually inverse. Each given formula γ is translated into an equation $\gamma = 1$, which is translated back into a formula $\gamma \leftrightarrow 1$. Conversely, if we start from an equation $s = t$ then it is translated into a formula $s \leftrightarrow t$ which is translated into an equation $(s \leftrightarrow t) = 1$. But in either case, the application of two translations produces the formula and the equation, respectively, equivalent to the original ones, as the following shows.

Lemma 5.4. *For any formula γ and any equation $s = t$,*

$$\gamma \dashv\vdash_{\mathbf{L}} \gamma \leftrightarrow 1$$

$$s = t \rightleftharpoons \mathcal{V} \ (s \leftrightarrow t) = 1$$

where $\dashv\vdash_{\mathbf{L}}$ and $\rightleftharpoons \mathcal{V}$ mean that the respective relations hold both ways.

In terms of abstract algebraic logic, these three lemmas as a whole say the algebraizability of the deducibility $\vdash_{\mathbf{L}}$ (in the sense of Blok-Pigozzi).

> for each superintuitionistic logic $\mathbf{L}$, $\vdash_{\mathbf{L}}$ is *algebraizable* and $\mathcal{V}_{\mathbf{L}}$ is an *equivalent algebraic semantics* for it.

6 Substructural logics

We have discussed algebraic approaches to logics, by taking superintuitionistic logics as examples. As a matter of fact, they work well for a much wider class of logics, in particular for substructural logics. Substructural logics include superintuitionistic logics, relevant logics, linear logic and logics over Łulasiewicz's infinite-valued logic.

In this section, we will introduce substructural logics and show how well algebraic methods are applied to them.

6.1 What are substructural logics

Study of substructural logics can be regarded as an enterprise of understanding various nonclassical logics in an uniform framework. Here, by nonclassical logics, roughly we mean logics weaker than classical logic. Thus logics with additional connectives like modal connectives are excluded. As mentioned already, superintuitionistic logics, relevant logics, linear logic and logics over Łulasiewicz's infinite-valued logic will be included. Here, relevant logics are logics of relevant implication, in which neither of $(\alpha \wedge \neg\alpha) \rightarrow \beta$ and $\alpha \rightarrow (\beta \rightarrow \alpha)$ are rejected in general since there may be no relevant connections between formulas α and β. On the other hand, the implication in classical logic is material implication, i.e., the implication $\alpha \rightarrow \beta$ is regarded as $\neg\alpha \vee \beta$, and therefore both of them are true formulas. Our attempt is to find something common to these logics that are introduced and studied from different backgroud and motivation.

But in retrospect, the study has developed in a different way. Around the middle of the 1980s, some people independently discussed logics which

are formalized in Gentzen-type sequent systems, like commutative version of Lambek calculus for categorial grammar, linear logic and logics lacking the weakening rule. A common feature of them is that these sequent systems lack some *structural rules*, which standard sequent system **LK** for classical logic or **LJ** intuitionistic one have. Gradually it has been discovered that many of nonclassical logics fall under this class. For instance, relevant logics do not allow the axiom $\alpha \to (\beta \to \alpha)$, which corresponds to the (left) *weakening rule* in sequent systems, and Łukasiewicz's many-valued logics do not allow the axiom $(\alpha \to (\alpha \to \beta)) \to (\alpha \to \beta)$, which corresponds to the *contraction rule*. Thus, the word "substructural logics" is introduced as a generic term to denote logics that may lack some of structural rules when formulated in sequent systems.

6.2 Sequent systems and structural rules

To explain structural rules in more details, we will give a brief explanation on sequent systems. In sequent systems, *sequents* are basic syntactical objects. We consider here sequents of the following form, where commas and the arrow $\Rightarrow$ are meta-logical symbols:

$$\alpha_1, \ldots, \alpha_m \Rightarrow \beta. \tag{1}$$

Such sequents are used in the system **LJ** for intuitionistic logic. Intuitively the above sequent means that "*β follows from assumptions from α_1 to α_m*." Sequent systems consist of initial sequents and rules. Intitial sequents correspond to axioms in Hilbert-style formal systems, and usually sequents of the form $\alpha \Rightarrow \alpha$ are taken for initial sequents. Each rule in a sequent system describes an inference of a new sequent from one or two given sequent(s). Rules are divided into three classes: rules for each logical connectives, cut rule and structural rules. Within the class of structural rules, there are three types: exchange rule, contraction rule and left- and right-weakening rules.

(e) exchange rule : $\dfrac{\Gamma, \alpha, \beta, \Delta \Rightarrow \varphi}{\Gamma, \beta, \alpha, \Delta \Rightarrow \varphi}$

(c) contraction rule : $\dfrac{\Gamma, \alpha, \alpha, \Delta \Rightarrow \varphi}{\Gamma, \alpha, \Delta \Rightarrow \varphi}$

(i) left weakening rule : $\dfrac{\Gamma, \Delta \Rightarrow \varphi}{\Gamma, \alpha, \Delta \Rightarrow \varphi}$

(o) right weakening rule : $\dfrac{\Gamma \Rightarrow}{\Gamma \Rightarrow \alpha}$

Structural rules control roles of commas, as described below. The meaning of commas will be changed if our system lacks some of them. The exchange rule allows us to use assumptions in an arbitrary order. The contraction rule allows us to contract two assumptions of the same formula into one. Hence, in the absence of the contraction rule each assumption is consumed once it is used. The left-weakening rule allows us to add any formula as an assumption. Thus, in the absence of the left-weakening no redundant assumptions are admitted.

In the sequent system **LJ** for intuitionistic logic, it is shown that the sequent (1) is provable if and only if the following sequent is provable.

$$\alpha_1 \wedge \ldots \wedge \alpha_m \Rightarrow \beta \tag{2}$$

It means that in **LJ** each comma of sequents is an external expression of conjunction. But, in the proof of the above equivalence, we need to use both the contraction and the left-weakening. In other words, we cannot expect that commas are interpreted as conjunctions in weaker systems.

To supplement this weakness, we introduce a new logical connective *fusion* ($\cdot$, in symbol), which expresses each comma internally, by introducing rules for fusion. Though we omit the details, we can show without using any structural rule that the sequent (1) is provable if and only if the following sequent is provable.

$$\alpha_1 \cdot \ldots \cdot \alpha_m \Rightarrow \beta \tag{3}$$

6.3 Various logics

The basic sequent system for substructural logics is **FL**, *Full Lambek Calculus*, which is obtained from **LJ** by deleting all of structural rules and by adding rules for fusion. But in the following we take the sequent system $\mathbf{FL_e}$, obtained from **FL** by adding the exchange rule, and consider mostly substructural logics over $\mathbf{FL_e}$, in order to avoid some technical complications caused by the deletion of the exchange rule. (In fact, two kinds of implication are introduced in logics without the exchange rule.) The following basic result holds in $\mathbf{FL_e}$, which says that implication is the *residual* of fusion.

$$\alpha \cdot \beta \Rightarrow \varphi \text{ is provable if and only if } \alpha \Rightarrow \beta \rightarrow \varphi \text{ is provable}$$

Now we define *substructural logics* over **FL** (and $\mathbf{FL_e}$) to be axiomatic extensions of **FL** (and $\mathbf{FL_e}$, respectively), i.e., systems obtained from **FL**

($\mathbf{FL_e}$) by adding a set of axiom schemes as additional initial sequents. As shown below, many important classes of nonclassical logics are included in substructural logics.

1) Lambek calculus: J. Lambek introduced it in 1958, as a calculus for categorial grammer, which is roughly equal to the logic **FL** without structural rules.

2) Relevant logics: A common feature to various relevant logics is the absence of the weakening rules.

3) Logics without the contraction rule: The paper by Ono and Komori in 1985 discussed logics in this class comprehensively. Łukasiewicz's many-valued logics are among them.

4) Linear logic: The logic **MALL** is introduced by J.-Y. Girard in 1987, which is the extension of $\mathbf{FL_e}$ with the law of double negation.

5) Johansson's minimal logic: This is the logic without the rule of right-weakening. Thus it rejects the principle that every formula follows from a contradiction.

6.4 Substructural logics and residuated lattices

Algebraic models of substructural logics are given by *residuated lattices*. We consider here only algebraic models of substructural logics over $\mathbf{FL_e}$ that are given by *commutative* residuated lattices. An algebra $\mathbf{A} = \langle A, \wedge, \vee, \rightarrow, \cdot, 1\rangle$ is a commutative residuated lattice if and only if $\langle A, \wedge, \vee\rangle$ is a lattice, and $\langle A, \cdot, 1\rangle$ is a commutative monoid satisfying the following *law of residuation*;

$$a \cdot b \leq c \text{ if and only if } a \leq b \rightarrow c, \text{ for all } a, b, c \in A.$$

The operation $\cdot$ is introduced to interpret the fusion, and also the meta-logical symbol "commas" in sequents. An $\mathbf{FL_e}$-algebra is a commutative residuated lattice with a fixed element 0. Using 0, negation is defined by $\neg x = x \rightarrow 0$. Heyting algebras are exactly $\mathbf{FL_e}$-algebras in which $a \cdot b = a \wedge b$ for all a, b and 0 is the least element. In models of Łukasiewicz's many-valued logics with the universe A, which is a subset of the interval $[0, 1]$, if we put $a \cdot b = \max\{0, a + b - 1\}$ then $\langle A, \cdot, 1\rangle$ forms a commutative monoid which satisfies that $a \cdot b \leq c$ if and only if $a \leq b \rightarrow c$, where $b \rightarrow c$ is defined by $\min\{1, 1 - b + c\}$. Thus, they are also $\mathbf{FL_e}$-algebras.

The class of $\mathbf{FL_e}$-algebras is shown to be defined by a set of equations, including the equation $x \cdot y = y \cdot x$ which expresses the commutativity. Similarly, other structural rules, i.e., the contraction, the left-

and right-weakening rules, are expressed by equations $x \leq x \cdot x$ (square-increasingness), $x \leq 1$ (integrality), and $0 \leq x$ (minimality of 0), respectively. (Note that each inequality $\leq$ is expressed as an equality, since we have lattice operations in the language.)

As a consequence of Theorem 3.2, the class $\mathcal{FL}_e$ of all $\mathbf{FL_e}$-algebras is a variety, and clearly the classes of all Boolean algebras and of all Heyting algebras are subvarieties. Similar correspondence to the one mentioned in Section 3.3 holds between the lattice of all substructural logics over $\mathbf{FL_e}$ and the lattice of all subvarieties of $\mathcal{FL}_e$. In this connection, substructural logics with the contraction rule (the weakening rules) correspond varieties of $\mathbf{FL_e}$-algebras satisfying square-increasingness (integrality and minimality of 0, respectively). In an algebra $\mathbf{A}$ without the integrality, the element 1 may not be the greatest element. In such a case, we modify the definition of the validity in such a way that a formula φ is valid in $\mathbf{A}$ if and only if $f(\varphi) \geq 1$ for each valuation f, or equivalently $f(\varphi) \wedge 1 = 1$ for each f.

Algebraization result discussed in Section 5.3 can be extended to the present case, but by the same reason as above, it is necessary to translate each formula γ into the equation $\gamma \wedge 1 = 1$. Then, we have:

> for each substructural logic $\mathbf{L}$ over $\mathbf{FL_e}$, in fact for each substructural logic $\mathbf{L}$ over $\mathbf{FL}$, $\vdash_{\mathbf{L}}$ is algebraizable.

6.5 Deducibility revisited

Deduction theorem (Theorem 5.1) says that deducibility in intuitionistic logic is reducible to provability in it. As the outline of the proof shows, the theorem depends on the existence of structural rules, and the proof cannot be applied to the case of $\mathbf{FL_e}$. But the following weak form of deduction theorem, called *local deduction theorem* holds for $\mathbf{FL_e}$, and hence it holds for all substructural logics over $\mathbf{FL_e}$.

Theorem 6.1. *For any set of formulas* Σ *and any formula* ϕ, $\Sigma \cup \{\alpha\} \vdash_{\mathbf{FL_e}} \beta$ *if and only if* $\Sigma \vdash_{\mathbf{FL_e}} ((\alpha \wedge 1)^m \to \beta)$ *for some* m.

Here, δ^m means the formula $\delta \cdot \ldots \cdot \delta$ with δ for m times. Deduction theorem of intuitionistic logic is obtained from this by applying the contraction and the weakening rules. The above deduction theorem is said to be *local*, since the number m in it cannot be determined in general, when Σ, α, β are given.

The deducibility relation for a logic $\mathbf{L}$ is *decidable*, if there is an effective procedure of deciding whether or not $\Sigma \vdash_{\mathbf{L}} \alpha$ holds for each finite set of

formulas Σ and each formula α. The decidability of provability for a logic $\mathbf{L}$ is defined by restricting to the case when Σ is empty. We can show the following. The first is shown by a syntactic way (cut elimination of the sequent system $\mathbf{FL_e}$) and the second was proved essentially by Lincoln, Mitchell, Scedrov and Shankar in the 1990s.

Theorem 6.2. *1. The provability problem of* $\mathbf{FL_e}$ *is decidable.*
2. The deducibility problem of $\mathbf{FL_e}$ *is undecidable.*

By using algebraization result and by translating logical relations into equational ones, we have the following algebraic results as an immediate consequence of Theorem 6.2.

Theorem 6.3. *1. The equational theory of* $\mathbf{FL_e}$*-algebras is decidable.*
2. The quasi-equational theory of $\mathbf{FL_e}$*-algebras is undecidable.*

6.6 Further notes

Study of substructural logics has opened a new branch in algebraic logic. It will take a middle position between abstract algebraic logic and study of individual nonclassical logics. The class of substructural logics is wide enough to cover many nonclassical logics, and at the same time the study has successfully offered interesting concrete examples to abstract algebraic logic.

Residuation is a key notion in both substructural logics and residuated lattices, by which the necessity of sequent formulation of substructural logics can be clarified. Residuated lattices are not only algebraic models of substructural logics, but also form an important class in algebra. For example, lattice-ordered groups are among them. Close relations between logic and algebra, and even between proof-theoretic notions and algebraic ones have been discovered in algebraic logic. For further information on substructural logics, see [7].

References

[1] Blok, W.J. and D. Pigozzi, *Algebraizable Logics*, Memoirs of the AMS, Americal Mathematical Society, 1989.

[2] Bull, R. and K. Segerberg, *Basic modal logic*, Handbook of Philosophical Logic 3, 2nd edition, Kluwer Academic Publishers, 2001, pp.1-81.

[3] Burris, S. and H.P. Sankappanavar, *A Course in Universal Algebra*, Graduate Texts in Mathematics, Springer, 1981, available on line.

[4] van Dalen, D., *Intuitionistic logic*, Handbook of Philosophical Logic 5, 2nd edition, Kluwer Academic Publishers, 2002, pp.1-114.

[5] Davey, B.A. and H.A. Priestley, *Introduction to Lattices and Order* 2nd edition, Cambridge University Press, Cambridge, 2002.

[6] Font, J.M., R. Jansana and D. Pigozzi, *A survey of abstract algebraic logic*, Studia Logica 74 (2003) pp.13-97.

[7] Galatos, N., P. Jipsen, T. Kowalski and H. Ono, *Residuated Lattices: an Algebraic Glimpse at Substructural Logics*, Studies in Logic and the Foundations of Mathematics 151, Elsevier, 2007.

[8] Goble L. ed., *The Blackwell Guide to Philosophical Logic*, Blackwell Publishing, 2001.

[9] Goldblatt, R., *Mathematical modal logic: a view of its evolution*, Handbook of the History of Logic 7, Elsevier, 2006, pp.1-98.

[10] Halmos, P. and S. Givant, *Logic as Algebra*, Dolciani Mathematical Expositions 21, The Mathematical Association of America, 1998.

[11] Rasiowa, H., *An Algebraic Introduction to Non-Classical Logics*, Studies in Logic and the Foundations of Mathematics 78, North-Holland, PWN, 1974.

[12] Rasiowa, H. and R. Sikorski, *The Mathematics of Metamathematics*, PWN-Polish Science Publishers, Warszawa, 1963.

PART III

Logics of Processes and Computation

Temporal and Dynamic Logic

FRANK WOLTER AND MICHAEL WOOLDRIDGE*

Abstract

We present an introductory survey of temporal and dynamic logics: logics for reasoning about how environments change over time, and how processes change their environments. We begin by introducing the historical development of temporal and dynamic logic, starting with the seminal work of Prior. This leads to a discussion of the use of temporal and dynamic logic in computer science. We describe three key formalisms used in computer science for reasoning about programs (LTL, CTL, and PDL), and illustrate how these formalisms may be used in the formal specification and verification of computer systems. We then discuss interval temporal logics. We conclude with some pointers for further reading.

1 Introduction

Mathematical logic was originally developed with the goal of formalising mathematical reasoning – to formalise notions such as truth and proof. One important property of mathematical expressions such as theorems and their proofs is that they are inherently *timeless*: a result such as Fermat's Theorem is true now, always has been true, and always will be true – irrespective of when it was actually proved. In this sense, mathematical logic was conceived with the goal of developing formal languages for representing a fixed, non-changing world, and the semantics of classical logic reflect this assumption. In the semantics of classical logic, it is assumed that there is exactly one world (called a model), which satisfies or refutes any given sentence. But this limits the applicability of such logics for reasoning about *dynamic* domains of discourse, where the truth status of statements can change over time.

It is of course possible to reason about time-varying domains using classical first-order logic. One obvious approach is to use a two-sorted language, in which we have one sort for the domain of discourse, and a second

*Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, UK

sort for points in time. Variables $t, t', \ldots$ are used to denote time points, and a binary "earlier than" relation, "$<$" is used to capture the temporal ordering of statements. Using this approach, for example, the English sentence "It is never hot in Liverpool" might be translated into the following first-order formula:

$$\forall t. \neg Hot(Liverpool, t)$$

This approach is sometimes called the method of temporal arguments [35], or simply the first-order approach [20]. The advantage of the approach is that no extra logical apparatus must be introduced to deal with time: the entire machinery of standard first-order logic can be brought to bear directly. The obvious disadvantages are that the approach is unnatural and awkward for humans to use. Formulae representing quite trivial temporal properties become large, complicated, and hard to understand. For example, when translated to first-order logic the English sentence "we are not friends until you apologise" becomes something like the following:

$$\begin{aligned} &\exists t. [(now < t) \wedge Apologise(you, t)] \wedge \\ &\quad \forall t'. [(now < t' < t) \rightarrow \neg Friends(us, t')]. \end{aligned}$$

The desire for logics that are capable of naturally and transparently capturing the meaning of statements such as those above, in dynamic environments, led to the development of specialised *temporal* and *dynamic* logics. This article is intended as a high-level survey of such logics. Contemporary research in temporal and dynamic logics is a huge and very active enterprise, with participation from disciplines ranging from philosophy and linguistics to computer science [24]. Within the latter, it is especially the application of temporal and dynamic logics to verifying the correctness of computer systems that had a huge impact on the field. In an amazing example of technology transfer, this application has transformed purely philosophical logics into industrial-strength software analysis tools [42].

Before taking a look at such recent developments, however, we reflect on the origins of temporal logic, and in particular, the contributions of Arthur Prior.

2 The Origins of Temporal Logic

As we noted in the introduction, in classical logic it is implicitly assumed that formulae are interpreted with respect to a single model. But this inherently static view of logic and its subject matter makes it awkward to apply classical logic to the analysis of everyday sentences such as "Barack

Obama will win the election", since this statement might be true if evaluated now, but false if evaluated next week. It was concerns like this that led Arthur Prior, a philosopher and logician born in New Zealand in 1914, to start working on logics intended to facilitate reasoning about such statements. In Prior's words [34]: "Certainly there are unchanging truths, but there are changing truths also, and it is a pity if logic ignores these, and leaves it ... to comparatively informal dialecticians to study the more dynamic aspects of reality." Prior's work, mainly carried out in the 1950s and 1960s, is regarded as the foundation of the area now called temporal and dynamic logic [32, 33].

To analyse sentences such as "Barack Obama will win the election", Prior proposed the idea of regarding tense as a species of *modality*. He took classical propositional logic with its connectives $\vee$ (for "or"), $\neg$ ("not"), and $\rightarrow$ ("if ..., then ..."') and extended it with modal *tense operators*, F and P. In Prior's notation, Fp stands for "it will be the case that p" and Pp stands for "it was the case that p". In a similar way as in classical logic, one can define other basic tense operators as composed formulae. For example, the expression Gp (read "generally, p") is defined as $\neg F \neg p$ (meaning "p will always be the case") and Hp ("heretofore, p") is defined as $\neg P \neg p$ (meaning "p has always been the case"). Other temporal operators are also sometimes used, such as $\neg F \neg F p$ ("p will be the case again and again").

Given this set-up, in addition to classical tautologies, (sentences like $p \rightarrow p$, which are true independently of the model under consideration), one should also consider temporal tautologies: formulae using tense operators that are true independently from the truth of its propositional atoms. Priors first axiomatization of temporal tautologies included formulae such as

$$FFp \rightarrow Fp, \quad \text{and} \quad F(p \vee q) \rightarrow (Fp \vee Fq).$$

The first formula, for example, states "if it will be the case that it will be the case that p, then it will be the case that p". A less obvious candidate for a temporal tautology is its "converse" $Fp \rightarrow FFp$. A moment's reflection should convince the reader that the truth of this formula depends on precisely how Fp is interpreted. In other words, to decide whether this formula is a temporal tautology one has to define a formal semantics for the temporal language. Of course, one cannot interpret Prior's formulae in the static, non-changing models of classical logic. Instead one has to develop models that capture the evolution of reality over time. Leaving aside the question of how a physicist might answer this question, one obvious and mathematically simple approach is to interpret time as a linear sequence of time points: t_0, t_1, t_2, ... These time points can be naturally interpreted

as, say, dates in a calendar. Mathematically, however, we can view the flow of time as the natural numbers $\mathbb{N}$, ordered by the usual "less than" relation, "$<$". Many other models of time are also possible, with correspondingly different properties.

Now, in contrast to classical logic, a formula can be true at one time point, and false at another time point. Thus, we obtain a time-dependent notion of truth: a formula might be true when evaluated at t_i and false when evaluated at t_{i+1}. Formally, each time point t_i comes with a truth assignment stating which propositional atoms p are true at t_i. The propositional connectives are interpreted as in classical logic (for example $\phi \wedge \psi$ is true at t_i if, and only if, ϕ and ψ are both true at t_i). Finally,

Fp is true at t_i if, and only if, there exists $j > i$ such that p is true at t_j;

Pp is true at t_i if, and only if, there exists $j < i$ such that p is true at t_j.

Assuming, for example, that p is true at t_{100} and no other time point, then Fp is true at t_0. In fact, it is true at all time points between (and including) t_0 and t_{99}, but not at t_{100} nor any time point after t_{100}. Let us check that $FFp \rightarrow Fp$ is true in every time point no matter at which points p is true. To this end, assume that FFp is true at, say, t_n. Then Fp is true at some time point after t_n, say t_m. Similarly, this means that p is true at some time point after t_m, say t_k. But then t_k is a time point after t_n and we obtain that Fp is true at t_n. We have shown that Fp is true at any given time point if FFp is true at that time point. Thus, $FFp \rightarrow Fp$ is true at every time point, independently from p. Note that what we have applied here is the natural assumption that temporal precedence is transitive: if t_k is after t_m and t_m is after t_n, then t_k is after t_n. In contrast, the truth of $Fp \rightarrow FFp$ in this model of time depends on p. For example, assume that p is true at t_5 and no other time point. Then Fp is true at t_4, but FFp is not true at t_4: it is not possible to find a time point after t_4 that is before t_5. In fact, it is not difficult to see that a time model $(T, <)$ with set of time points T and temporal precedence relation $<$ validates $Fp \rightarrow FFp$ if, and only if, it is *dense*: for any two time points $t_1 < t_2$ there is a time point t' such that $t_1 < t' < t_2$. Thus, if we move from the *discrete* time model $t_0, t_1, \ldots$ to a *dense* model of time that resembles the rational or real numbers, we obtain new temporal tautologies (and loose others).

This small example illustrates one of the main distinctions between classical and temporal logics: even if the language is fixed and as simple as the basic tense logic with operators F and P, there remains a number of choices to be made with respect to the model used to represent time. Depending

on the temporal domain and discourse of interest, one can choose between flows of time that are discrete, dense, or continuous; time can be cyclic or cycle-free; and time can be endless or start with a "big bang". The are many possibilities, and in the late 1960s and early 1970s it became something of an industry to axiomatize and analyse the temporal tautologies of such time models.

In addition to varying the flow of time, also more tense operators were introduced and investigated. Of particular interest are the binary operators S (for *since*) and U (for *until*) whose semantics is defined as follows:

> pUq is true in t if, and only if, there exists $t' > t$ such that q is true in t' and p is true in t'' for all t'' such that $t < t'' < t'$;
>
> $pS\,q$ is true in t if, and only if, there exists $t' < t$ such that q is true in t' and p is true in t'' for all t'' such that $t' < t'' < t$.

Note that Fp and Pp can be expressed using since and until as

$$Fp = \top Up \quad \text{and} \quad Pp = \top S\,p,$$

where $\top$ is a propositional constant standing for a propositional tautology. For axiomatizations of temporal tautologies for languages with operators F, P, S, and U for various time flows see [8].

3 Temporal *versus* Predicate Logic

In our introduction to Prior's basic tense logic, we emphasised that the main difference between classical and temporal logic is time-dependence: in classical logic truth is time-independent, whereas in temporal logic it is not. Under this view, temporal languages are extensions of propositional logic by means of temporal operators. There is, however, a very different and equally important interpretation of temporal logic, namely as a fragment of predicate logic (see [4] for a general discussion of these two different views in modal logic). To achieve this, propositional atoms are identified with unary predicates and complex temporal formulae become predicate logic sentences with exactly one free variable x that ranges over time points. Consider, for example, the sentence "the mail will be delivered". In Prior's tense logic, this is formalised as Fp, where p stands for "the mail is delivered". In contrast, in predicate logic one introduces a unary predicate P ranging over time points for "the mail is delivered" and

the sentence is formalised as $\exists y(x < y \wedge P(y))$, where $<$ stands for temporal precedence.

To make the connection between temporal and predicate logic precise, consider a flow of time $(T, <)$ which can be, for example, the discrete time flow $t_0, t_1, \ldots$ or a copy of the rational or real numbers. A valuation v determines at which time points $t \in T$ an atom p from a given set Φ of propositional atoms is true. Equivalently, we can describe the resulting model by identifying $(T, <, v)$ with the first-order relational model

$$M = (T, <, (p^M \mid p \in \Phi)),$$

where $p^M \subseteq T$ denotes the set of time points at which p is true. Thus, now we regard the $p \in \Phi$ as unary predicates that have as extensions a set of time points. Inductively, we can translate every temporal logic formula ϕ in the language with, say, S and U, as a first-order predicate logic formula $p^\sharp(x)$, where x is a fixed individual variable:

$$\begin{aligned}
p^\sharp &= p(x) \\
(\phi \wedge \psi)^\sharp &= \phi^\sharp(x) \wedge \psi^\sharp(x) \\
(\neg\phi)^\sharp &= \neg\phi^\sharp(x) \\
(\phi S \psi)^\sharp &= \exists y(y < x \wedge \psi^\sharp(y) \wedge \forall z((y < z < x) \rightarrow \phi^\sharp(z))) \\
(\phi U \psi)^\sharp &= \exists y(y > x \wedge \psi^\sharp(y) \wedge \forall z((x < z < y) \rightarrow \phi^\sharp(z)))
\end{aligned}$$

where y, z are fresh individual variables and $\phi^\sharp(y)$ and $\phi^\sharp(z)$ are obtained from $\phi^\sharp(x)$ by replacing x with y and z, respectively (see [24] for details).

By definition of $\cdot^\sharp$, we have that a temporal logic formula ϕ is true at a time point t in a model M if, and only if $M \models \phi^\sharp[t]$, where $\models$ is the standard truth-relation of first-order predicate logic. Thus, modulo the translation $\cdot^\sharp$, we can regard the temporal language with operators S and U as a fragment of first-order predicate logic. The formulae obtained as translations of temporal formulae are, of course, only a tiny subset of the set of all first-order predicate logic formulae (even those with one free variable x using only unary predicates $p \in \Phi$ and the binary predicate $<$). Moreover, for arbitrary time flows, there are many first-order predicate formulae of this form that are *not equivalent* to any translation of a temporal formula. However, a fundamental result proved by Hans Kamp [25] established that, for some important flows of time, every first-order predicate logic formula in the language with unary predicates $p \in \Phi$ and the binary predicate $<$ and having exactly one free variable is indeed equivalent to a temporal formula using since and until. The precise formulation is as follows:

Theorem 1 (Kamp). *Let $(T, <)$ be a flow of time consisting of the natural or real numbers. Then one can construct for every first-order formula $\varphi(x)$ using $<$ and $p \in \Phi$ and with one free variable x, a temporal formula φ^T using the operators S and U such that the following holds for every model $M = (T, <, (p^M \mid p \in \Phi))$ and every $t \in T$:*

$$\varphi^T \textit{ is true at } t \quad \Leftrightarrow \quad M \models \varphi[t]$$

Kamp's theorem explains why there are only very few important distinct temporal operators for linear time: any first-order definable temporal operator can be expressed using just since and until. We would like to stress here that it would be wrong to conclude from Kamp's result that temporal logic has nothing useful to offer compared to predicate logic. As we pointed out in the introduction, the crucial difference between temporal and predicate logic is that temporal logic is much closer to natural language than predicate logic and, therefore, much easier for people to read and understand. Thus, for the same reason that programming languages such as C or JAVA are not useless just because any program in C or JAVA is equivalent to a Turing Machine, temporal logics do not become useless just because they have the same expressive power as first-order formulae.

Interestingly, the difference between temporal and first-order logic can also be described in technical terms. The translation from first-order predicate logic to temporal logic introduces temporal formulae of non-elementary size (i.e., their size cannot be bounded by a tower of exponentials) [18] and for standard linear time flows the satisfiability problem for temporal logic with S and U is PSPACE-complete (and even coNP-complete with operators F and P only), but it is non-elementary for the corresponding fragment of first-order predicate logic [18, 24, 27].

Kamp's result was the beginning of a long and ongoing research tradition. Results such as Kamp's are nowadays known as *expressive completeness* results. A typical expressive completeness result states that a certain temporal language is equivalent to (some fragment of) first-order predicate logic over a certain class of time flows. For example, Prior's original language with the tense operators F and P only and without since and until is expressively complete for the *two-variable fragment* of first-order predicate logic (i.e., first-order predicate sentences using only two individual variables) on arbitrary linear time flows [12, 28]. An overview of recent extensions and variations of Kamp's theorem is given in [24].

4 Temporal Logic in Computer Science

At this point in our story, computer science enters the scene. A key research topic in computer science is the *correctness problem*: crudely, the problem of showing that computer programs operate correctly [6]. Two key issues associated with correctness are the related problems of *specification* and *verification*. A specification is an exact description of the behaviour that we want a particular computer system to exhibit. Verification is the problem of demonstrating that a particular program does or does not behave as a particular specification says it should.

Temporal logic has proved to be an extremely valuable formalism for the specification and verification of computer systems. This application of temporal logic is largely due to the work of Amir Pnueli, an Israeli logician born in 1941. In 1977, Pnueli was considering the problem of specifying a class of computer programs known as *reactive systems*. A reactive system is one that does not simply compute some function and terminate, but rather has to maintain an *ongoing interaction* with its environment. Examples of reactive systems include computer operating systems and process control systems. Typical properties that we might find in the specification of a reactive system are *liveness* and *safety* properties. Intuitively, liveness properties relate to programs correctly progressing, while safety properties relate to programs avoiding undesirable situations. In a seminal paper [31], Pnueli observed that temporal logics of the type introduced by Prior provide an elegant and natural formal framework with which to specify and verify liveness and safety of reactive systems. For example, suppose p is a predicate describing a particular undesirable property of a program (a "system crash", for example). Then the temporal formula $G\neg p$ formally expresses the requirement that the program does not crash – this is an example of a safety property.

Thus, Pnueli's idea was to formally specify the desirable behaviour of a reactive computer system as a formula Ψ of temporal logic. Such a *formal* specification is valuable in its own right, as a precise, mathematical description of the intended behaviour of the program. But it also opens up the possibility of formal verification, as follows. Suppose we are given a computer program $\mathcal{P}$, (written in a programming language such as PASCAL, C, or JAVA), and a specification Ψ, expressed as a formula of temporal logic. Then the idea of *deductive temporal verification* is to first derive the *temporal theory* $Th(\mathcal{P})$ of $\mathcal{P}$, i.e., a logical theory which expresses the actual behaviour of $\mathcal{P}$. The temporal theory $Th(\mathcal{P})$ of the program $\mathcal{P}$ is derived from the text of the program $\mathcal{P}$. For example, for each program statement

in $\mathcal{P}$, there will typically be a collection of axioms in the temporal theory $Th(\mathcal{P})$, which collectively characterise the effect of that statement. To do verification, we attempt to show that $Th(\mathcal{P}) \vdash \Psi$. If we succeed, then we say that $\mathcal{P}$ *satisfies* the specification Ψ; it is *correct* with respect to Ψ. In this way, verification reduces to a proof problem in temporal logic.

Pnueli's insight led to enormous interest in the use of temporal logic

in the formal specification and verification of reactive systems, and ultimately, to software verification tools that are used in industry today. One question that attracted considerable interest in the 1980s was that of exactly what kind of temporal logic is best suited to program specification and verification. Although one can use Prior's logics for reasoning about programs, they are not well suited to talking about the "fine structure" of the state sequences generated by programs as they execute. For this reason, a great many different proposals were made with respect to temporal logics for reasoning about programs (see, e.g., [26, 3, 37, 44, 1, 11]). In the end, two key formalisms emerged from this debate: *Linear Temporal Logic* (LTL) and *Computation Tree Logic* (CTL). These two formalisms are intended to capture different aspects of computation. LTL describes properties of a single run of a reactive program. Hence it is interpreted over a *linear* sequence of successive machine states. In contrast, CTL describes properties of the *branching* structure of the set of all possible runs of the program.

4.1 From Programs to Flows of Time

Before presenting the semantics of LTL and CTL, let us pause to consider in a little more detail exactly how computer programs give rise to flows of time. Consider the following (admittedly rather pointless) computer program, written in a PASCAL/C-like language.

```
x = y = true;
while (true) do
  if x == false then
    x = true;
  else
    x = false;
  end-if;
end-while;
```

Thus, this program manipulates two Boolean-valued program variables, `x` and `y`; the variable `x` is initialised to the value `true`, and then its value is

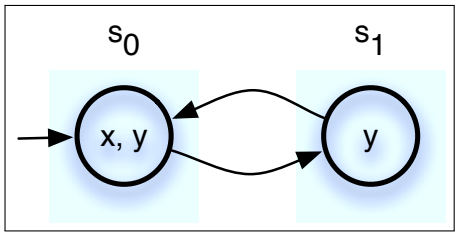


Figure 1: A state transition graph.

subsequently flipped between the values `true` and `false`. The variable `y` is initialised to `true` and remains unchanged subsequently. Notice that the program never terminates – it is an infinite loop. Now, we can understand the behaviour of this program as a *state transition system*, as illustrated in Figure 1. The state transition graph contains the possible states, or configurations, of the program; edges between states correspond to the execution of individual program instructions. For the program above, there are just two possible system states, labelled s_0 and s_1. The variables `x` and `y` are both true in s_0, while `x` is false and `y` is true in s_1; the arrow to state s_0 indicates that this is the *initial* state of the system. The other edges in the graph indicate that, when the program is in state s_0, then the only possible next state of the system is s_1, while when the program is in state s_1, the only possible next state of the system is s_0. Now, if we want to reason about the program given above, then we can focus on the state transition graph: this graph completely captures the behaviour of the program.

A little more formally, a state transition system is a triple:

$$M = (S, R, V)$$

where:

- S is a non-empty set of *states*;
- $R \subseteq S \times S$ is a total[1] binary relation on S, which we refer to as the *transition relation*; and
- $V : S \to 2^{\Phi}$ labels each state with the set of propositional atoms true in that state.

State transition systems are fundamental to the use of temporal logics for reasoning about programs. To make the link to temporal logic, we need

[1] By totality we here mean the property that every state has a successor, i.e., for every $s \in S$ there is a $t \in S$ such that $(s, t) \in R$.

a little more notation and terminology. A *path*, ρ, over M is an infinite sequence of states $\rho = s_0, s_1, \ldots$ which must satisfy that $(s_n, s_{n+1}) \in R$ for all natural numbers n. In the program given above, the state transition system is completely deterministic, in the sense that there is only ever one possible next state of the system, and so there is in fact only one path possible through the transition system of the program:

$$\rho : s_0, s_1, s_0, s_1, \ldots$$

However, in general, state transition systems are *non-deterministic*, in the sense that, for any given state s_i in the state transition graph, there can be multiple outgoing edges from s_i. This non-determinism can be thought of as reflecting the choices available to the program itself, or as the program's environment (e.g., its user) interacting with the program. So, in general, there may be more than one possible path through a transition system. The set of all possible paths through a state transition system will completely characterise the behaviour of the program: the paths in a transition system are exactly the possible *runs* of the program. Figure 2 shows how a transition system (Figure 2(a)) can be "unravelled" into a set of paths (Figure 2(b)). Of course, we do not show all the paths of the transition system in Figure 2 – the reader should be able to easily convince themselves that there are an infinite number of such paths, and these paths are infinitely long – even though the transition system that generates them is finite. Now, going back to temporal logic, a path is simply a linear, discrete sequence of time points (now called states), and we can think of such a path as a flow of time, in exactly the way that we discussed earlier. Thus temporal formulae of the kind studied by Prior's logics can be used to express properties of the runs of programs.

So far, we have three different views of programs: (*i*) the original program text, written in a programming language like PASCAL or C, above; (*ii*) the state transition diagram of the program, as in Figure 1 and Figure 2(a); and (*iii*) the runs obtained from the state transition diagram by "unravelling" it (Figure 2(b)). However, a third view is also possible: we can unravel the transition system into a *computation tree*, as shown in Figure 2(c). The key difference between the logics LTL and CTL is that the language of LTL is intended for representing properties of individual computation paths, while the language of CTL is intended for representing properties of computation trees of the type shown in Figure 2(c).

In the following subsections, we will take a closer look at the technical frameworks of LTL and CTL.

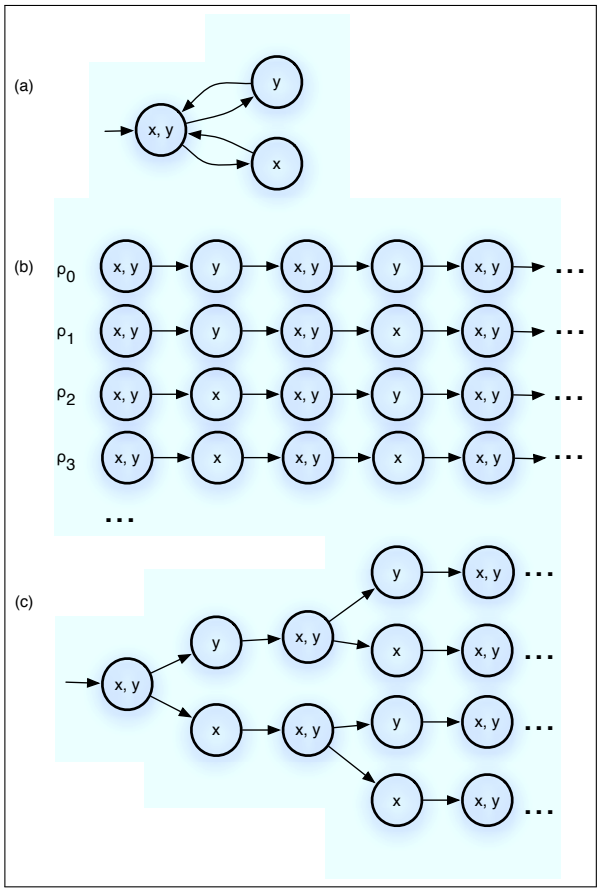


Figure 2: A state transition diagram (a), can be "unravelled" into a set of runs (b), or viewed as a tree-like branching model of time (c).

4.2 Linear Temporal Logic – LTL

In this section, we will present and investigate the framework of LTL in a little more detail. In the particular version that we work with, we will only consider temporal operators that refer to the future; it is also possible to consider LTL operators that refer to the past, although we will not do so here [9]. LTL extends classical propositional logic with the unary modal operator X ("next") and the binary operator U ("until"). Formally, starting with a set Φ of propositional atoms, the syntax of LTL is defined by the following grammar:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi \vee \phi \mid X\phi \mid \phi U \phi.$$

where $p \in \Phi$. A model for LTL is a path ρ in a transition system $M = (S, R, V)$. If $u \in \mathbb{N}$, then we denote by $\rho[u]$ the element indexed by u in ρ (thus $\rho[0]$ denotes the first element, $\rho[1]$ the second, and so on). The satisfaction relation "$\rho, u \models \phi$" between pairs ρ, u and LTL formulae ϕ formalises the condition "after u steps of the computation given by ρ, the formula ϕ holds" and is inductively defined via the following rules:

$\rho, u \models \top$

$\rho, u \models p$ iff $p \in V(\rho[u])$ (where $p \in \Phi$)

$\rho, u \models \neg\phi$ iff $\rho, u \not\models \phi$

$\rho, u \models \phi \vee \psi$ iff $\rho, u \models \phi$ or $\rho, u \models \psi$

$\rho, u \models X\phi$ iff $\rho, u + 1 \models \phi$

$\rho, u \models \phi U \psi$ iff there exists $v \geq u$ such that $\rho, v \models \psi$ and for all w such that $u \leq w < v$, we have $\rho, w \models \psi$.

(It is worth mentioning that the semantics of LTL can equivalently be defined using a flow of time $t_1, t_2, \ldots$ and without introducing an underlying transition system M. This is the viewpoint taken in our discussion of Prior's temporal logics. The main purpose of introducing state transition systems is to make explicit the computational interpretation of LTL and to enable the comparison with CTL in the next section.)

The reader will have noticed that the temporal operator U has been given a slightly different interpretation here: previously we had the "strict" interpretation of $\phi U \psi$ according to which $\phi U \psi$ is true at t if ψ is true some time t' later than t and ϕ is true at all time points properly between t and t'. This interpretation is typically seen in philosophical temporal logic motivated by capturing the semantics of natural language tense constructs. In contrast, in typical computer science temporal logic t' can be t itself or later. The same applies to the definition of F and G in terms of U. As before we define the temporal connectives F and G by setting

$$F\phi = \top U \phi \qquad G\phi = \neg F \neg \phi.$$

As the truth condition of U has changed, so have the truth conditions of $F\phi$ and $G\phi$. $G\phi$ means "either now or at some time later ϕ" and $G\phi$ means "now and always in the future ϕ". In what follows these distinctions will not play any important role.

Referring back to Figure 2(a), consider the path ρ_0. The following temporal properties may be seen to hold:

- $\rho_0, 0 \models x \wedge y$
- $\rho_0, 0 \models X(y \wedge \neg x)$
- $\rho_0, 0 \models XXX(y \wedge \neg x)$
- $\rho_0, 1 \models y \wedge \neg x$
- $\rho_0, 0 \models F(y \wedge \neg x)$
- $\rho_0, 0 \models GF(x \wedge y)$

However, considering path ρ_1, we have for example:

- $\rho_1, 0 \models X(y \wedge \neg x)$
- $\rho_1, 0 \models XXX(x \wedge \neg y)$

At this point, let us take a look at the types of properties that LTL may be used to specify. As we noted above, it is generally accepted that such properties fall into two categories: *safety* and *liveness* properties[2]. Informally, a safety property can be interpreted as saying that "something bad won't happen". For obvious reasons, safety properties are sometimes called invariance properties. The simplest kind of safety property is *global invariant*, expressed by a formula of the form: $G\phi$. A *mutual exclusion* property is a global invariant of the form: $G\left(\sum_{i=1}^{n} \phi_i \leq 1\right)$. This formula states that at most one of the properties $\phi_i \in \{\phi_1, \ldots, \phi_n\}$ should hold at any one time. (The Σ notation is readily understood if one thinks of truth being valued at 1, falsity at 0.) A *local invariant*, stating that whenever ϕ holds, ψ must hold also, is given by the following formula: $G(\phi \rightarrow \psi)$. Where a system terminates, *partial correctness* may be specified in terms of a precondition ϕ, which must hold initially, a postcondition ψ, which must hold on termination, and a condition φ, which indicates when termination has been reached: $\phi \rightarrow G(\varphi \rightarrow \psi)$. A *liveness* property is one that states that "something good will eventually happen". The simplest liveness properties have the form $F\phi$, stating that eventually, ϕ will hold. *Termination* is an example of liveness. The basic termination property is: $\phi \rightarrow F\varphi$ which states that every run which initially satisfied the property ϕ eventually satisfied the property φ, where φ is the property which holds when a run has terminated. Another useful liveness property is *temporal implication*: $G(\phi \rightarrow F\psi)$ which states that "every ϕ is followed by a ψ". *Responsiveness* is a classic example of temporal implication: suppose ϕ

[2]The material in this section has been adapted from [9, p1049–1054].

Axioms:	
$(LAX1)$	propositional tautologies
$(LAX2)$	$\neg X\phi \leftrightarrow X\neg\phi$
$(LAX3)$	$X(\phi \rightarrow \psi) \rightarrow (X\phi \rightarrow X\psi)$
$(LAX4)$	$G(\phi \rightarrow \psi) \rightarrow (G\phi \rightarrow G\psi)$
$(LAX5)$	$G\phi \rightarrow (\phi \wedge XG\phi)$
$(LAX6)$	$G(\phi \rightarrow X\phi) \rightarrow (\phi \rightarrow G\phi)$
$(LAX7)$	$(\phi U\psi) \rightarrow F\psi$
$(LAX8)$	$(\phi U\psi) \leftrightarrow (\psi \vee (\phi \wedge X(\phi U\psi))$
Inference Rules:	
$(LIR1)$	From $\vdash \phi \rightarrow \psi$ and $\vdash \phi$ infer $\vdash \psi$
$(LIR2)$	From $\vdash \phi$ infer $\vdash G\psi$

Table 1: A complete axiomatization for LTL.

represents a "request", and ψ a "response". The above temporal implication would then state that every request is followed by a response.

A great many technical results have been obtained with respect to LTL. A complete axiomatization was given in [19], and several different axiomatizations have subsequently been presented (see Table 1 for one). Tableau-based proof methods for LTL were introduced by Wolper [45], and resolution proof methods were developed by Fisher [15]. The computational complexity of satisfiability checking for LTL was investigated by Sistla and Clarke, who showed that the problem is PSPACE-complete [38].

One interesting aspect of temporal languages over the natural numbers (and, in particular, LTL) which has turned out to be of great practical and theoretical value in computer science, is the relationship to *automata for infinite words* [30]. Of particular interest in temporal logic are a class of automata known as *Büchi automata*. Büchi automata are those that can recognise *ω-regular expressions*: regular expressions that may contain infinite repetition. A fundamental result in temporal logic theory is that for every LTL formula ϕ one can construct a Büchi automaton A_ϕ that accepts exactly the models of ϕ (more precisely, the infinite words corresponding to models of ϕ). The technique for constructing A_ϕ from ϕ is closely related to Wolper's tableau proof method for temporal logic [45]. This result yields a decision procedure for satisfability of LTL formulae: to determine whether a formula ϕ is satisfiable, construct the automaton A_ϕ and check whether this automaton accepts at least one word (the latter problem

is well-understood and can be solved in polynomial time). We refer the reader to [30] for an overview of automata-based techniques for temporal reasoning.

4.3 Branching Temporal Logic – CTL

If we are interested in the branching structure of reactive programs and their possible computations, then LTL does not seem a very appropriate language. With linear time, there is just one path, so an event either happens or it doesn't happen. But for a reactive program there may be *multiple* possible computations, and an event may occur on some of these, but not on others. How to capture this type of situation? The basic insight in CTL is to talk about possible computations (or futures) by introducing two operators "A" ("on all paths ..."") and "E" ("on some path ..."), called *path quantifiers*, which can be prefixed to a temporal (LTL) formula. For example, the CTL formula $\mathsf{A}Fp$ says "on all possible futures, p will eventually occur", while the CTL formula $\mathsf{E}Fq$ says "there is at least one possible future on which q eventually occurs". CTL imposes one important syntactic restriction on the structure of formulae: a temporal (LTL) operator must be prefixed by a path quantifier. The language without this restriction is known as CTL* [11]: it is much more expressive than CTL, but also much more complex. For simplicity, we will here stick with CTL.

Starting from a set Φ of propositional atoms, the syntax of CTL is defined by the following grammar:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi \vee \phi \mid \mathsf{E}X\phi \mid \mathsf{E}(\phi U \phi) \mid \mathsf{A}X\phi \mid \mathsf{A}(\phi U \phi)$$

where $p \in \Phi$. Given these operators, we can derive the remaining CTL temporal operators as follows:

$$\begin{array}{lcllcl} \mathsf{A}F\phi & \equiv & \mathsf{A}(\top U \phi) & \mathsf{E}F\phi & \equiv & \mathsf{E}(\top U \phi) \\ \mathsf{A}G\phi & \equiv & \neg\mathsf{E}F\neg\phi & \mathsf{E}G\phi & \equiv & \neg\mathsf{A}F\neg\phi \end{array}$$

As in the case of LTL, the semantics of CTL is defined with respect to transition systems. For a state s in a transition system $M = (S, R, V)$ we say that a path ρ is a *s-path* if $\rho[0] = s$. Let $paths(s)$ denote the set of s-paths over M.

The satisfaction relation "$M, s \models \phi$" between pairs M, s and CTL formulae ϕ formalizes the condition "at state s in the transition system M the formula ϕ holds" and is inductively defined via the following rules:

$$M, s \models \top$$

$M, s \models p$ iff $p \in V(s)$ (where $p \in \Phi$);

$M, s \models \neg\phi$ iff $M, s \not\models \phi$

$M, s \models \phi \vee \psi$ iff $M, s \models \phi$ or $M, s \models \psi$

$M, s \models \mathsf{A}X\phi$ iff $\forall \rho \in paths(s) : M, \rho[1] \models \phi$

$M, s \models \mathsf{E}X\phi$ iff $\exists \rho \in paths(s) : M, \rho[1] \models \phi$

$M, s \models \mathsf{A}(\phi U \psi)$ iff $\forall \rho \in paths(s), \exists u \in \mathbb{N}$, s.t. $M, \rho[u] \models \psi$ and $\forall v, (0 \leq v < u) : M, \rho[v] \models \phi$

$M, s \models \mathsf{E}(\phi U \psi)$ iff $\exists \rho \in paths(s), \exists u \in \mathbb{N}$, s.t. $M, \rho[u] \models \psi$ and $\forall v, (0 \leq v < u) : M, \rho[v] \models \phi$

Referring back to the branching time model given in Figure 2(c), we leave the reader to verify that in the initial state, the following formulae are satisfied:

- $\mathsf{E}Xx$
- $\neg\mathsf{A}Xx$
- $\mathsf{A}Fy$
- $\mathsf{E}(xUy)$

At this point, it is a useful exercise to convince oneself that one cannot interpret in any meaningful way LTL formulae in pairs M, s consisting of a model and a state: a run is required to interpret pure temporal formulae without prefixed path quantifiers. Conversely, one cannot interpret CTL formulae in pairs ρ, u consisting of a run and a number: the whole transition system is required to interpret path quantifiers.

As with LTL, many technical results have been obtained with respect to CTL. A complete axiomatization is given in Table 2 (see [10]; for discussion and further references, see [9, p.1040]). Satisfiability of CTL formulas is harder than that of LTL: the problem is EXPTIME-complete [9, p.1037]. As with linear time temporal logic, the relationship to automata is fundamental for understanding the behaviour of CTL and other branching time logics. Here one employs automata for *infinite trees* rather than infinite words. We refer the reader to [30] for an overview.

Axioms:	
$(BAX1)$	propositional tautologies
$(BAX2)$	$\mathsf{E}F\phi \leftrightarrow \mathsf{E}(\top U \phi)$
$(BAX2b)$	$\mathsf{A}G\phi \leftrightarrow \neg\mathsf{E}F\neg\phi$
$(BAX3)$	$\mathsf{A}F\phi \leftrightarrow \mathsf{A}(\top U \phi)$
$(BAX3b)$	$\mathsf{E}G\phi \leftrightarrow \neg\mathsf{A}F\neg\phi$
$(BAX4)$	$\mathsf{E}X(\phi \vee \psi) \leftrightarrow (\mathsf{E}X\phi \vee \mathsf{E}X\psi)$
$(BAX5)$	$\mathsf{A}X\phi \leftrightarrow \neg\mathsf{E}X\neg\phi$
$(BAX6)$	$\mathsf{E}(\phi U \psi) \leftrightarrow (\psi \vee (\phi \wedge \mathsf{E}X\mathsf{E}(\phi U \psi)))$
$(BAX7)$	$\mathsf{A}(\phi U \psi) \leftrightarrow (\psi \vee (\phi \wedge \mathsf{A}X\mathsf{A}(\phi U \psi)))$
$(BAX8)$	$\mathsf{E}X\top \wedge \mathsf{A}X\top$
$(BAX9)$	$\mathsf{A}G(\phi \rightarrow (\neg\psi \wedge \mathsf{E}X\phi)) \rightarrow (\phi \rightarrow \neg\mathsf{A}(\gamma U \psi))$
$(BAX9b)$	$\mathsf{A}G(\phi \rightarrow (\neg\psi \wedge \mathsf{E}X\phi)) \rightarrow (\phi \rightarrow \neg\mathsf{A}F\psi)$
$(BAX10)$	$\mathsf{A}G(\phi \rightarrow (\neg\psi \wedge (\gamma \rightarrow \mathsf{A}X\phi))) \rightarrow (\phi \rightarrow \neg\mathsf{E}(\gamma U \psi))$
$(BAX10b)$	$\mathsf{A}G(\phi \rightarrow (\neg\psi \wedge \mathsf{A}X\phi)) \rightarrow (\phi \rightarrow \neg\mathsf{E}F\psi)$
$(BAX11)$	$\mathsf{A}G(\phi \rightarrow \psi) \rightarrow (\mathsf{E}X\phi \rightarrow \mathsf{E}X\psi)$
Inference Rules:	
$(BIR1)$	From $\vdash \phi$ and $\vdash \phi \rightarrow \psi$ infer $\vdash \psi$
$(BIR2)$	From $\vdash \phi$ infer $\vdash \mathsf{A}G\phi$

Table 2: A complete axiomatization for CTL.

5 Interval Temporal Logics

In all temporal logics we considered so far, temporal formulae were evaluated at *time points* or *states*. An alternative, and typically more powerful, way of evaluating formulae is in *intervals*, sets I of time points with the property that if $t_1 < t_2 < t_3$ and $t_1, t_3 \in I$, then $t_2 \in I$. For example, the sentence "Mary often visits her mother" can be true in a certain interval, say from 2006 to 2008, but it does not make sense to say that it is true at a certain time point or state. Many different interval based temporal logics have been introduced [22, 39, 29]. When designing such a language, the first decision to take is the set of temporal operators. Between time points, there are only three distinct qualitative relations: before, after, and equal. This might be the reason that point-based temporal logics typically employ (some subset) of the rather small set of operators discussed above (Kamp's Theorem provides another explanation). In contrast, there are thirteen distinct qualitative relations between time intervals, known as Allen's rela-

tions [2]. To give just four obvious examples: interval I can be before interval J (for all $t \in I, t' \in J$ we have $t < t'$), I and J can overlap, I can be during J, and I can finish J. The possible choices of temporal operators for interval-based logics reflect these relations. For example, for the relation "before" one can introduce a temporal operator <before> whose truth condition is as follows:

> <before> ϕ is true in interval I if, and only if, there exists an interval J before I such that ϕ is true in J.

Operators <overlap>, <during>, etc. can be introduced in the same way. The resulting temporal language with operators for all 13 Allen relations (or some subset thereof of equal expressivity) has been investigated extensively [22, 39]. In contrast to most point-based temporal logics, the resulting temporal tautologies are typically undecidable. Often (e.g., for the discrete time flow consisting of a copy of the natural numbers and for the time flow consisting of the reals) they are even not recursively enumerable and, therefore, non-axiomatizable [22]. Such a negative result can give rise to an interesting new research program: to classify the fragments of the undecidable/non-axiomatizable logic into those that are decidable/axiomatizable and those that are still undecidable/non-axiomatizable. The paradigmatic example of such a program is the undecidability of classical predicate logic that has transformed Hibert's original Entscheidungs problem into a classification problem asking which fragments of classical predicate logic are decidable [5]. For interval temporal logics a similar (but smaller scale) program has been launched. The recent state of the art for the classification problem for interval temporal logics is described in [7].

From a philosophical as well as mathematical perspective it is also of interest to regard intervals not as derived objects from time points but as primitive objects [39]. The models in which temporal formulae are interpreted are then not collections of time points, but collections of intervals with temporal relations between them. Typical temporal relations one can consider are (subsets of) the set of thirteen Allen relations, however, now one has to explicitly axiomatize their properties rather than derive them from the underlying point based structure. This then opens the door for representation theorems: when is an abstract structure of primitive intervals representable as a concrete structure of intervals induced by time points? Can one describe those structures axiomatically? We refer the reader to [39, 40] for a discussion of this approach and results.

6 Dynamic Logic

At about the same time that Pnueli was first investigating temporal logic, a different class of logics, also based on modal logic, were being developed for reasoning about actions in general, and computer programs in particular. The starting point for this research is the following observation. Temporal logics allow us to describe the time-varying properties of dynamic domains, but they have nothing to say about the *actions* that *cause* these changes; that is, in the language, we have no direct way of expressing things like "Michael turned the motor on". Here "turning the motor on" is an action, and the performance of this action changes the state of the world. There are many situations, however, where it is desirable to be able to explicitly refer to actions and the effects that they have. One such domain that is particularly well-suited for formal representation and reasoning is computer programs. A computer program can be regarded as a list of actions which the computer must execute one after the other. Note that in contrast to many other domains, there is little or no ambiguity about what the actions are; the computer programming language makes the meaning and effect of such actions precise, and this enables us to develop and use formalisms for reasoning about them. Dynamic logics arose from the desire to establish the correctness of computer programs using a logic that explicitly refers to the actions the computer is executing.

An important question to ask about terminating programs is what properties they guarantee, i.e., what properties are guaranteed to hold after they have finished executing. This type of reasoning can be captured using modal operators: we might interpret the formula $[\mathcal{P}]\phi$ to mean that "after all possible terminating runs of the program $\mathcal{P}$, the formula ϕ holds". Given a conventional Kripke semantics, possible worlds are naturally interpreted as the states of a machine executing a program. However, this modal treatment of programs has one key limitation. It treats programs as *atomic*, whereas in reality, programs are highly structured, and this structure is central to understanding their behaviour. So, rather than using a *single* modal "box" operator, the idea in dynamic logic is to use a collection of operators $[\pi]$, one for each program π, where $[\pi]\phi$ then means "on all terminating executions of program π, the property ϕ holds". Crucially, π is allowed to contain *program constructs* such as selection ("if") statements, loops, and the like; the overall behaviour of programs is derived from their component programs. The resulting formalism is known as *dynamic logic*; it was originally formulated by Vaughan Pratt in the late 1970s. Here, we will introduce the best-known variant of dynamic logic, known as *Propositional*

Dynamic Logic (PDL), introduced by Fischer and Ladner [14].

Formally, we define the syntax of programs π and formulae ϕ with respect to a set A of atomic actions and a set Φ of propositional atoms by mutual induction through the following grammar:

$$\begin{array}{lcl} \pi & ::= & \alpha \mid \pi;\pi \mid \pi^* \mid \pi \cup \pi \mid \phi? \\ \phi & ::= & p \mid \neg\phi \mid \phi \vee \phi \mid [\pi]\phi \end{array}$$

where $\alpha \in A$ and $p \in \Phi$.

The program constructs ";", "$\cup$", "$\cdot^*$", and "?" are known as *sequence*, *choice*, *iteration*, and *test*, and closely reflect the basic constructs found in programming languages:

$\pi_1;\pi_2$ means "execute program π_1 and then execute program π_2";

$\pi_1 \cup \pi_2$ means "either execute program π_1 or execute program π_2";

π^* means "repeatedly execute π (an undetermined number of times)";

and

$\phi?$ means "only proceed if ϕ is true."

Let us see a few examples of PDL formulae, and the program properties that they capture.

$$p \rightarrow [\pi]q$$

This asserts that if p is true, then after we have executed program π, we are guaranteed to have q true.

$$p \rightarrow [\pi_1 \cup \pi_2]q$$

This asserts that if p is true, then no matter whether we execute program π_1 or program π_2, ϕ will be true.

The program constructs provided in PDL may seem rather strange to those familiar with programming languages such as C, PASCAL, or JAVA. In particular, we do not seem to have in PDL operators for `if...else` constructs, or the familiar loop constructs such as `while` and `repeat`. However, this is not the case: they can be defined in terms of PDL constructs, as follows. First, consider the following definition of an `if...else` construction:

$$\texttt{if}\ \phi\ \texttt{then}\ \pi_1\ \texttt{else}\ \pi_2 \quad = \quad ((\phi?;\pi_1) \cup (\neg\phi?;\pi_2))$$

A `while` construct can be defined:

$$\texttt{while}\ \phi\ \texttt{do}\ \pi \quad = \quad ((\phi?;\pi)*;\neg\phi?)$$

A `repeat` construct can be defined:

$$\begin{array}{rcl} \texttt{repeat}\ \pi\ \texttt{until}\ \phi & = & \pi;\texttt{while}\ \neg\phi\ \texttt{do}\ \pi \\ & = & \pi;((\neg\phi?;\pi)*;\phi?) \end{array}$$

We invite the reader to convince themselves that these definition do indeed capture the meaning of these operators in C/JAVA-like languages.

The semantics of PDL are somewhat more involved than the logics we have looked at previously. It is based on the semantics of normal modal logic; we have a set S of states, and for each atomic action α we have a relation $R_\alpha \subseteq S \times S$, defining the behaviour of α, with the idea being that $(s, s') \in R_\alpha$ if state s' is one of the possible outcomes that could result from performing action α in state s. Given these atomic relations, we can then obtain accessibility relations for arbitrary programs π, as follows. Let the composition of relations R_1 and R_2 be denoted by $R_1 \circ R_2$, and the reflexive transitive closure (ancestral) of relation R by R^*. Then the accessibility relations for complex programs are defined [23]:

$$\begin{array}{rcl} R_{\pi_1;\pi_2} & = & R_{\pi_1} \circ R_{\pi_2} \\ R_{\pi_1 \cup \pi_2} & = & R_{\pi_1} \cup R_{\pi_2} \\ R_{\pi^*} & = & (R_\pi)^* \\ R_{\phi?} & = & \{(s, s) \mid M, s \models \phi\}. \end{array}$$

Notice that the final clause refers to a satisfaction relation for PDL, $\models$, which has not yet been defined. So let us define this relation. A model for PDL is a structure:

$$M = \langle S, \{R_\alpha\}_{\alpha \in A}, V \rangle$$

where:

- S is a set of states;
- $\{R_\alpha\}_{\alpha \in A}$ is a collection of accessibility relations, one for each atomic program $\alpha \in A$; and
- $V : S \rightarrow 2^\Phi$ gives the set of propositional atoms true in each state.

Given these definitions, the satisfaction relation $\models$ for PDL holds between pairs M, s and formulae:

Axioms:	
$(PAX1)$	propositional tautologies
$(PAX2)$	$[\pi](\phi \rightarrow \psi) \rightarrow ([\pi]\phi \rightarrow [\pi]\psi)$
$(PAX3)$	$[\pi](\phi \wedge \psi) \leftrightarrow [\pi]\phi \wedge [\pi]\psi$
$(PAX4)$	$[\pi_1 \cup \pi_2]\phi \leftrightarrow [\pi_1]\phi \wedge [\pi_2]\phi$
$(PAX5)$	$[\pi_1;\pi_2]\phi \leftrightarrow [\pi_1][\pi_2]\phi$
$(PAX6)$	$[\psi?]\phi \leftrightarrow (\psi \rightarrow \phi)$
$(PAX7)$	$\phi \wedge [\pi][\pi*]\phi \leftrightarrow [\pi^*]\phi$
$(PAX8)$	$\phi \wedge [\pi^*](\phi \rightarrow [\pi]\phi) \rightarrow [\pi^*]\phi$
Inference Rules:	
$(PIR1)$	From $\vdash \phi \rightarrow \psi$ and $\vdash \phi$ infer $\vdash \psi$
$(PIR2)$	From $\vdash \phi$ infer $\vdash [\pi]\psi$

Table 3: A complete axiomatization for PDL.

$M, s \models p$ iff $p \in V(s)$ (where $p \in \Phi$);

$M, s \models \neg\phi$ iff not $M, s \models \phi$

$M, s \models \phi \vee \psi$ iff $M, s \models \phi$ or $M, s \models \psi$

$M, s \models [\pi]$ iff $\forall s' \in S$ such that $(s, s') \in R_\pi$ we have $M, s' \models \phi$

A complete axiomatization of PDL was first given by Segerberg [36] – see Table 3. The satisfiability problem for PDL is EXPTIME-complete [13]. Thus PDL has the same computational complexity as CTL. Numerous extensions of PDL with various additional program constructors such as loop, intersection, and converse have been considered. For an overview, we refer the reader to [23]. There also exist first-order versions of dynamic logics in which the abstract atomic actions of PDL are replaced by "real" atomic programs such as, for example, $x := x + 4$ [23].

7 Further Reading

We emphasise that temporal and dynamic logics are major research areas, with a vast literature behind them. In this short paper, we have been able to do no more than sketch some of the major directions and developments. For more reading, we recommend [43] as a gentle and short introduction to temporal logic, [21] as a mathematical introduction to temporal

and dynamic logic, with particular emphasis on the use of such logics for reasoning about programs, and [9] for an excellent technical introduction to LTL and CTL. A recent collection of papers on temporal reasoning in AI is [17]; a comprehensive overview article, providing many pointers to further reading on temporal logic may be found in [16]. The debate on the relative merits of linear versus branching time logics to a certain extent continues today; see, e.g., [41] for a relatively recent contribution to the debate, with extensive references. The definitive reference to dynamic logic is [23].

References

[1] M. Abadi. *Temporal Logic Theorem Proving*. PhD thesis, Computer Science Department, Stanford University, Stanford, CA 94305, 1987.

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983.

[3] M. Ben-Ari, Z. Manna, and A. Pnueli. The temporal logic of branching time. In *Proceedings of the Eighth ACM Symposium on the Principles of Programming Languages (POPL)*, pages 164–176, 1981.

[4] P. Blackburn, J. van Benthem, and F. Wolter. Preface. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*. Elsevier Science, 2006.

[5] E. Boerger, E. Graedel, and Y. Gurevich. *The Classical Decision Problem*. Springer, 1997.

[6] R. S. Boyer and J. S. Moore, editors. *The Correctness Problem in Computer Science*. The Academic Press: London, England, 1981.

[7] D. Bresolin, D. D. Monica, V. Goranko, A. Montanari, and G. Sciavicco. Decidable and undecidable fragments of halpern and shoham's interval temporal logic: Towards a complete classification. In *LPAR*, pages 590–604, 2008.

[8] J. Burgess. Basic tense logics. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical logic*. Reidel, 1984.

[9] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science Volume B: Formal*

Models and Semantics, pages 996–1072. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1990.

[10] E. A. Emerson and J. Y. Halpern. Decision procedures and expressiveness in the temporal logic of branching time. *Journal of Computer and System Sciences*, 30(1):1–24, 1985.

[11] E. A. Emerson and J. Y. Halpern. 'Sometimes' and 'not never' revisited: on branching time versus linear time temporal logic. *Journal of the ACM*, 33(1):151–178, 1986.

[12] K. Etessami, M. Y. Vardi, and T. Wilke. First-order logic with two variables and unary temporal logic. *Inf. Comput.*, 179(2):279–295, 2002.

[13] M. Fischer and N. J. Pippenger. Relations among complexity measures. *Journal of the ACM*, 26:361–381, 1979.

[14] M. J. Fischer and R. E. Ladner. Propositional dynamic logic of regular programs. *J. Comput. Syst. Sci.*, 18(2):194–211, 1979.

[15] M. Fisher. A resolution method for temporal logic. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, Sydney, Australia, Aug. 1991.

[16] M. Fisher. Temporal representation and reasoning. In F. van Harmelen, V. Lifschitz, and B. Porter, editors, *Handbook of Knowledge Representation*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 2008.

[17] M. Fisher, D. Gabbay, and L. Vila, editors. *Handbook of Temporal Reasoning in Artificial Intelligence*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 2005.

[18] D. Gabbay, I. Hodkinson, and M. Reynolds. *Temporal Logic: Mathematical Foundations and Computational Aspects*. Clarendon Press, 1994.

[19] D. Gabbay, A. Pnueli, S. Shelah, and J. Stavi. On the temporal analysis of fairness. In *Conference Record of the Seventh ACM Symposium on Principles of Programming Languages (POPL '80)*, pages 163–173, New York, USA, Jan. 1980. ACM Press.

[20] A. Galton. Temporal logic and computer science: An overview. In A. Galton, editor, *Temporal Logics and their Applications*, pages 1–52. The Academic Press: London, England, 1987.

[21] R. Goldblatt. *Logics of Time and Computation (CSLI Lecture Notes Number 7)*. Center for the Study of Language and Information, Ventura Hall, Stanford, CA 94305, 1987. (Distributed by Chicago University Press).

[22] J. Y. Halpern and Y. Shoham. A propositional modal logic of time intervals. *J. ACM*, 38(4):935–962, 1991.

[23] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. The MIT Press: Cambridge, MA, 2000.

[24] I. Hodkinson and M. Reynolds. Temporal logic. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*. Elsevier Science, 2006.

[25] H. Kamp. *Tense logic and the theory of linear order*. PhD thesis, University of California, 1968.

[26] L. Lamport. Sometimes is sometimes not never — but not always. In *Proceedings of the Seventh ACM Symposium on the Principles of Programming Languages (POPL)*, 1980.

[27] T. Litak and F. Wolter. All finitely axiomatizable tense logics of linear time flows are conp-complete. *Studia Logica*, 81(2):153–165, 2005.

[28] C. Lutz, U. Sattler, and F. Wolter. Modal logic and the two-variable fragment. In *CSL*, pages 247–261, 2001.

[29] B. C. Moszkowski. A complete axiomatization of interval temporal logic with infinite time. In *LICS*, pages 241–252, 2000.

[30] M.Y.Vardi. Automata-theoretic techniques for temporal reasoning. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*. Elsevier Science, 2006.

[31] A. Pnueli. The temporal logic of programs. In *Proceedings of the Eighteenth IEEE Symposium on the Foundations of Computer Science*, pages 46–57, 1977.

[32] A. Prior. *Time and Modality*. Oxford University Press, 1957.

[33] A. Prior. *Past, Present, and Future*. Oxford University Press, 1967.

[34] A. Prior. A statement of temporal realism. In B.J.Copeland, editor, *Logic and Reality: Essays on the Legacy of Arthur Prior*. Clarendon Press, 1996.

[35] H. Reichgelt. A comparison of first-order and modal logics of time. In P. Jackson, H. Reichgelt, and F. van Harmelen, editors, *Logic Based Knowledge Representation*, pages 143–176. The MIT Press: Cambridge, MA, 1989.

[36] K. Segerberg. A completeness theorem in the modal logic of programs. *Not. Amer. Math. Soc.*, 24(6), 1977.

[37] A. Sistla. Theoretical issues in the design and verification of distributed systems. Technical Report CMU-CS-83-146, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1983.

[38] A. P. Sistla and E. M. Clarke. The complexity of propositional linear temporal logics. *Journal of the ACM*, 32(3):733–749, 1985.

[39] J. van Benthem. *The Logic of Time*. Kluwer, 1983.

[40] J. van Benthem. Temporal logic. In D. Gabbay, C. Hogger, and J. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4, pages 241–351. Oxford University Press, 1995.

[41] M. Y. Vardi. Branching vs. linear time: Final showdown. In T. Margaria and W. Yi, editors, *Proceedings of the 2001 Conference on Tools and Algorithms for the Construction and Analysis of Systems, TACAS 2001 (LNCS Volume 2031)*, pages 1–22. Springer-Verlag: Berlin, Germany, Apr. 2001.

[42] M. Y. Vardi. From philosophical to industrial logics. In *ICLA*, pages 89–115, 2009.

[43] Y. Venema. Temporal logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, pages 203–223. Blackwell Publishers, 2001.

[44] P. Wolper. Temporal logic can be more expressive. *Information and Control*, 56, 1983.

[45] P. Wolper. The tableau method for temporal logic: An overview. *Logique et Analyse*, 110–111, 1985.

Logic and Categories as Tools for Building Theories

SAMSON ABRAMSKY*

1 Introduction

My aim in this short article is to provide an impression of some of the ideas emerging at the interface of logic and computer science, in a form which I hope will be accessible to philosophers.

Why is this even a good idea? Because there has been a huge interaction of logic and computer science over the past half-century which has not only played an important rôle in shaping Computer Science, but has also greatly broadened the scope and enriched the content of logic itself.[1]

This huge effect of Computer Science on Logic over the past five decades has several aspects: new ways of *using* logic, new attitudes to logic, new questions and methods. These lead to new perspectives on the question:

What logic is — and should be!

Our main concern is with method and attitude rather than matter; nevertheless, we shall base the general points we wish to make on a case study: *Category theory*. Many other examples could have been used to illustrate our theme, but this will serve to illustrate some of the points we wish to make.

2 Category Theory

Category theory is a vast subject. It has enormous potential for any serious version of 'formal philosophy' — and yet this has hardly been realized.

We shall begin with introduction to some basic elements of category theory, focussing on the fascinating conceptual issues which arise even at

*Oxford University Computing Laboratory

[1]This view is not universally held, either among Computer Scientists or logicians, but I and many of my colleagues do believe it — and we are right!

the most elementary level of the subject, and then discuss some its consequences and philosophical ramifications.

2.1 Some Basic Notions of Category Theory

We briefly recall the basic definitions. A category has a collections of *objects* $A, B, C, \ldots$, and a collection of *arrows* (or *morphisms*) $f, g\ h \ldots$. Each arrow has specified objects as its *domain* and *codomain*.[2] We write $f : A \to B$ for an arrow with domain A and codomain B. For any triple of objects A, B, C there is an operation of *composition*: given $f : A \to B$ and $g : B \to C$, we can form $g \circ f : A \to C$. Note that the codomain of f has to match with the domain of g. Moreover, for each object A, there is an *identity arrow* $\mathsf{id}_A : A \to A$. These data are subject to the following axioms:

$$h \circ (g \circ f) = (h \circ g) \circ f \qquad f \circ \mathsf{id}_A = f = \mathsf{id}_B \circ f$$

whenever the indicated compositions make sense, *i.e.* the domains and codomains match appropriately.

These definitions appear at first sight fairly innocuous: some kind of algebraic structure, reminiscent of monoids (groups without inverses), but with the clumsy-looking apparatus of objects, domains and codomains restricting the possibilities for composition of arrows. These first appearances are deceptive, as we shall see, although in a few pages we can only convey a glimpse of the richness of the notions which arise as the theory unfolds.

Let us now see some first examples of categories.

- The most basic example of a category is **Set**: the objects are sets, and the arrows are functions. Composition and identities have their usual meaning for functions.
- Any kind of mathematical structure, together with structure preserving functions, forms a category. E.g.
  - **Mon** (monoids and monoid homomorphisms)
  - **Grp** (groups and group homomorphisms)
  - $\mathbf{Vect}_k$ (vector spaces over a field k, and linear maps)
  - **Pos** (partially ordered sets and monotone functions)

[2] More formally, there are operations dom, cod from arrows to objects.

  - **Top** (topological spaces and continuous functions)

- **Rel**: objects are sets, arrows $R : X \to Y$ are *relations* $R \subseteq X \times Y$. Relational composition:

$$R; S(x, z) \iff \exists y.\, R(x, y) \ \wedge \ S(y, z)$$

- Let k be a field (for example, the real or complex numbers). Consider the following category $\mathbf{Mat}_k$. The objects are natural numbers. A morphism $M : \mathbf{n} \to \mathbf{m}$ is an $\mathbf{n} \times \mathbf{m}$ matrix with entries in k. Composition is matrix multiplication, and the identity on $\mathbf{n}$ is the $\mathbf{n} \times \mathbf{n}$ diagonal matrix.

- Monoids are one-object categories. Arrows correspond to the elements of the monoid, composition of arrows to the monoid multiplication, and the identity arrow to the monoid unit.

- A category in which for each pair of objects A, B there is at most one morphism from A to B is the same thing as a *preorder*, *i.e.* a reflexive and transitive relation. Note that the identity arrows correspond to reflexivity, and composition to transitivity.

2.1.1 Categories as Contexts and as Structures

Note that our first class of examples illustrate the idea of categories as *mathematical contexts*; settings in which various mathematical theories can be developed. Thus for example, **Top** is the context for general topology, **Grp** is the context for group theory, etc.

This issue of "mathematics in context" should be emphasized. The idea that any mathematical discussion is relative to the category we happen to be working in is pervasive and fundamental. It allows us simultaneously to be both properly specific and general: specific, in that statements about mathematical structures are not really precise until we have specified which structures we are dealing with, *and* which morphisms we are considering — *i.e.* which category we are working in. At the same time, the awareness that we are working in some category allows us to extract the proper generality for any definition or theorem, by identifying exactly which properties of the ambient category we are using.

On the other hand, the last two examples illustrate that many important mathematical structures *themselves appear as categories of particular kinds*. The fact that two such different kinds of mathematical structures

as monoids and posets should appear as extremal versions of categories is also rather striking.

This ability to capture mathematics both "in the large" and "in the small" is a first indication of the flexibility and power of categories.

2.1.2 Arrows vs. Elements

Notice that the axioms for cateories are formulated purely in terms of the algebraic operations on arrows, without any reference to 'elements' of the objects. Indeed, in general elements are not available in a category. We will refer to any concept which can be defined purely in terms of composition and identities as *arrow-theoretic*. We will now take a first step towards learning to "think with arrows" by seeing how we can replace some familiar definitions for functions between sets couched in terms of elements by arrow-theoretic equivalents.

We say that a function $f : X \longrightarrow Y$ is:

$$\begin{array}{lll} \textit{injective} & \text{if} & \forall x, x' \in X.\ f(x) = f(x') \implies x = x' \,, \\ \textit{surjective} & \text{if} & \forall y \in Y.\ \exists x \in X.\ f(x) = y \,, \end{array}$$

$$\begin{array}{lll} \textit{monic} & \text{if} & \forall g, h : Z \to X.\ f \circ g = f \circ h \implies g = h \,, \\ \textit{epic} & \text{if} & \forall g, h : Y \to Z.\ g \circ f = h \circ f \implies g = h \,. \end{array}$$

Note that injectivity and surjectivity are formulated in terms of elements, while epic and monic are arrow-theoretic.

Proposition 1. *Let $f : X \to Y$. Then:*

1. *f is injective iff f is monic.*
2. *f is surjective iff f is epic.*

Proof We show 1. Suppose $f : X \to Y$ is injective, and that $f \circ g = f \circ h$, where $g, h : Z \to X$. Then for all $z \in Z$:

$$f(g(z)) = f \circ g(z) = f \circ h(z) = f(h(z)) \,.$$

Since f is injective, this implies $g(z) = h(z)$. Hence we have shown that

$$\forall z \in Z.\ g(z) = h(z) \,,$$

and so we can conclude that $g = h$. So f injective implies f monic. For the converse, fix a one-element set $\mathbf{1} = \{\bullet\}$. Note that elements $x \in X$ are in

1–1 correspondence with functions $\bar{x} : \mathbf{1} \to X$, where $\bar{x}(\bullet) := x$. Moreover, if $f(x) = y$ then $\bar{y} = f \circ \bar{x}$. Writing injectivity in these terms, it amounts to the following:

$$\forall x, x' \in X.\ f \circ \bar{x} = f \circ \bar{x}' \implies \bar{x} = \bar{x}'.$$

Thus we see that being injective is a *special case* of being monic. □

The reader will enjoy — and learn from — proving the equivalence for functions of the conditions of being surjective and epic.

2.1.3 Generality of Notions

Since the concepts of monic and epic are defined in purely arrow-theoretic terms, *they make sense in any category*. This possibility for making definitions in vast generality by formulating them in purely arrow-theoretic terms can be applied to virtually all the fundamental notions and constructions which pervade mathematics.

As an utterly elementary, indeed "trivial" example, consider the notion of isomorphism. What is an isomorphism *in general*? On might try a definition at the level of generality of model theory, or Bourbaki-style structures, but this is really both unnecessarily elaborate, and still insufficiently general. Category theory has exactly the language needed to give a perfectly general answer to the question, in *any mathematical context, as specified by a category*. An isomorphism in a category C is an arrow $f : A \to B$ with a two-sided inverse: an arrow $g : B \to A$ such that

$$g \circ f = \mathsf{id}_A, \qquad f \circ g = \mathsf{id}_B.$$

One can check that in **Set** this yields the notion of bijection; in **Grp** it yields isomorphism of groups; in **Top** it yields homeomorphism; in $\mathbf{Mat}_k$, it yields the usual notion of invertible matrix; and so on throughout the range of mathematical structures. In a monoid considered as a category, an isomorphism is an invertible element. Thus a group is exactly a one-object category in which every arrow is an isomorphism! This cries out for generalization; and the notion of a category in which every arrow is an isomorphism is indeed significant — it is the idea of a *groupoid*, which plays a key rôle in modern geometry and topology.

We also see here a first indication of the *prescriptive nature* of categorical concepts. Having defined a category, what the notion of isomorphism means inside that category is now *fixed* by the general definition. We can observe and characterize what that notion is; if it isn't right for our purposes, we need to work in a different category.

2.1.4 Replacing Coding by Intrinsic Properties

We now consider one of the most common constructions in mathematics: the formation of "direct products". Once again, rather than giving a case-by-case construction of direct products in each mathematical context we encounter, we can express once and for all a general notion of product, meaningful in any category — and such that, if a product exists, it is characterised uniquely up to unique isomorphism. Given a particular mathematical context, *i.e.* a category, we can then verify whether on not the product exists in that category. The concrete construction appropriate to the context will enter only into the proof of *existence*; all of the useful *properties* of the product follow from the general definition. Moreover, the categorical notion of product has a *normative* force; we can test whether a concrete construction works as intended by verifying that it satisfies the general definition.

In set theory, the cartesian product is defined in terms of the ordered pair:

$$X \times Y := \{(x, y) \mid x \in X \ \wedge \ y \in Y\}.$$

It turns out that ordered pairs can be *defined* in set theory, e.g. as

$$(x, y) := \{\{x, y\}, y\}.$$

Note that in no sense is such a definition canonical. The essential *properties* of ordered pairs are:

1. We can retrieve the first and second components x, y of the ordered pair (x, y), allowing *projection functions* to be defined:

$$\pi_1 : (x, y) \mapsto x, \qquad \pi_2 : (x, y) \mapsto y\,.$$

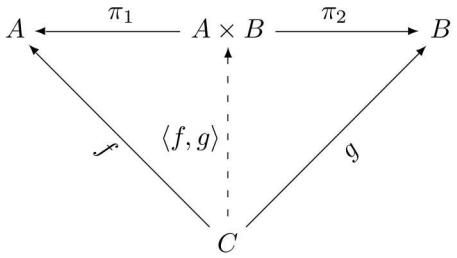
2. The information about first and second components completely determines the ordered pair:

$$(x_1, x_2) = (y_1, y_2) \iff x_1 = y_1 \ \wedge \ x_2 = y_2.$$

The categorical definition expresses these properties in arrow-theoretic terms, meaningful in any category.

Let A, B be objects in a category C. A *product* of A and B is an object $A \times B$ together with a pair of arrows $A \xleftarrow{\pi_1} A \times B \xrightarrow{\pi_2} B$ such that for *every* such triple $A \xleftarrow{f} C \xrightarrow{g} B$ there exists a *unique* morphism

$$\langle f, g \rangle : C \longrightarrow A \times B$$



such that the following diagram commutes.

Writing the equations corresponding to this commuting diagram explicitly, one obtains:

$$\pi_1 \circ \langle f, g \rangle = f, \qquad \pi_2 \circ \langle f, g \rangle = g.$$

Moreover, $\langle f, g \rangle$ is *the unique morphism* $h : C \to A \times B$ satisfying these equations.

To relate this definition to our earlier discussion of definitons of pairing for sets, note that a 'pairing' $A \xleftarrow{f} C \xrightarrow{g} B$ offers a decomposition of C into components in A and B, at the level of arrows rather than elements. The fact that pairs are uniquely determined by their components is expressed in arrow-theoretic terms by the *universal property* of the product; the fact that for *every* candidate pairing, there is a unique arrow into the product, which commutes with taking components.

As immediate evidence that this definition works in the right way, we note the following properties of the categorical product (which of course hold in *any* category):

- The product is determined *uniquely up to unique isomorphism*. That is, if there are two pairings satisfying the universal property, there is a unique isomorphism between them which commutes with taking components. This sweeps away all issues of coding and concrete representation, and shows that we have isolated the essential content of the notion of product. We shall prove this property for the related case of terminal objects in the next subsection.

- We can also express the universal property in purely equational terms. This equational specification of products requires that we have a pairing $A \xleftarrow{\pi_1} A \times B \xrightarrow{\pi_2} B$ satisfying the equation

$$\pi_1 \circ \langle f, g \rangle = f, \qquad \pi_2 \circ \langle f, g \rangle = g$$

as before, and additionally, for any $h : C \to A \times B$:

$$h = \langle \pi_1 \circ h, \pi_2 \circ h \rangle.$$

This says that any map into the product is uniquely determined by its components. This equational specification is equivalent to the definition given previously.

We look at how this definition works in some of our example categories.

- In **Set**, products are the usual cartesian products.
- In **Pos**, products are cartesian products with the pointwise order.
- In **Top**, products are cartesian products with the product topology.
- In $\mathbf{Vect}_k$, products are direct sums.
- In a poset, seen as a category, products are *greatest lower bounds*.

2.1.5 Terminal Objects

Our discussion in the previous sub-section was for *binary* products. The same idea can be extended to define the product of any family of objects in a category. In particular, the apparently trivial idea of the product of an empty family of objects turns out to be important. The product of an empty family of objects in a category C will be an object $\mathbf{1}$; there are no projections, since there is nothing in the family to project to! The universal property turns into the following: for each object A in C, there is a unique arrow from C to $\mathbf{1}$. Note that compatibility with the projections trivially holds, since there are no projections! This 'empty product' is the notion of *terminal object*, which again makes sense in any category.

Examples

- In **Set**, any one-element set $\{\bullet\}$ is terminal.
- In **Pos**, the poset $(\{\bullet\}, \{(\bullet, \bullet)\})$ is terminal.
- In **Top**, the space $(\{\bullet\}, \{\emptyset, \{\bullet\}\})$ is terminal.
- In $\mathbf{Vect}_k$, the one-element space $\{0\}$ is terminal.
- In a poset, seen as a category, a terminal object is a greatest element.

We shall now prove that terminal objects are *unique up to (unique) isomorphism*. This property is characteristic of all such "universal" definitions. For example, the apparent arbitrariness in the fact that any singleton set is a terminal object in **Set** is answered by the fact that what counts is the property of being terminal; and this suffices to ensure that any two objects having this property must be isomorphic to each other.

The proof of the proposition, while elementary, is a first example of distinctively categorical reasoning.

Proposition 2. *If T and T' are terminal objects in the category C then there exists a unique isomorphism $T \cong T'$.*

Proof Since T is terminal and T' is an object of C, there is a unique arrow $\tau_{T'} : T' \to T$. We claim that $\tau_{T'}$ is an isomorphism. Since T' is terminal and T is an object in C, there is an arrow $\tau'_T : T \to T'$. Thus we obtain $\tau_{T'} \circ \tau'_T : T \to T$, while we also have the identity morphism $\mathsf{id}_T : T \to T$. But T is terminal, and therefore there exists a *unique* arrow from T to T, which means that $\tau_{T'} \circ \tau'_T = \mathsf{id}_T$. Similarly, $\tau'_T \circ \tau_{T'} = \mathsf{id}_{T'}$, so $\tau_{T'}$ is indeed an isomorphism. □

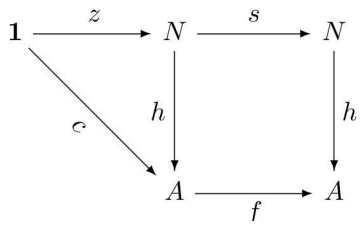
One can reduce the corresponding property for binary products to this one, since the definition of binary product is equivalently expressed by saying that the pairing $A \xleftarrow{\pi_1} A \times B \xrightarrow{\pi_2} B$ is terminal in the category of such pairings, where the morphisms are arrows preserving the components.

It is straightforward to show that if a category has a terminal object, and all binary products, then it has products of all finite families of objects. Thus these are the two cases usually considered.

2.1.6 Natural Numbers

We might suppose that category theory, while suitable for formulating general notions and structures, would not work well for specific mathematical objects such as the number systems. In fact, this is not the case, and the idea of *universal definition*, which we just caught a first glimpse of in the categorical notion of product, provides a powerful tool for specifying the basic discrete number systems of mathematics. We shall illustrate this with the most basic number system of all — the natural numbers (*i.e.* the non-negative integers).

Suppose that C is a category with a terminal object $\mathbf{1}$. We define a *natural numbers object* in C to be an object N together with arrows $z : \mathbf{1} \to N$ and $s : N \to N$ such that, for every such triple of an object A and arrows $c : \mathbf{1} \to A$, $f : A \to A$, *there exists a unique arrow* (note this characteristic



property of universal definitions again) $h : N \to A$ such that the following diagram commutes:

Equivalently, this means that the following equations hold:

$$h \circ z = c, \qquad h \circ s = f \circ h.$$

Once again, the universal property implies that if a natural numbers object exists in C, it is unique up to unique isomorphism. We are not committed to any particular representation of natural numbers; we have specified the properties a structure with a constant and a unary operation must have in order to function as the natural numbers in a particular mathematical context.

In **Set**, we can verify that $\mathbb{N} = \{0, 1, 2, \ldots\}$ equipped with

$$z : \{\bullet\} \to \mathbb{N} :: \bullet \mapsto 0, \qquad s : \mathbb{N} \to \mathbb{N} :: n \mapsto n + 1$$

does indeed form a natural numbers object. But we are not committed to any particular set-theoretic representation of $\mathbb{N}$: whether as von Neumann ordinals, Zermelo numerals [10] or anything else. Indeed, any countable set X with a particular element x picked out by a map z, and a unary operation $s : X \to X$ which is injective and has $X \setminus \{x\}$ as its image, will fulfil the definition; and any two such systems will be canonically isomorphic.

Note that, if we are given a natural numbers object (N, z, s) in an abstract category C, the resources of *definition by primitive recursion* are available to us. Indeed, we can define *numerals relative to* N: $\bar{n} : \mathbf{1} \to N := s^n \circ z$. Here s^n is defined inductively: $s^1 = s$, $s^{n+1} = s \circ s^n$.[3] Given any (A, c, f), with the unique arrow $h : N \to A$ given by the universal property, we can check that $h \circ \bar{n} = f^n \circ c$. In fact, if we assume that C has finite products, and refine the definition of natural numbers object to allow for parameters, or

[3]There is a metainduction going on here, using a natural number object *outside* the category under discussion. This is not essential, but is a useful device for seeing what is going on.

if we keep the definition of natural numbers object as it is but assume that C is *cartesian closed* [19], then all primitive recursive function definitions can be interpreted in C, and will have their usual equational properties.

2.1.7 Functors: category theory takes its own medicine

Part of the "categorical philosophy" is:

Don't just look at the objects; take the morphisms into account too.

We can also apply this to categories! A "morphism of categories" is a *functor*. A functor $F : C \to \mathcal{D}$ is given by:

- An object map, assigning an object FA of $\mathcal{D}$ to every object A of C.
- An arrow map, assigning an arrow $Ff : FA \to FB$ of $\mathcal{D}$ to every arrow $f : A \to B$ of C, in such a way that composition and identities are preserved:

$$F(g \circ f) = Fg \circ Ff\,, \qquad F\mathsf{id}_A = \mathsf{id}_{FA]}.$$

Note that we use the same symbol to denote the object and arrow maps; in practice, this never causes confusion. The conditions expressing preservation of composition and identities are called *functoriality*.

As a first glimpse as to the importance of functoriality, the following fact can be noted:

Proposition 3. *Functors preserve isomorphisms; if $f : A \to B$ is an isomorphism, so is Ff.*

Proof Suppose that f is an isomorphism, with inverse f^{-1}. Then

$$F(f^{-1}) \circ F(f) = F(f^{-1} \circ f) = F(\mathsf{id}_A) = \mathsf{id}_{FA}$$

and similarly $F(f) \circ F(f^{-1}) = \mathsf{id}_{FB}$. So $F(f^{-1})$ is a two-sided inverse for Ff, which is thus an isomorphism. □

Examples

- Let $(P, \leq)$, $(Q, \leq)$ be preorders (seen as categories). A functor $F : (P, \leq) \longrightarrow (Q, \leq)$ is specified by an object-map, say $F : P \to Q$, and an appropriate arrow-map. The arrow-map corresponds to the condition

$$\forall p_1, p_2 \in P.\, p_1 \leq p_2 \;\Rightarrow\; F(p_1) \leq F(p_2)\,,$$

  i.e. to monotonicity of F. Moreover, the functoriality conditions are trivial since in the codomain $(Q, \leq)$ all hom-sets are singletons.Hence, a functor between preorders is just a monotone map.

- Let $(M, \cdot, 1)$, $(N, \cdot, 1)$ be monoids. A functor $F : (M, \cdot, 1) \longrightarrow (N, \cdot, 1)$ is specified by a trivial object map (monoids are categories with a single object) and an arrow-map, say $F : M \to N$. The functoriality conditions correspond to

$$\forall m_1, m_2 \in M.\, F(m_1 \cdot m_2) = F(m_1) \cdot F(m_2)\,, \qquad F(1) = 1\,,$$

  i.e. to F being a monoid homomorphism.Hence, a functor between monoids is just a monoid homomorphism.

Some further examples:

- The *covariant* powerset functor $\mathcal{P} : \mathbf{Set} \to \mathbf{Set}$:

$$X \mapsto \mathcal{P}(X)\,, \qquad (f : X \to Y) \mapsto \mathcal{P}(f) := S \mapsto \{f(x) \mid x \in S\}.$$

- More sophisticated examples: e.g. *homology*. The basic idea of algebraic topology is that there are functorial assignments of algebraic objects (e.g. groups) to topological spaces. The fact that functoriality implies that isomorphisms are preserved shows that these assignments are *topological invariants*. Variants of this idea ('(co)homology theories') are pervasive throughout modern pure mathematics.

2.1.8 The category of categories

There is a category **Cat** whose objects are categories, and whose arrows are functors. Identities in **Cat** are given by identity functors:

$$\mathsf{Id}_C : C \longrightarrow C := A \mapsto A,\; f \mapsto f.$$

Composition of functors is defined in the evident fashion. Note that if $F : C \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{E}$ then, for $f : A \to B$ in C,

$$G \circ F(f) := G(F(f)) : G(F(A)) \longrightarrow G(F(B))$$

so the types work out. A category of categories sounds (and is) circular, but in practice is harmless: one usually makes some size restriction on the categories, and then **Cat** will be too 'big' to be an object of itself.

2.1.9 Logical notions as adjunctions

Finally, we shall take a glimpse at the fundamental notion of *adjunction*; not in its most general form, but in some examples arising from logic [22], which also give a first impression of the deep connections which exist between category theory and logic.

We begin with implication. Implication and conjunction — whether classical or intuitionistic — are related by the following bidirectional inference rule:

$$\frac{\phi \ \wedge \ \psi \vdash \theta}{\phi \vdash \psi \to \theta}\,.$$

If we form the preorder of formulas related by entailment as a category, this rule becomes a relationship between arrows which holds in this category. In fact, it can be shown that this *uniquely characterizes* implication, and is a form of universal definition. Note that it gives the essence of what implication is. The way one proves an implication—essentially the *only* way—is to add the antecedent to one's assumptions and then prove the consequent. This is justified by the above rule.

In terms of the boolean algebra of sets, define $X \to Y = X^c \cup Y$, where X^c is the set complement. Then we have, for any sets X, Y, Z:

$$X \cap Y \subseteq Z \quad \Longleftrightarrow \quad X \subseteq Y \to Z.$$

The same algebraic relation holds in any Heyting algebra, and defines intuitionistic implication.

Now we show that this same formal structure underpins quantification. This is the fundamental insight due to Lawvere [22], that *quantifiers are adjoints to substitution*.

Consider a function $f : X \to Y$. This induces a function

$$f^{-1} : \mathcal{P}(Y) \longrightarrow \mathcal{P}(X) :: T \mapsto \{x \in X \mid f(x) \in T\}.$$

This function f^{-1} has both a left adjoint $\exists(f) : \mathcal{P}(X) \longrightarrow \mathcal{P}(Y)$, and a right adjoint $\forall(f) : \mathcal{P}(X) \longrightarrow \mathcal{P}(Y)$. These adjoints are uniquely specified by the following conditions. For all $S \subseteq X$, $T \subseteq Y$:

$$\exists(f)(S) \subseteq T \iff S \subseteq f^{-1}(T), \qquad f^{-1}(T) \subseteq S \iff T \subseteq \forall(f)(S).$$

The unique functions satisfying these conditions can be defined explicitly as follows:

$$\begin{aligned} \exists(f)(S) &:= \{y \in Y \mid \exists x \in X.\, f(x) = y \ \wedge \ x \in S\}, \\ \forall(f)(S) &:= \{y \in Y \mid \forall x \in X.\, f(x) = y \ \Rightarrow \ x \in S\}. \end{aligned}$$

Given a formula ϕ with free variables in $\{v_1, \ldots, v_{n+1}\}$, it will receive its Tarskian denotation $[\![\phi]\!]$ in $\mathcal{P}(A^{n+1})$ as the set of satisfying assignments:

$$[\![\phi]\!] = \{s \in A^{n+1} \mid s \models_X \phi\}.$$

We have a projection function

$$\pi : A^{n+1} \longrightarrow A^n \ :: (a_1, \ldots, a_{n+1}) \mapsto (a_1, \ldots, a_n).$$

Note that this projection is the Tarskian denotation of the tuple of terms $(v_1, \ldots, v_n)$. We can characterize the standard quantifiers as *adjoints to this projection*:

$$[\![\forall v_{n+1}.\, \phi]\!] = \forall(\pi)([\![\phi]\!]), \qquad [\![\exists v_{n+1}.\, \phi]\!] = \exists(\pi)([\![\phi]\!]).$$

More explicitly, the Tarski semantics over a structure $\mathcal{M} = (A, \ldots)$ assigns such formulas values in $\mathcal{P}(A^{n+1})$. We can regard the quantifiers $\exists v_{n+1}$, $\forall v_{n+1}$ as functions

$$\exists(\pi), \forall(\pi) : \mathcal{P}(A^{n+1}) \longrightarrow \mathcal{P}(A^n)$$

$$\begin{aligned} \exists(\pi)(S) &= \{s \in A^n \mid \exists a \in A.\, s[v_{n+1} \mapsto a] \in S\} \\ \forall(\pi)(S) &= \{s \in A^n \mid \forall a \in A.\, s[v_{n+1} \mapsto a] \in S\} \end{aligned}$$

If we unpack the adjunction conditions for the universal quantifier, they yield the following bidirectional inference rule:

$$\frac{\Gamma \vdash_X \phi}{\Gamma \vdash_X \forall v_{n+1}.\, \phi} \qquad X = \{v_1, \ldots, v_n\}.$$

Here the set X keeps track of the free variables in the assumptions Γ. Note that the usual "eigenvariable condition" is automatically taken care of in this way.

Since adjoints are uniquely determined, this characterization completely captures the meaning of the quantifiers.

2.2 Discussion: the significance of category theory

We turn from this all too brief glimpse at the basics of category theory to discuss its conceptual significance, and why it might matter to philosophy.

The basic feature of category theory which makes it conceptually fascinating and worthy of philosophical study is that it is not just another mathematical theory, but a way of mathematical thinking, and of doing mathematics, which is genuinely distinctive, and in particular very different to the prevailing set-theoretic style which preceded it. If one wanted a clear-cut example of a paradigm-shift in the Kuhnian sense within mathematics, involving a new way of looking at the mathematical universe, then the shift from the set-theoretic to the categorical perspective provides the most dramatic example we possess.

This has been widely misunderstood. Category theory has been portrayed, sometimes by its proponents, but more often by its detractors, as offering an alternative foundational scheme for mathematics to set theory. But this is to miss the point. What category theory offers is an alternative to *foundational schemes in the traditional sense themselves*. This point has been argued with great clarity and cogency in a forceful and compelling essay by Steve Awodey [7]. We shall not attempt to replicate his arguments, but will just make some basic observations.

Firstly, it must be emphasized that the formalization of mathematics within the language of set theory, as developed in the first half of the twentieth century, has been extremely successful, and has enabled the formulation of mathematical definitions and arguments with a previously unparalleled degree of precision and rigour. However, the set-theoretical paradigm has some deficiencies.

The set-theoretical formalization of mathematics rests on the idea of representing mathematical objects as sets which can be defined within a formal set theory, typically ZFC. It is indeed a significant empirical observation, as remarked by Blass [11], that mathematical objects can be thus represented, and mathematical proofs carried out using the axioms of set theory. This leads to claims such as the recent one by Kunen [18] (p. 14), that

> All abstract mathematical concepts are set-theoretic.
> All concrete mathematical objects are specific sets.

This claim fails to ring true, for several reasons.

- Firstly, the set-theoretic representation is *too concrete*. it involves irrelevant details and choices — it is a coding rather than a structural representation of the concepts at hand. We saw this illustrated

with the issue of defining ordered pairs in set theory, and the contrast with the categorical definition of product, which extracted the essential structural features of pairing at the right level of abstraction. Even if we think of number systems, the representation say of the natural numbers as the finite ordinals in set theory is just a particular coding — there are many others. The essential features of the natural numbers are, rather, conveyed by the universal definition of *natural numbers object* — which makes sense in any category. We should not ask what natural numbers *are*, but rather what they *do* — or what we can do with them. Set theoretic representations of mathematical objects give us too much information — and information of the wrong kind.

- Furthermore, by being too specific, set theoretic representations lose much of the generality that mathematical concepts, as used by mathematicians, naturally have. Indeed, the notion of natural numbers object makes sense in any category with a terminal object. Moreover, as a universal consruction, if it exists in a given category, it is unique up to unique isomorphism. Once we are in a particular mathematical context specified by a category, we can *look and see* what the natural numbers object is — while knowing that the standard reasoning principles such as proof by induction and definition by primitive recursion will hold.

- When one passes to more inherently structural notions, such as 'cohomology theory' or 'coalgebra' the assertion that 'all abstract mathematical concepts are set-theoretic' becomes staggeringly implausible, unless we replace 'are' by 'are codable into'. The crudity of the pure set-theoretic language becomes all too apparent. One might indeed say that insensitivity to the distortions of coding is a tell-tale feature of the set-theoretic cast of thought.

  It may be useful to draw an analogy here with geometry. A major theme of 20th century geometry was the replacement of coordinate-based definitions of geometrical notions (such as tensors or varieties) with 'intrinsic' definitions. Coordinates are still very useful for calculations, but the intrinsic definitions are more fundamental and more illuminating — and ultimately more powerful. The move from set-theoretical encodings, which identify mathematical structures with specific entities in the set-theoretical universe, to universal characterizations which make sense in any mathematical context

(category) satisfying some given background conditions, similarly leads to greater insight and technical power.

The foundationalist critique of category theory proceeds as follows:

1. Category theory cannot emancipate itself completely from set theory, and indeed relies on set theory at certain points.

2. Hence it is not truly fundamental, and cannot serve as a foundation for mathematics.

On the first point, one can discern two main arguments.

- Firstly the very definition of category and functor presuppose the notion of a *collection* of things, and of *operations* on these things. So one needs an underlying theory of collections and operations as a substrate for category theory.

  This is true enough; but the required 'theory of collections and operations' is quite rudimentary. Certainly nothing like formal set theory is presupposed. In fact, the basic notions of categories are *essentially algebraic* in form [14]; that is, they can be formalized as partial algebras, in which the domains of definition of the operations can themselves be defined equationally, in terms of operations which have already been specified. For example, if we consider composition as a partial binary operation comp on arrows, then $\mathsf{comp}(g, f)$ is defined just when $\mathsf{cod}(f) = \mathsf{dom}(g)$.

- The second argument is that at various points, issues of *size* enter into category theory. We saw an example of this in considering the category **Cat** of categories and functors. Is **Cat** an object of **Cat**? To avoid such issues, one usually defines a version of **Cat** with some size restriction; for example, one only considers categories whose underlying collection of arrows form a set in Zermelo-Fraenkel set theory. Then **Cat** will be too large (a proper class) to be an object of itself.

  There are various technical elaborations of this point. One can consider categories of arbitrary size in a stratified fashion, by assuming a sufficient supply of inaccessible cardinals (and hence of 'Grothendieck universes' [11]). One can also formalize notions of size relative to an ambient category one is 'working inside'; which actually describes what one is doing when formalizing category-thoretic notions in set theory.

> Again, the point that in practice category theory is not completely emancipated from set theory is fair enough. What should be borne in mind, though, is how innocuous this residue of set theory in category theory actually is in practice. The strongly typed nature of category theory means that one rarely — one is tempted to say 'never' — stumbles over these size issues; they serve more as a form of type-checking than as a substantial topic in their own right. Moreover, category theoretic arguments typically work generically in relation to size; thus in practice, one argument fits all cases, despite the stratification.

All this is to say that, while category theory is not completely disentangled from set theory, it is quite misleading to see this as the main issue in considering the philosophical significance of categories. The temptation to do so comes from the foundationalist attitude expressed in (2) above.

The form of categorical structuralism sketched by Awodey in [7] stands in contrast to this set-theoretic foundationalism. It is a much better representation of mathematical practice, and it directs attention towards the kind of issues we have been discussing, and away from the well-worn tracks of traditional thought in the philosophy of mathematics, which after more than a century have surely reached, and passed, the point of diminishing returns.

2.2.1 Categories and Logic

Our brief introduction to category theory did not reach the rich and deep connections which exist between category theory and logic. Categorical logic is a well-developed area, with several different branches. The most prominent of these is *topos theory*.

Topos theory is an enormous field in its own right, now magisterially presented in Peter Johnstone's *magnum opus* [16]. Because, among other things, it provides a categorical formulation of a form of set theory, it is often seen as the main or even the only part of category theory relevant to philosophy. Topos theory is seen as an alternative or rival to standard versions of set theory, and the relevance of category theory to the foundations of mathematics is judged in these terms.

There are many things within topos theory of great conceptual interest; but topos theory is far from covering all of categorical logic, let alone all of category theory. From our perspective, there is a great deal of 'logic' in the elementary parts of category theory which we have diacussed. The

overemphasis on topos theory in this context arises from the wish to understand the novel perspectives of category theory in terms of the traditional concepts of logic and set theory. This impulse is understandable, but misguided. As we have already argued, learning to look at mathematics from a category-theoretic viewpoint is a real and deep-seated paradigm shift. It is only by embracing it that we will reap the full benefits.

Thus while we heartily recommend learning about topos theory, this should build on having already absorbed the lessons to be learnt from category theory in general, and with the awareness that there are other important connections between category theory and logic, in particular *categorical proof theory* and *type theory*.

2.2.2 Applications of Category Theory

As we have argued, category theory has a great deal of intrinsic conceptual interest. Beyond this, it offers great potential for applications in formal philosophy, as a powerful and versatile tool for building theories. The best evidence for this comes from Theoretical Computer Science, which has seen an extensive development of applications of category theory over the past four decades.

Some of the main areas where category theory has been applied in Computer Science include:

- **Semantics of Computation**. Denotational semantics of programming languages makes extensive use of categories. In particular, categories of domains have been widely studied [28, 5]. An important topic has been the study of *recursive domain equations* such as

  $$D \cong [D \to D]$$

  which is a space isomorphic to its own function space. Such spaces do not arise in ordinary mathematics, but are just what is needed to provide models for the type-free λ-calculus [9], in which one has self-application, leading to expressions such as the $\mathbf{Y}$ combinator

  $$\lambda f.(\lambda x.f(xx))(\lambda x.f(xx))$$

  which produces fixpoints from arbitrary terms: $\mathbf{Y}M = M(\mathbf{Y}M)$.

  The solution of such domain equations is expressed in terms of fixpoints of functors:

  $$FX \cong X.$$

This approach to the consistent interpretation of a large class of recursive data types has proved very powerful and expressive, in allowing a wide range of reflexive and recursive process behaviours to be modelled.

Another form of categorical structure which has proved very useful in articulating the semantic structure of programs are monads. Various 'notions of computation' can be encapsulated as monads [25]. This has proved a fruitful idea, not only in semantics, but also in the development of functional programming languages.

- **Type Theories**. An important point of contact between category theory and logic is in the realm of proof theory and type theory. Logical systems can be represented as categories in which formulas are objects, proofs are arrows, and equality of arrows reflects equality of proofs [19]. There are deep connections between cut-elimination in proof systems, and coherence theorems in category theory. Moreover, this paradigm extends to type theories of various kinds, which have played an important rôle in computer science as core calculi for programming languages, and as the basis for automated proof systems.

- **Coalgebra**. Over the past couple of decades, a very lively research area has developed in the field of *coalgebra*. In particular, 'universal coalgebra' has been quite extensively developed as a very attractive theory of systems [27]. This entire area is a good witness to the possibilities afforded by categorical thinking. The idea of an algebra as a set equipped with some operations is familiar, and readily generalizes to the usual setting for universal algebra. Category theory allows us to *dualize* the usual discussion of algebras to obtain a very general notion of *coalgebras of an endofunctor*. Coalgebras open up a new and quite unexpected territory, and provides an effective abstraction and mathematical theory for a central class of computational phenomena:

  - Programming over *infinite data structures*, such as streams, lazy lists, infinite trees, etc.
  - A novel notion of *coinduction*
  - Modelling *state-based computations* of all kinds
  - A general notion of *observation equivalence* between processes.

– A general form of *coalgebraic logic*, which can be seen as a wide-ranging generalization of modal logic.

In fact, coalgebra provides the basis for a very expressive and flexible theory of discrete, state-based dynamical systems, which seem ripe for much wider application than has been considered thus far; for a recent application to the representation of physical systems, see [2].

- **Monoidal Categories**.

  Monoidal categories impart a geometrical flavour to category theory. They have a beautiful description in terms of 'string diagrams' [29], which allows equational proofs to be carried out in a visually compelling way. There are precise correspondences between free monoidal categories of various kinds, and constructions of braids, tangles, links, and other basic structures in knot theory and low-dimensional topology. Monoidal categories are also the appropriate general setting for the discussion of multilinear algebra, and, as has recently been shown, for much of the basic apparatus of quantum mechanics and quantum information: tensor products, traces, kets, bras and scalars, map-state duality, Bell states, teleportation and more [3, 4]. There are also deep links to linear logic and other substructural, 'resource-sensitive' logics, and to diagrammatic representations of proofs. For a paper showing links between all these topics, see [1]. Monoidal categories are used in the modelling of concurrent processes [24], and are beginning to be employed in 'computational systems biology' [17].

  Altogether, the development of structures based on monoidal categories, and their use in modelling a wide range of computational, physical, and even biological phenomena, is one of the liveliest areas in current logically and semantically oriented Theoretical Computer Science.

  It is interesting to compare and contrast the two rich realms of monoidal categories and the structures built upon them, on the one hand; and topos theory, on the other. One might say: the *linear* world, and the *cartesian* world. It is still not clear how these two worlds should be related. A clearer understanding of the mathematical and structural issues here may shed light on difficult questions such as the relation of quantum and classical in physics.

Having surveyed some of the ways in which category theory has been

used within Computer Science, we shall now consider some of the features and qualities of category theory which have made it particularly suitable for these applications, and which may suggest a wider range of possible applications within the scope of formal philosophy.

Modelling at the right level of abstraction As we have discussed, category theory goes beyond coding to extract the essential features of concepts in terms of universal characterizations, which are then uniquely specified up to isomorphism. This is not just aesthetically pleasing; as experience in Computer Science has shown, working at the right level of abstraction is *essential* if large and complex systems are to be described and reasoned about in a managable fashion. Formal philosophy will benefit enormously by learning this lesson — among others! — from Computer Science.

Compositionality Another deep lesson to be learned from Computer Science is the importance of *compositionality*, in the general sense of a form of description of complex systems in terms of their parts. This notion originates in logic, but has been greatly widened in scope and applicability in its use in computer science.

The traditional approach to systems modelling in the sciences has been *monolithic*; one considers a whole system, models it with a system of differential equations or some other formalism, and then analyzes the model.

In the compositional approach, one starts with a fixed set of basic, simple building blocks, and *constructions* for building new, more complex systems out of given sub-systems, and builds up the required complex system with these. This typically leads to some form of algebraic description of complex systems:

$$S \;=\; \omega(S_1, \ldots, S_n)$$

where ω is an operation corresponding to one of the system-building constructions.

In order to understand the logical properties of such a system, one can develop a matching compositional view:

$$\frac{S_1 \models \phi_1, \ldots, S_n \models \phi_n}{\omega(S_1, \ldots, S_n) \models \phi}$$

One searches for a rule that will allow one to reduce the verification of a property of a complex system to verifications of suitable sub-properties for its components.

The compositional methods for description and analysis of systems which have been developed in Computer Science are ripe for application in a much wider range of scientific contexts — and in formal philosophy.

Mappings between representations Another familiar theme in Computer Science is the need for multiple levels of abstraction in describing and analyzing complex systems, and for mappings between them. Functorial methods provide the most general and powerful basis for such mappings. Particular cases, such as Galois connections, which specialize the categorical notion of adjoint functors to posets, are widely used in abstract interpretation [13].

Normative criteria for definitions As we have already remarked on a couple of occasions, category theory has a strong normative force. If we devise a mapping from one kind of structure to another, category theory tells us that we should demand that it maps morphisms as well as objects, and that it should be *functorial*. Similarly, if we devise some kind of product for a certain type of structures, category theory tells us which properties our construction should satisfy to indeed be a product in the corresponding category. More generally, constructions, if they are 'canonical', should satisfy a suitable universal property; and if they do, then they are unique up to isomorphism. There are other important criteria too, such as *naturality* (which we have not discussed).

These demands and criteria to be satisfied should be seen as providing valuable *guidance*, as we seek to develop a suitable theory to capture some phenomenon. If we have no such guidance, it is all too likely that we may make various ad hoc definitions, not really knowing what we are doing. As it is, once we have specified a category, there are an enormous range of well-posed questions about its structure which we can ask. Does the category have products? Other kinds of limits and colimits? Is it cartesian closed? Is it a topos? And so on. By the time we have answered these questions, we will already know a great deal about the structure of the category, and what we can do with it. We can also then focus on the more distinctive features of the category, which may in turn lead to a characterization of it, or perhaps to a classification of categories of that kind.

2.3 Logic and Category Theory as Tools for Building Theories

The project of scientific or formal philosophy, which seems to be gathering new energy in recent times, can surely benefit from the methods and tools offered by Category theory. Indeed, it can surely not afford to neglect them. Logic has been used as the work-horse of formal philosophy for many years, but the limitations of logic as traditionally conceived become apparent as soon as one takes a wider view of the intellectual landscape. In particular, Computer Science has led the way in finding new ways of applying logic — and new forms of logic and structural mathematics which can be fruitfully applied.

Philosophers and foundational thinkers who are willing and able to grasp these opportunities will find a rich realm of possibilities opening up before them. Perhaps this brief essay, modest in scope as it is, will point someone along this road. If so, the author will feel handsomely rewarded.

3 Guide to Further Reading

The lecture notes [6] are a natural follow-up to this article.

The short book [26] is nicely written and gently paced. A very clear, thorough, and essentially self-contained introduction to basic category theory is given in [8].

Another very nicely written text, focussing on the connections between categories and logic, and especially topos theory, is [15], recently reissued by Dover Books. The book [23] is pitched at an elementary level, but offers insights by one of the key contributors to category theory.

The text [20] is a classic by one of the founders of category theory. It assumes considerable background knowledge of mathematics to fully appreciate its wide-ranging examples, but it provides invaluable coverage of the key topics. The 3-volume handbook [12] provides coverage of a broad range of topics in category theory.

A classic text on categorical logic and type theory is [19]. A more advanced text on topos theory is [21]; while [16] is a comprehensive treatise, of which Volume 3 is still to appear.

References

[1] S. Abramsky. Temperley-Lieb algebra: From knot theory to logic and computation via quantum mechanics. In Goong Chen, Louis Kauff-

man, and Sam Lomonaco, editors, *Mathematics of Quantum Computing and Technology*, pages 415–458. Taylor and Francis, 2007.

[2] S. Abramsky. Coalgebras, Chu spaces, and representations of physical systems. In J.-P. Jouannaud, editor, *Proceedings of the 25th Annual IEEE Symposium on Logic in Computer Science: LiCS 2010*. IEEE Computer Society, 2010. To appear.

[3] S. Abramsky and B. Coecke. A categorical semantics of quantum protocols. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science: LiCS 2004*, pages 415–425. IEEE Computer Society, 2004.

[4] S. Abramsky and B. Coecke. Categorical quantum mechanics. In K. Engesser, D. Gabbay, and D. Lehmann, editors, *Handbook of Quantum Logic and Quantum Structures: Quantum Logic*, pages 261–324. Elsevier, 2009.

[5] S. Abramsky and A. Jung. Domain theory. In S. Abramsky, D. Gabbay, and T. S. E. Maibaum, editors, *Handbook of Logic in Computer Science*, pages 1–168. Oxford University Press, 1994.

[6] S. Abramsky and N. Tzevelekos. Introduction to category theory and categorical logic. In B. Coecke, editor, *New Structures for Physics*, Lecture Notes in Physics. Springer-Verlag, 2010. To appear.

[7] S. Awodey. An answer to G. Hellman's question "Does category theory provide a framework for mathematical structuralism?". *Philosophia Mathematica*, 12:54–64, 2004.

[8] S. Awodey. *Category theory*. Oxford University Press, 2010.

[9] H. P. Barendregt. *The Lambda Calculus*, volume 103 of *Studies in Logic and Foundations of Mathematics*. North-Holland, 1984.

[10] Paul Benacerraf. What numbers could not be. *The Philosophical Review*, 74:47–73, 1965.

[11] Andreas Blass. The interaction between category theory and set theory. In J. Gray, editor, *Mathematical Applications of Category Theory*, volume 30 of *Contemporary Mathematics*, pages 5–29. AMS, 1984.

[12] F. Borceux. *Handbook of Categorical Algebra Volumes 1–3*. Cambridge University Press, 1994.

[13] Patrick Cousot and Radhia Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *POPL*, pages 238–252, 1977.

[14] P. J. Freyd. Aspects of topoi. *Bulletin of the Australian Mathematical Society*, 7(1):1–76, 1972.

[15] R. I. Goldblatt. *Topoi, the Categorial Analysis of Logic*. North-Holland, 1984. Reprinted by Dover Books, 2006.

[16] P. T. Johnstone. *Sketches of an Elephant: A Topos Theory Compendium. I, II*. Oxford University Press, 2002.

[17] Jean Krivine, Robin Milner, and Angelo Troina. Stochastic bigraphs. In *Proceedings of MFPS XXIV: Mathematical Foundations of Programming Semantics*, volume 218 of *ENTCS*, page 7396, 2008.

[18] Kenneth Kunen. *Foundations of Mathematics*. College Publications, 2009.

[19] J. Lambek and P. J. Scott. *Introduction to higher-order categorical logic*. Cambridge University Press, 1986.

[20] S. Mac Lane. *Categories for the Working Mathematician, Second Edition*. Springer, 1998.

[21] S. Mac Lane and I. Moerdijk. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Springer, 1994.

[22] F. W. Lawvere. Adjointness in foundations. *Dialectica*, 23:281–296, 1969.

[23] F. W. Lawvere and S. Schanuel. *Conceptual Mathematics: A First Introduction to Categories*. Cambridge University Press, 1997.

[24] Robin Milner. *The Space and Motion of Communicating Agents*. Cambridge University Press, 2009.

[25] Eugenio Moggi. Notions of computation and monads. *Inf. Comput.*, 93(1):55–92, 1991.

[26] Benjamin C. Pierce. *Basic Category Theory for Computer Scientists*. MIT Press, 1991.

[27] Jan J. M. M. Rutten. Universal coalgebra: a theory of systems. *Theor. Comput. Sci.*, 249(1):3–80, 2000.

[28] D. S. Scott. Outline of a mathematical theory of computation. Technical report, Oxford University Computing Laboratory, 1970. Technical Monograph PRG-2 OUCL.

[29] Peter Selinger. A survey of graphical languages for monoidal categories. In B. Coecke, editor, *New Structures for Physics*, Lecture Notes in Physics. Springer-Verlag, 2010. To appear.

Memory and Logic: a Tale from Automata Theory

R. RAMANUJAM*

1 Infrastructure of reasoning

> *Please assume ... that there is in our souls a block of wax, in one case larger, in another smaller, in one case the wax is purer, in another more impure and harder, in some cases softer, and in some of proper quality ... Let us, then, say that this is the gift of* **Memory***, the mother of the* Muses*, and that whenever we wish to remember anything we see or hear or think of in our own minds, we hold this wax under the perceptions and thoughts and imprint them upon it, just as we make impressions from seal rings; and whatever is imprinted we remember and know as long as its image lasts, but whatever is rubbed out or cannot be imprinted we forget and do not know.*
>
> (Socrates to Theaetetus. Plato, Theaetetus 191d).

I once had the pleasure of reading this quote to some school children. I was discussing memory in the context of playing games, and the quote was to set the context. One child in the audience stood up to ask me: "If this is right, we would need a huge block of wax inside us, how can we carry it?" Everyone laughed, but she was a true philosopher and had grasped one important problem underlying reasoning that we often tend to overlook. In fact, the *theory of computation* is her natural home, where such questions are *de rigueur*.

Logicians typically ask, what do we need to know about something to reason about it? A supplementary question is lurking here: how much memory is implied (for acquiring and maintaining such knowledge)? For an all-powerful being, all reasoning may be instanteneous. But the compositional language of logic carries with it a notion of a being who assembles

*The Institute of Mathematical Sciences, Chennai, India - 600 113

together pieces of propositional knowledge to construct epistemic edifices by processes of inference. This is a moving image and uses memory for retrieval of pieces encountered earlier to combine with pieces constructed later. So one might well ask: what are the *mechanisms* that underlie the process of reasoning, that the reasoner needs to have access to? In other words, what is the infrastructure of reasoning?

Rather interestingly, such were the considerations that led Alan Turing to devise the mathematical model of a digital computer, which answered an important question raised by David Hilbert: what is the notion of an effective procedure? If we were to talk of the existence of procedures to solve mathematical problems, their construction, or show that such procedures cannot exist, we need to formalize the notion of procedure. In a very readable and amusing account, Turing takes this head-on: he envisages a mathematician sitting with pen and paper, writing symbols on paper and performing calculations. He takes questions of how much ink and paper would be required very seriously indeed. The model he goes on to define, what we now call **Turing machines**, has turned out to be robust: over the last century, a consensus (based on strong mathematical argument) has emerged, and 'solvable' has come to mean 'computable' by a Turing machine. (See [Dav06] for a lucid account of this history.)

Theory of computation offers the paradigm of **computational complexity** to study these 'infrastructural' questions, so that we can sensibly ask how much memory is needed to solve problems, and sometimes even find good answers to such questions. This is done by postulating the problem solver to be a Turing machine which processes instances of the problem and produces answers, and studying how much memory such a machine would require, expressed as a function of the input description size. This is studied under the rubric of space complexity, and solvable problems are classified as being in logarithmic space, polynomial space, exponential space, and so on (where the terms refer to the function that reflects the growth of memory requirement on input size). Why this is a good idea may not be clear, but since that would take us too far afield, we refer the reader to texts on theory of computation and complexity such as [HU79], [Koz97] and [Sip05], and move on. Complexity classes of time and space and relationships between them occupy a central place in computation theory.

In the context of logic, we could translate the discussion above to ask: *how much memory is needed for solving a problem expressed in a (specified) formal logical language?* Finding a model for the formula (if it is satisfiable) would amount to solving the problem, so we consider a Turing machine that takes a formula as input and after some finite time, outputs

a 'Yes' (and perhaps a model for the formula) or a 'No' to assert that the formula is not satisfiable. One then addresses the question of how much memory would be necessary and sufficient for the hypothesized Turing machine. This is the well-known question of decision procedures for logics and their complexity. But there are more questions of this nature, leading to logical foundations of complexity theory, and in relation to computational logics.

This is only one of many questions related to memory infrastructure for reasoners. A natural one, and closer to the spirit of the preceding discussion on reasoning, would be: *how much memory is needed for proving a property in a formal deduction system?* This is related to the notion of proof complexity which attempts to measure how hard it is to prove a property in a formal system. Consider an arithmetical property such as $\varphi(x) = \exists y : y * y = x$, where $*$ is the multipication operation. The formula asserts that x is a perfect square. A proof of $\varphi(1024)$ is also a computation of the square-root of 1024, but the formula itself offers no clue as to how hard such a computation would be, even within a specified formal system. [Par71] offered an early account, distinguishing existence assertions from the *feasibility* of proving them in a fragment of Peano arithmetic. Subsequently, bounded arithmetic has provided many interesting insights of this kind.

Proof complexity arises from the view of logic as a system of deductive reasoning. Another natural question relates to the view of logic as a language. *How hard is it to describe a structure in a (specified) logical language?* Termed descriptive complexity this is the study of expressiveness of logical languages. Consider a fixed formula (sentence) in a logical language and its "input" to be a structure over which the sentence can be interpreted. The formula then acts as a machine that takes structures as input and answers 'True' or 'False'. We can then ask, how much memory does this machine require? Note that this line of thought is very fruitful: the next step is to formalize notions of complexity in logic, and thus attempt to capture complexity classes by logical means. Indeed, it is then natural to go further and ask meta-theoretical questions: what is the complexity of reasoning about complexity theoretic statements?

These are fertile areas of formal study, with rich notions and theorems. However, formal connections between notions of hardness arising from these different viewpoints have not been clearly established (yet) and form a subject of very interesting current research. We refer the interested reader to [ABSRW02] and [Kol08] for a fascinating account; we will not take up this theme here.

There is yet another, and more direct, question relating memory and logic, and that is the theme we will pursue here. Consider any logic for reasoning about processes that occur in time. Examples are tense logics and temporal logics, but also dynamic logics, epistemic logics, process logics, game logics and logics of action and agency. These are patterns of reasoning that naturally involve phrases such as 'before' and 'after', past and future, always and never. The infrastructural concern is very relevant in these contexts: what does the reasoning entail, in terms of memory needed to support such reasoning? Plato's wax is just what we need: form impressions of events when they occur and recall events by referring to the traces.

Already, we see that there is more to be said than speak only of how much memory is needed. Rather, we can consider *what* we need to remember (and hence what we may afford to forget), how such memory can be *structured* or organized (for efficient retrieval), and so on. These memory mechanisms can well be thought of as infrastructure underlying logics for reasoning about processes.

We need some language for describing such mechanisms, and theory of computation provides such a language. This is a programme that can be taken up in general, but since this aims to be a presentation at an elementary level, we will confine ourselves to **finite memory** mechanisms and logics that describe seqential behaviour (into which many of the process logics listed above can be translated). We present two formalisms: one of first order and monadic second order logics over sequences; the other, of finite state automata that encapsulate memory mechanisms. We then see that these are two ways of looking at the same picture, in the sense that the formalisms are equivalent in expressive power, with the proof giving us some insight into what we have referred to as the memory infrastructure underlying the logics. We discuss how this approach can be generalized and extended in many directions.

Why should a philosopher and logician interested in inference bother about automata models at all? While inference can be discussed in descriptive terms, it is clear that there are underlying procedures which together make up inference, and theory of computation provides a formal basis for studying such procedures. However, such procedures make infrastuctural assumptions which need logical analyses, and this interplay between logic and procedures (computations) is complex. One way of understanding this is to ask what the addition of a new connective to a logic implies: in terms of how existing underlying procedures need to be modified, and what new procedures need to be created. The answer to this is complex and con-

sidered to be difficult for even very simple logics, which shows that this interaction is little understood as yet.

Another important dimension is the study of inference by "real" agents, who have limited capacity to observe, to recall facts, to decide which inference mechanism to apply when, and to pursue assumptions to their conclusions. Many of these limitations can be easily modelled as memory limitations which is what automata models achieve. Whatever an agent knows, if the agent is to recall and make use of such knowledge at different points in time, needs to be remembered, and hence part of the agent's memory. Explicit modelling of such memory can then incorporate agency more directly.

However, the account developed in this article also shows us roads untravelled yet: these automata models of memory mechanisms incorporate agency at a "hardware level" in the design of the machines, as it were. When we take agency more seriously, and study the interaction of memory and epistemic attitudes, further structuring of memory (in terms of selection and retention) becomes important, and this needs more sophisticated modelling. When we consider systems of many reasoners, the situation is much more interesting: in such systems, individual memory is traded off against communication, and hence notions like Halbwachs' *collective memory* ([Hal80]) become important. The fact that memory is distributed across agents, and that it evolves with time is well known to us. When inference systems treat memory as being monolithic and immediately available as a whole, they miss the influence of such temporal and distributed structure which, in turn, limits the reasoning being studied. Distributed automata models can be seen as initial attempts in this direction but we are far from the sort of logical analyses discussed here.

2 Finite state automata

The memory mechanism that we wish to describe is very simple: it has some fixed number of fixed size registers. This means that the set of values that each register can hold is finite and fixed *a priori*. For instance, a boolean variable is such a register, its value can be 0 or 1. Date (within this century) would be one such register. An integer valued variable cannot be a register since its value space is infinite.

Formally, a finite memory is a tuple $M = (< R_1, D_1 >, \ldots, < R_m, D_m >)$, where, for $j \in \{1, \ldots, m\}$, R_j is a register that can hold a value in the finite set D_j. A configuration of M is an m-tuple $(d_1, \ldots, d_m)$, where, for

$j \in \{1, \ldots, m\}$, $d_j \in D_j$ is the current value of register R_j.

How would any mechanism use such a memory? Given any configuration, the mechanism would perform an *action* which causes a potential change in the configuration. The action may be reading a register, performing some computation, updating a register, interact with some external agent and based on the interaction, store some value in the register, etc. Since we can use as many registers as we wish, all information that the mechanism needs (as long as it is finite) can be coded up into registers. Thus an action is in general of the form (e, χ, μ) where e is some event occurrence, χ is some condition that is checked to hold of the current configuration, and μ is an instruction to update one or more of the registers. A process would then be simply a finite sequence of such actions, or more generally, a (finite or infinite) set of action sequences.

We can design a suitable programming language in which we can express such event stimuli, interactions, tests on register values, instructions to update registers etc. Such a language would have variables, declarations of types of variables (to specify range of values). It can have conditional branching and looping. What it would not have are variables that can take infinitely many values.

How would we present the *semantics* of such processes? We would need to specify the initial memory configuration of the mechanism, describe which action is enabled in any configuration, and define how each action may modify any given configuration. A process can then be presented as a set of sequences of configuration changes. Any particular historical trajectory would be given by the trace which is the sequence of configurations. The objective or goal of the process can then be given by labelling sequences as good (those that satisfy the objective) or bad (those that don't).

It should be clear that however informal this description may be, it can be quite easily formalized. Further we can also see that the notion of process is general enough to cover a wide range of processes that we wish to reason about in propositional modal logics, at least in terms of what we might need to remember.

One benefit of such an analysis, however superficial, is that we immediately see that there is no need at all to design a mechanism language as above and work out its detailed semantics. A very simple abstraction suffices.

For M, let C_M denote the space of *all possible* configurations. It is clear that C_M is finite, and that the cardinality of C_M is given by: $|C_M| = |D_1| \times \ldots |D_m|$. Let $|C_M| = K$. We might as well enumerate C_M as $\{c_1, c_2, \ldots c_K\}$. Similarly, as far as the actions are concerned, whatever the actions may

be, we are only concerned with how they change configurations. Clearly there can be only finitely many such changes that we can describe (there are at most K^K functions from C_M to C_M), we can work with a finite set of event types E and specify, for each $e \in E$, a function $\delta_e : C_M \to C_M$. An initial configuration, and a set of *final* configurations, declaring whether the desired "objective" has been attained, and we are done.

We mentioned that the semantics of a process can be given by a set of action sequences. Since actions are abstracted into event types above, we would have a set of sequences of event occurrences, those that start from initial configurations and end in one of the acceptable final configurations. In general, behaviours are given by sets of finite words (sequences of letters) called *languages*.

The theory of computation formalizes these intuitive suggestions in the well-known model of finite state automata.

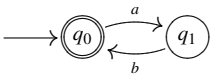
2.1 Automata and languages

Let E be a fixed finite set of event types. We use letters a, b, e etc (with or without subscripts) to denote elements of E. By E^* we denote the set of all finite sequences over E, and ϵ to denote the null sequence in E^*. We use letters u, v, w etc (with or wihout subscripts) to denote sequences over E (called words). $|w|$ denotes the length of the sequence w. We call $L \subseteq E^*$ a language over E. We use L, L' etc for languages.

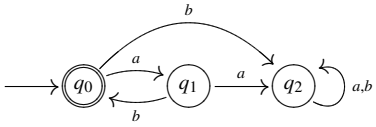
Suppose $u = a_1 a_2 \ldots a_m$ and $v = b_1 b_2 \ldots b_n$ be sequences. By uv we denote the concatenated sequence $w = c_1 c_2 \ldots c_k$, where $k = m + n$, and $c_i = a_i$, for $1 \leq i \leq m$, and $c_{j+m} = b_j$ for $1 \leq j \leq n$. Note that $w\epsilon = \epsilon w = w$, for all $w \in E^*$. We can extend this to define concatenation of languages by $L_1 L_2 = \{uv \mid u \in L_1, v \in L_2\}$. By u^n we mean the n-fold concatenation of u. Given $u \in E^*$ and $L \subseteq E^*$, we also define $u \cdot L = \{uv \mid v \in L\}$.

Let $E = \{a, b, c\}$ and $u = abbaccb$. We can then talk about the first occurrence of a, the k^{th} occurrence of b, the lack of occurrences of c in u and so on. Note that when we talk of a property of a sequence over E, we implicitly define a subset of E^* (and thus a language). For instance, consider the property that every occurrence of b is preceded by some earlier occurrence of a. u above satisfies the property but $v = cbaaccb$ does not. This defines the following set: $L_0 = \{a, c\}^* \cup \{uavbw \mid u \in \{c\}^*, v \in \{a, c\}^*, w \in E^*\}$. Note that we can alternatively define L_0 by: $E^* - \{ubv \mid u \in \{c\}^*, v \in E^*\}$.

The rationale for invoking such a property is: a might be the signal that enables the occurrences of b's thereafter. This formalism can be used to express many temporal and causal relationships between event occurrences.



(a) Automaton for $(ab)^n$, $n \geq 0$



(b) Deterministic automaton for $(ab)^n$, $n \geq 0$

Figure 1:

A finite state automaton over E is a tuple $A = (Q, \delta, I, F)$, where Q is a finite set of states, $I, F \subseteq Q$ are the sets of initial and accepting or final states, respectively, and $\delta \subseteq (Q \times E \times Q)$ is the transition relation.

A is said to be deterministic if I is a singleton set, and δ is the graph of a function $(Q \times E) \to Q$.

We say that $a \in E$ is *enabled* at $q \in Q$ if there exists $q' \in Q$ such that $(q, a, q') \in \delta$ (for which we usually write $q \xrightarrow{a} q'$).

We present finite state automata in graphical form as below. States are represented by circles, and transitions by edges. Initial states are marked by incoming arrows without source, and final states by two concentric circles. Consider the two automata on $E = \{a, b\}$ pictured in Figure 1 below. In the one in Figure 1(a), b is not enabled at q_1 and a is not enabled at q_2. The automaton in Figure 1(b) is deterministic.

Let $w = e_1 e_2 \ldots e_k \in E^*$. A run of A on w from $q_0 \in Q$ is a sequence $\rho = q_0 q_1 \ldots q_k$, where for all $i : 1 \leq i \leq k$, we have: $(q_{i-1}, a_i, q_i) \in \delta$. We say that ρ is accepting and that A accepts w if $q_0 \in I$ and $q_k \in F$. We define the language accepted by A to be the set $L(A) = \{w \in E^* \mid$ there exists an accepting run of A on $w\}$. Note that a deterministic automaton has a unique run on every word in E^*.

Let $L \subseteq E^*$. We say that L is **recognizable** if there exists a finite state automaton A such that $L(A) = L$. The class of recognizable languages over E is denoted Rec_E.

We can check that both automata in Figure 1 accept the same language: $\{(ab)^n \mid n \geq 0\}$. Now consider the language L_0 above, the set of all sequences in which every occurrence of b is preceded by some earlier occurrence of a. It is easily seen that the automaton below in Figure 2 accepts the language L_0.

As one more example, consider the language L_{even} over $E = \{a, b\}$ con-

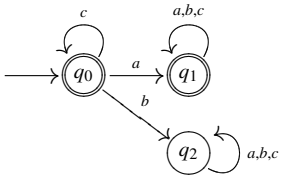


Figure 2: Deterministic automaton accepting L_0

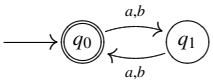


Figure 3: Automaton for L_{even}

sisting of all sequences of even length. The automaton below in Figure 3 accepts this language.

2.2 Deterministic automata

In the few examples we have considered so far, deterministic automata have been used to accept a given language. The following theorem, due to Rabin and Scott ([RS59], which in some sense initiated automata theory) asserts that this is no accident.

Theorem 2.1. *Every recognizable language is accepted by a deterministic automaton.*

Proof Suppose that $L \subseteq E^*$ is recognizable and $A = (Q, \delta, I, F)$ is an automaton such that $L(A) = L$. We now construct a deterministic automaton B such that $L(B) = L$. Define $B = (Q_B, \delta_B, I_B, F_B)$, where:

- $Q_B = 2^{Q_A}$.
- $I_B = \{I\}$.
- $F_B = \{X \subseteq Q \mid (X \cap F) \neq \emptyset|\}$.
- $\delta_B = \{(X, a, Y) \mid Y = \{q' \mid (q, a, q') \in \delta, q \in X\}\}$.

We can easily check that B is indeed deterministic. To see that $L(A) = L(B)$, consider a word $w = e_1 \dots e_k \in E^*$ and the unique run of B on it, say $\rho_B = X_0 X_1 \dots X_k$ where $X_j \xrightarrow{a_j}_B X_{j+1}$. If $w \in L(B)$, $X_0 = I$ and

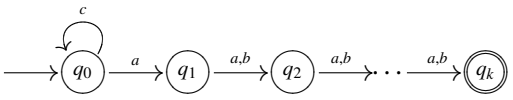


Figure 4: Automaton for "accepting k^{th} last letter is a"

$X_k \cap F \neq \emptyset$. Then we can find $q_i \in X_i$, $i \in \{0, \ldots, k\}$ such that the sequence $\rho_A = q_0 q_1 \ldots q_k$ is a run of A on w and show that it must be accepting. Conversely, given such an accepting run ρ_A of A on w, we can use it to show that ρ_B must be accepting. □

Thus every automaton is language-equivalent to a deterministic one. The construction above causes an exponential blow-up, and it can be shown that this is unavoidable. Consider the language $\{w \in \{a,b\}^* \mid w = ue_1e_2 \ldots e_k, k > 0, e_1 = a, u \in \{a,b\}^*, e_2, \ldots e_k \in \{a,b\}\}$, that consists of words in which the k^{th} last letter is an a. This is accepted by an automaton with $k+1$ states as in Figure below, whereas one can argue that any *deterministic* automaton requires more than 2^k states.

How do we show such lower bounds on number of states? Given a language (specified somehow), how do we argue that any automaton for it must use so many states? This is important, since the number of states tells us how much memory we have in the system, to realize the behaviour specified by the language. n states corresponds to $\log n$ bits of memory (roughly speaking). While we are about it, we might as well ask: how do we know that any automaton *exists at all* which accepts the given language?

2.3 Existence of non-recognizable languages

Consider an automaton $A = (Q, \delta, I, F)$ on $E = \{a,b\}$, and consider a run of A on $w = a_1 \ldots a_k \in E^*$, say $\rho = q_0 q_1 \ldots q_k$. Now suppose that $k > |Q|$. Then, some state must repeat in this sequence: there exist $i < j$ such that $q_i = q_j$. (So the automaton has a loop from q_i to itself via $q_{i+1} \ldots q_{j-1}$.) We can write $w = uvu'$, where $u = a_1 \ldots a_i$, $v = a_{i+1} \ldots a_j$, $u' = a_{j+1} \ldots a_k$. Therefore, if w is accepted by A then so also is a strictly shorter word uu' and a strictly longer word $uvvu'$. Indeed, for any $n > 0$, $uv^n u'$ is accepted, and hence the language accepted by A is infinite.

This observation leads us to an important limitation of finite state automata. All information about the past has to be "hard-wired" into its states. Two different pasts that end up in the same state are equivalent for the automaton. When it considers words that are 'too long', it cannot tell them

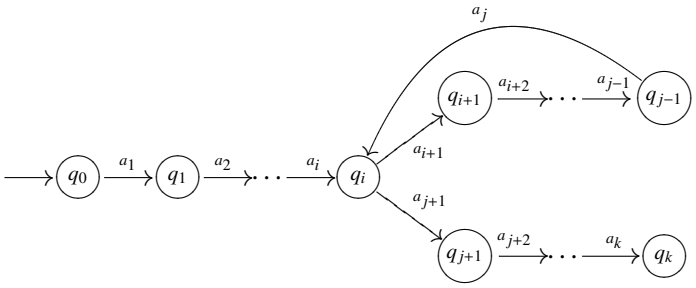


Figure 5:

apart. Such 'forgetfulness' is a critical feature of the automaton, measured by the number of its states, and gives us a handle on what it can and what it cannot do.

The remark above is formalized as follows. Let $A = (Q, \delta, q_0, F)$ be a deterministic automaton on E, and let $u, v \in E^*$. A has a unique run on each word from q_0, ending in say, q_u and q_v. Now suppose that $q_u = q_v = q$. Consider any $w \in E^*$. A has a unique run from q on w, ending in q'. Now either $q' \in F$ or not; in the former case, both uw and vw are accepted, and in the latter, both are rejected. Therefore the role of q is to simply 'tie up together' all words (pasts) after which a residue can be accepted (or not). This intuition leads us to the following definition.

Fix E and consider $u \in E^*, L \subseteq E^*$. Define $u/L = \{v \in E^* \mid uv \in L\}$. We call u/L the residue of u in L. Note that $\epsilon/L = L$ (hence $L \in Res(L)$ always) and for any $u \in E^*$, we have: $u \in L$ iff $\epsilon \in u/L$.

Define $Res(L) = \{u/L \mid u \in E^*\}$. The cardinality of $Res(L)$ is called the rank of L. We call $L \subseteq E^*$ a regular language when $Res(L)$ is finite.

Theorem 2.2. *L is recognizable iff L is regular. Moreover if A is any deterministic automaton with m states accepting L then rank$(L) \leq m$.*

Before we go on to proving the theorem (due to Myhill and Nerode), consider what it says. Finite state automata accept precisely the finite residue languages, and the rank of a recognizable language is the number of states of a minimal deterministic automaton accepting it. Thus residues (defined for any language, independent of machines) correspond abstractly to states of the most economical constructions possible.

The implications are clear. To show that a given language L is not recognizable, we only need to show that it has infinitely many residues. To

prove a lower bound on number of states required for a recognizable language, we only need to compute its rank. Note an immediate consequence of the theorem: every finite language is recognizable.

As an example consider the language $L = \{a^n b^n \mid n \geq 0\}$ over $\{a, b\}$. Let $i > 0, j > 0, i \neq j$. Then $a^i/L = \{b^i\} \neq \{b^j\} = a^j/L$. Thus $\{a^i/L \mid i > 0\}$ is infinite and hence $Res(L)$ is infinite. By the theorem, we cannot find any automaton that accepts this language. As an exercise, consider the language that consists of words over $\{a, b\}$ which have the same number of a's as that of b's (but in any order), and show that it is not recognizable.

Recall the language over $\{a, b\}$ with words where the k^{th} last letter is an a. Showing that its rank is exponential in k suffices to establish the lower bound we were after.

The proof of the Theorem proceeds by constructing an automaton in case $Res(L)$ is finite. For the converse, when L is recognizable, the proof shows that any automaton for L has fewer than $rank(L)$ many states, thus establishing the converse as well as the stronger statement.

Proof Let $L \subseteq E^*$ and suppose that $Res(L)$ is finite. Define $A = (Q, \delta, q_0, F)$, where $Q = Res(L)$, $q_0 = L$, $F = \{X \in Res(L) \mid \epsilon \in X\}$ and $\delta : (Q \times E) \to Q$ is defined by: $\delta(X, a) = Y$ if there exists $u \in E^*$ such that $X = u/L$ and $Y = (ua)/L$.

To show that δ is well-defined, we need to prove that for any $u, v \in E^*$, if $u/L = v/L$ then $(ua)/L = (va)/L$. But this is easy: if $w \in (ua)/L$, then $uaw \in L$, so $aw \in u/L$, hence $aw \in v/L$ so $w \in (va)/L$. Thus A is indeed a deterministic automaton. To show that $L(A) = L$, let $u = a_1 \ldots a_k \in E^*$. The unique run of A is of the form

$$q_0 = L = \epsilon/L \;\; a_1/L \;\; (a_1a_2)/L \ldots (a_1 \ldots a_k)/L = u/L.$$

u is accepted by A iff $u/L \in F$ iff $\epsilon \in u/L$ iff $u \in L$, as required.

For the converse direction, assume that L is recognizable and that $A = (Q, \delta, q_0, F)$ accepts L. For $q, q' \in Q$ and $u \in E^*$, define $\widehat{\delta}(q, u) = q'$ when q is the last state of the unique run of A on u from q. With no loss of generality, we can assume that every $q \in Q$ is reachable from q_0, that is, for some $u \in E^*$, $\widehat{\delta}(q_0, u) = q$. (Otherwise, simply remove all the unreachable states and consider the automaton with remaining states.)

Now consider the map $f : Q \to Res(L)$ defined by $f(q) = u/L$, such that $u \in E^*$ and $\widehat{\delta}(q_0, u) = q$. We need to show that f is well-defined. Let $u, v \in E^*$ such that $\widehat{\delta}(q_0, u) = \widehat{\delta}(q_0, v) = q$. Then we need to prove that $u/L = v/L$. Let $w \in u/L$, then $uw \in L = L(A)$. Hence $\widehat{\delta}(q, w) \in F$. But then $vw \in L(A) = L$ as well, and hence $w \in v/L$ as well.

Now since A is deterministic, for every $u \in E^*$, $\widehat{\delta}(q_0, u)$ is defined and hence f is *onto*. Thus we have a surjective map from a finite set Q to $Res(L)$, which means that $|Res(L)| \leq |Q|$, proving the theorem. □

2.4 Programs and machines

We have shown that the two notions on languages, that of recognizability (based on automaton mechanisms) and regularity (based on behavioural notions) coincide. We have relied on informal specifications of languages and discussed their recognizability. In the next section, we seek a *logical notation* in which such languages can be described. As a preparation, we note some properties of regular languages and describe a different notation here.

Proposition 2.3. *Recognizable languages are closed under the boolean operations.*

Proof Let $A = (Q, \delta, q_0, F)$ be a deterministic automaton over E. Consider $A' = (Q, \delta, q_0, Q - F)$. It is easily seen from the definitions that $L(A') = E^* - L(A)$. From the Rabin - Scott theorem it follows that recognizable languages are closed under complementation.

Let $A_1 = (Q_1, \delta_1, I_1, F_1)$ and $A_2 = (Q_2, \delta_2, I_2, F_2)$ be automata over E. Without loss of generality assume that Q_1 and Q_2 are disjoint. (Otherwise, rename states.) Then define $A = (Q_1 \cup Q_2, \delta_1 \cup \delta_2, I_1 \cup I_2, F_1 \cup F_2)$ by point-wise union. It is easily seen that $L(A) = L(A_1) \cup L(A_2)$. Thus recognizable languages are closed under union (and hence, from above, under intersection as well). □

Suppose that $E = E_1 \times E_2$ and let $u \in E^*$. Define the projection operation on E_1 as follows: let $u = (a_1, b_1) \ldots (a_k, b_k) \in E^*$; then $u\lceil E_1 = a_1 \ldots a_k$. For $L \subseteq E^*$ define $L\lceil E_1 = \{u\lceil E_1 \mid u \in L\}$. $u\lceil E_2$ and $L\lceil E_2$ are defined in the obvious way.

Proposition 2.4. *Recognizable languages are closed under projection.*

Proof Let $A = (Q, \delta, I, F)$ be an automaton over $E = E_1 \times E_2$ accepting $L \subseteq E^*$. Define $A' = (Q, \delta', I, F)$ where $\delta' \subseteq (Q \times E_1 \times Q)$ is the erasure map: $(q, a, q') \in \delta'$ if there exists $b \in E_2$ such that $(q, (a, b), q') \in \delta$. It is then easily seen that $L(A') = L(A)\lceil E_1$. □

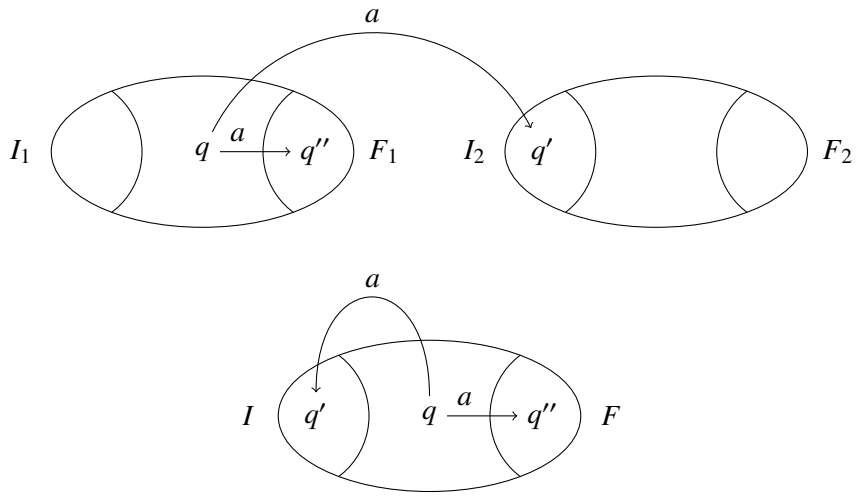
Indeed, recognizable languages are closed under many other interesting operations such as concatenation, iteration, mirror-reversal, prefixing etc (and some exotic ones such as 'halving'). For theory of computation this offers the possibility of 'programming notations' for finite state devices.

Let $L^n, n \geq 0$ be defined inductively as follows: $L^0 = \{\epsilon\}$; $L^{k+1} = LL^k$, for $k \geq 0$. Let $L^* = \bigcup_{k \geq 0} L^k$, the Kleene iteration of L.

Proposition 2.5. *Recognizable languages are closed under concatenation and Kleene iteration.*

Proof Let $A_1 = (Q_1, \delta_1, I_1, F_1)$ and $A_2 = (Q_2, \delta_2, I_2, F_2)$ be automata over E. Without loss of generality assume that Q_1 and Q_2 are disjoint. Then define $A = (Q_1 \cup Q_2, \delta, I_1, F_2)$, where $\delta = \delta_1 \cup \delta_2 \cup \{(q, a, q') \mid q' \in I_2, \exists q'' \in F_1 : (q, a, q'') \in \delta_1\}$. Then it is easily seen that $L(A) = L(A_1) \cdot L(A_2)$. Thus recognizable languages are closed under concatenation.

Let $A = (Q, \delta, I, F)$ an automaton over E and define $A' = (Q, \delta', I, I)$ where $\delta' = \delta \cup \{(q, a, q') \mid q' \in I, \exists q'' \in F : (q, a, q'') \in \delta\}$. Then it is easily seen that $L(A') = L(A)^*$, proving closure of recognizable languages under iteration. □



A celebrated theorem of Kleene asserts that these operations can indeed be used to characterize the class of recognizable languages. We state the theorem below without proof.

Theorem 2.6. *Recognizable languages over E constitute the least class of languages that contains all finite languages over E and closed under the operation of union, concatenation and Kleene iteration.*

This theorem serves as a basis for a syntactic presentation of regular languages. Consider the following syntax:

$$r \in R_E ::= \emptyset \mid a \in E \mid r_1 + r_2 \mid r_1 \cdot r_2 \mid r^*$$

The semantics of the constructs is given inductively. $\emptyset$ stands for the empty set, a for the singleton word language $\{a\}$; $+$, $\cdot$ and $*$ respectively denote union, concatenation and iteration. An expression in the syntax above is called a rational expression over E, and $L \subseteq E^*$ is said to be a rational language if there exists a rational expression r whose semantics is L.

Then Kleene's theorem can be restated as follows:

Theorem 2.7. *L is recognizable iff it is rational.*

Rational expressions are very similar to basic programming notations. Note the absence of complementation in the language of rational expressions, as in the case of programs as well. Since regular languages are closed under complementation, adding it will not change expressive power. However, it turns out that this can give a more succinct description. Consider the following syntax:

$$r \in SF_E ::= \emptyset \mid a \in E \mid r_1 + r_2 \mid r_1 \cdot r_2 \mid \ \oplus r$$

The semantics is as before, the meaning of $\oplus r$ is the complement of the meaning of r. Note that there is no iteration in this expression language. We call L star free if there exists $r \in SF_E$ such that the semantics of r is L.

E^*, for all its appearance, is star free, since it is $\oplus\emptyset$. $(ab)^*$ (over $E = \{a, b\}$) is also star free: it is the complement of $(E^* \cdot a) + (b \cdot E^*) + (E^* \cdot aa \cdot E^*) + (E^* \cdot bb \cdot E^*)$. On the other hand, $(aa)^*$, which is deceptively similar, is not star free, a fact that is by no means easy to prove and well beyond the technical means of this article.

As we will see soon, regular languages and star free languages have important rôles to play in logic.

3 Logics on sequences

As we mentioned in the introductory section, we consider logics for reasoning about processes that happen in time. Modal logics provide the general framework in which philosphical logic tends to discuss such reasoning, but since our aim here is not to discuss specific modalities, we look for an abstract logic in which several such logics can be embedded.

3.1 First order

Let $E = \{a, b, c\}$ and the property that every occurrence of b is preceded by some earlier occurrence of a. A logical specification of this property would be as follows. We use variables to denote positions in sequences and unary predicates P_e for talking of event occurrences at specific positions: $\forall x.P_b(x) \supset (\exists y.y < x \wedge P_a(y))$. Note that the alternative is asserted by: $\exists x.P_b(x) \wedge \forall y.(y < x \supset P_c(y))$, though this is not immediately clear, looking at the sentences only syntactically. This is precisely the sort of reasoning we wish to carry out in a logic for sequences.

The logical language introduced by way of illustration above is the familiar language of first order predicate calculus. However, we will work with only a small fragment, which we define formally now.

Fix E, a finite event alphabet. Let $\mathrm{L}(E)$ denote the first order vocabulary $(\{min, max\}; <^2, \{P_e^1 \mid e \in E\})$, where we have two constant symbols *min* and *max*, one binary predicate symbol $<$ and a unary predicate symbol P_e for each e in E. There are no function symbols in the vocabulary. This defines the (monadic) first order language, often denoted $FO(<)$.

The syntax of the logic can be presented in inductive form as follows. Let V be a countable set of variables, with x, y etc denoting elements of V. A term is either a constant (*min* or *max*), or a variable. t, t' denote terms.

$$\varphi ::= P_e(t),\ e \in E \mid t = t' \mid t < t' \mid \neg\varphi \mid \varphi \vee \varphi' \mid \exists x.\varphi$$

The other boolean connectives $\wedge$, $\supset$, $\equiv$ etc are defined using $\neg$ and $\vee$. $\forall x.\varphi$ is defined as $\neg\exists x.\neg\varphi$. The notion of bound and free occurrences of a variable in a formula are defined as usual. We often write $\varphi(x)$ to signify that x occurs free in φ. A formula with no free variables is referred to as a sentence.

To give the semantics of formulas in the logic, we need first order structures on which to interpet them. Every $u \in E^*$ has a natural presentation as a structure in our language. For example, let $E = \{a, b\}$ and $w = abbab$. We can view this as a structure $(\{1, 5\}; <_N, P_a, P_b)$ where $<_N$ is the standard ordering of the natural numbers, $P_a = \{1, 4\}$ and $P_b = \{2, 3, 5\}$.

In general, given a word $w = e_1 \ldots e_m \in E^*$, let D_w denote the set $\{1, \ldots, |w|\}$. $\widehat{w}$ denotes the structure $(D_w; \{j, |w|\}; <_N, (P_e)_{e \in E})$, where $j \in \{0, 1\}$ and $j = 1$ iff $|w| > 0$; $P_e = \{i \mid 1 \leq i \leq |w|, e_i = e\}$. (Note that in $\widehat{\epsilon}$, P_e is the empty set for all e and $|\epsilon| = 0$.) We use P_e both to denote the predicate symbol in the syntax as well as the set in the structure; this should cause no confusion in usage.

A model is a pair $M = (\widehat{w}, \pi)$ where $w \in E^*$, and $\pi : V \to D_w$ is the interpretation map that assigns positions in the sequence to variables. We can now define the truth of a formula φ in the model M by induction on the structure of M. Note that *min* and *max* are interpreted as 1 (0 if $w = \epsilon$) and $|w|$ respectively, and $<$ is interpreted as the standard ordering on natural numbers. π is easily lifted so that it is defined on all terms.

- $M \models P_e(t)$ if $\pi(t) \in P_e$.
- $M \models t = t'$ if $\pi(t) = \pi(t')$.
- $M \models t < t'$ if $\pi(t) <_N \pi(t')$.
- $M \models \neg\varphi$ if $M \not\models \varphi$.
- $M \models \varphi \vee \varphi'$ if $M \models \varphi$ or $M \models \varphi'$.
- $M \models \exists x.\varphi$ if for some $i \in D_w$, $M'(i) \models \varphi$, where $M'(i) = (\widehat{w}, \pi')$ where $\pi' : V \to D_w$ is defined by: $\pi'(x) = i$, and $\pi'(y) = \pi(y)$ for $y \neq x$.

It is easily seen that $M \models \forall x.\varphi$ if for every $i \in D_w$, $M'(i) \models \varphi$, where $M'(i)$ is defined as above.

We say that a formula φ is satisfiable if there exists a model M such that $M \models \varphi$. We say φ is valid if $\neg\varphi$ is not satisfiable.

$\forall x. \bigvee_{e \in E} (P_e(x) \wedge \bigwedge_{e' \in E, e' \neq e} \neg P_{e'}(x))$ is a valid formula, whereas $min < max$ is not (since ϵ falsifies it). $min = max$ is satisfiable but $max < min$ is not.

We can define the *successor* relation: $S(x, y) = x < y \wedge \forall z.(x < z \wedge z < y) \supset (x = z \vee z = y)$. We often denote $S(x, y)$ by $y = x + 1$ though we do not have function symbols in the logic.

Let $FV(\varphi)$ denote the set of free variables in φ. Let M be a model $M = (\widehat{w}, \pi)$. To define $M \models \varphi$, it suffices to consider the model $M = (\widehat{w}, \pi')$ where π' is defined to agree with φ on $FV(\varphi)$ and arbitrary otherwise. That is, the truth of the formula in a structure is determined by the interpretation of its free variables. We will in general write $\widehat{w}, k_1, \ldots, k_m \models \varphi$, where $FV(\varphi) \subseteq \{x_1, \ldots, x_m\}$ and $\pi'(x_i) = k_i$.

Therefore, for a *sentence* φ, we have the notion of a structure making it true or false, and we can speak of $\widehat{w} \models \varphi$. Thus, given a sentence φ, we can define the language of φ: $L(\varphi) = \{w \in E^* \mid \widehat{w} \models \varphi\}$.

Let $L \subseteq E^*$. We say that L is first order definable if there exists a sentence φ such that $L = L(\varphi)$.

Note that *min* and *max* can be easily eliminated from the logic. Let $\varphi_0(x) = \forall y.(x = y \vee x < y)$ and $\varphi_1(x) = \forall y.(x = y \vee y < x)$. Now, given some formula φ, let y, z be variables that do not occur in φ, and let $\varphi[y/min][z/max]$ be the result of replacing every occurrence of *min* in φ by y and *max* in φ by z. Then the formula $\exists y.\exists z.(\varphi_0(y) \wedge \varphi_1(z) \wedge \varphi[y/min][z/max])$ is the one we want. Thus we can work with a purely relational vocabulary, with no constants or functions, but only variables, unary predicates and one binary relation ($<$).

Example: Let $E = \{a, b, c\}$, $w = babba$. Now consider the following formulas:

- Let $\varphi_1 = \forall x.P_b(x) \supset \exists y.(x < y \wedge P_a(y))$. Then $\widehat{w} \models \varphi_1$.
- $L(\varphi_1) = \{a, c\}^* \cup \{ubvaw \mid u, v \in E^*, w \in \{a, c\}^*, v \neq \epsilon\}$.
- Let $\varphi_2(y) = (\forall z.(y < z \supset P_a(z))$. Then $\widehat{w}, 5 \models \varphi_2$, $\widehat{w}, 4 \models \varphi_2$ but $\widehat{w}, 3 \not\models \varphi_2$.
- Let $\varphi_3(x) = \forall y.(x < y \wedge P_a(y) \supset (\exists z.x \leq z \wedge z < y \wedge P_b(z)))$. Then $\widehat{w}, j \models \varphi_3$ for every j.

  $L(\forall x.\varphi_3)$ is the language of sequences in which *every* suffix has the property that an a is preceded by a b. As one would expect, $L(\exists x.\varphi_3)$ is one where *some* a is preceded by a b. $L(\forall x.(P_a(x) \wedge \varphi_3))$ talks of sequences where there is a b between two a's.
- Let $\varphi_4 = \exists x.S(x, max) \wedge P_a(x)$. Then $\widehat{w} \not\models \varphi_4$.

Example: Once again let $E = \{a, b, c\}$. Now consider the following languages:

- L_1: Between any two occurrences of b's there are only a's.
- L_2: There are two occurrences of b's between which there are only a's.
- L_3: No two occurrences of b's are such that between them there are only a's.

Let $\varphi(x, y) = x < y \wedge P_b(x) \wedge P_b(y) \wedge \forall z.((x < z \wedge z < y) \supset P_a(z))$. Now define $\varphi_1 = \forall x.\forall y.\varphi$, $\varphi_2 = \exists x.\exists y.\varphi$, $\varphi_3 = \neg\varphi_2$. It is easy that φ_i defines L_i for $i = 1, 2, 3$. More importantly, the compositionality inherent in a logical description helps us construct the sentences easily, one using the other. Note how closely the logical description matches the loose description above in natural language.

Is every language first-order definable? If it were, the logic would not be very interesting, would it? But what is rather surprising is that even some regular languages are not first-order definable.

Now consider the language consisting of all sequences of *even length*. To keep matters simple, let us assume that $E = \{a\}$. Thus we want the language $\{\epsilon, aa, aaaa, aaaaaa, \ldots\}$. A deterministic automaton for the language was given in Figure 3.

If we had addition on positions, this would be easily defined: $\exists x.x + x = max$. But addition is not definable in the logic, and indeed, this 'little' language is not first order definable. The proof of this assertion requires algebraic techniques beyond the scope of this article. On the other hand, this has to do with memory (or lack of it) in first order descriptions; we will discuss this later on.

3.2 Monadic second order

If addition on positions is not definable in the first order logic we have, why not simply "add" it to the logic? The extended logic is indeed interesting and termed *Presburger arithmetic*. In such a logic, we can define not only even-ness, but also other interesting languages like the following one. Let $E = \{a, b\}$, and $\varphi = \exists x.(x + x = max \wedge P_a(x) \wedge \forall y.(y < x \supset P_a(x) \wedge x < y \supset P_b(x))$. It defines the set $\{a^n b^n \mid n \geq 0\}$, which we have already seen to be non-recognizable as it requires strong infrastructure, namely unbounded memory.

It turns out that we have another option, one that admits languages like the even-length one, but not the 'unbounded memory' one. This is to add set quantification to the logic.

Consider the following logic. Let V be a countable set of first order variables, and U be a countable set of set variables. We use x, y etc for elements of V and X, Y etc for elements of U.

$$\varphi ::= P_e(x),\ e \in E \mid x = y \mid x < y \mid x \in X \mid \neg\varphi \mid \varphi \vee \varphi' \mid \exists x.\varphi \mid \exists X.\varphi$$

We have added set membership and set quantification in the logic. The semantics of the logic is extended naturally.

A model is a triple $M = (\widehat{w}, \pi, \theta)$ where $w \in E^*$, $\pi : V \to D_w$ is the interpretation map for first order variables, and $\theta : V \to 2^{D_w}$ is the interpretation map for set variables. Then the semantics of formulas is defined inductively as before; we only present the new cases below.

- $M \models x \in X$ if $\pi(x) \in \theta(X)$.
- $M \models \exists X.\varphi$ if for some $D \subseteq D_w$, $M'(D) \models \varphi$, where $M'(D) = (\widehat{w}, \pi, \theta')$ where $\theta' : V \to 2^{D_w}$ is defined by: $\theta'(X) = D$, and $\theta'(Y) = \theta(Y)$ for $Y \neq X$.

As before, we will use the notation $\varphi(x_1, \ldots, x_m, Y_1, \ldots, Y_n)$ to denote a formula all whose free first order variables are among $x_1, \ldots, x_m$ and all whose second order variables are among $Y_1, \ldots, Y_n$. Its models are denoted by the tuple $(\widehat{w}, k_1, \ldots, k_m, D_1, \ldots, D_n)$, where each $k_i \in D_w$ and each $D_j \subseteq D_w$. When φ is a *sentence*, we define $L(\varphi) = \{w \in E^* \mid \widehat{w} \models \varphi\}$.

Note that we do not have the constants *min* and *max* in the logic, but we will freely use them in formulas with the understanding that we can always eliminate them. We will also use other abbreviations in particular:

$$X \subseteq Y = \forall x.x \in X \supset x \in Y.$$
$$Empty(X) = \neg\exists x.x \in X.$$
$$Sing(X) = \exists x.x \in X \wedge (\forall y.y \in X \supset x = y).$$

We showed earlier that the successor relation $S(x, y)$ can be defined from the order relation $x < y$ in first order logic. In MSO, the converse also holds. Consider $\varphi(x, y) = \neg x = y \wedge \exists X.(x \in X \wedge y \in X \wedge \forall z.\forall z'.((z \in X \wedge S(z, z')) \supset z' \in X)$. Then it is easily seen that for any $w \in E^*$, $(\widehat{w}, i, j) \models \varphi$ iff $i < j$.

This logic is referred to as Monadic Second Order logic (abbreviated MSO), since it is second order (quantification over relations rather than only over individual positions) and monadic second order since such quantification is only over unary relations (sets). Formally it is denoted by $MSO_E(<)$.

We say that $L \subseteq E^*$ is MSO-definable if there exists a sentence φ in the logic $MSO_E(<)$ such that $L = L(\varphi)$.

The 'even-length' language is defined by the following MSO sentence.

$$\exists X.(min \in X) \wedge (max \in X) \wedge \forall x.\forall y.S(x, y) \supset (x \in X \equiv y \notin X)$$

Let $E = \{a, b\}$ and $L = \{(ab)^n \mid n \geq 0\}$. We can define this by the MSO sentence: $\exists X.(min \in X) \wedge (max \in X) \wedge \forall x.$:

$$\forall y.S(x, y) \supset ((x \in X \wedge y \notin X \wedge P_a(x) \wedge P_b(y)) \vee (x \notin X \wedge P_b(x) \wedge y \in X \wedge P_a(y)))$$

But then this language is already first order definable and by a simpler sentence. In general, it is not easy to determine when an MSO definable language is FO definable.

Are there languages that are not MSO definable either? Yes, and these are languages that require unbounded memory. This is what we now proceed to establish.

4 The logic - automata connection

Automata over E specify collections of E-sequences, and so do formulas in $FO_E(<)$ and $MSO_E(<)$. What is the difference in expressive power between the two formalisms? The following celebrated theorem of **Büchi, Elgot and Trakhtenbrot** (1960) asserts that there is none.

Below, we use the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$, where $n > 0$.

Theorem 4.1. *$L \subseteq E^*$ is recognizable iff it is $MSO_E(<)$-definable.*

How can one hope to prove this theorem? In one direction, we are given an automaton recognizing L, say $A = (Q, \delta, I, F)$ over E. We wish to construct a sentence φ such that $L(\varphi) = L$. This means that the models of the sentence should be exactly accepting runs of A. An accepting run is a sequence $q_1 q_2 \ldots q_k$, where $q_1 \in I$, $q_k \in F$ and for all $i \in \{1, \ldots, k-1\}$ there exists $a_i \in E$ such that $(q_i, a_i, q_{i+1}) \in \delta$. How do we refer to states in formulas when we do not have Q in the syntax of the logic? But then the names of the states are quite irrelevant. There is a canonical representation of the automaton: suppose $|Q| = m$. Then we might as well refer to the states by the set $[m]$ and δ, I, F are redescribed appropriately.

Consider the accepting run $j_1 j_2 \ldots j_k$, where all the j_i are from the set $[m]$. The sequence can then be given by an m-way partition of the set $[k]$: $X_j = \{i \mid j_i = j\}$, with $j \in [m]$. The set X_j is indeed the state q_j of the automaton.

Now we can specify any accepting run of A on a word $a_1 \ldots a_k$ by an m-way partition $X_1, \ldots, X_m$ such that:

- if $min \in X_j$ then $j \in I$.

- if $i \in X_j$ and $i+1 \in X_\ell$ then $(j, a_i, \ell) \in \delta$.
- if $max \in X_j$ then $j \in F$.

Since each of these conditions can be expressed in logic, we are almost done. Since we have set quantification, we only need to say $\exists X_1 \ldots \exists X_m$ and then express the above. There is a slight modification required, since models of the logic are words $a_1 a_2 \ldots a_k$ but associated runs would be of length $k+1$: so the last condition needs to be changed appropriately.

Thus the sentence we seek can be given as follows:

$$\varphi_A = \exists X_1 \ldots \exists X_m : States \wedge Init \wedge Trans \wedge Fin$$

where *States* asserts that these sets partition the set of positions, and the other formulas correspond to conditions above:

- $States = \forall x.((x \in X_1 \vee \ldots \vee x \in X_m) \wedge \bigwedge_{i,j:i \neq j} \neg(x \in X_i \wedge x \in X_j))$.
- $Init = \bigvee_{j \in I} min \in X_j$.
- $Trans = \forall x.\forall y.(S(x,y) \supset (\bigvee_{(i,a,j) \in \delta} x \in X_i \wedge Q_a(x) \wedge y \in X_j))$.
- $Fin = \bigvee_{(i,a,j) \in \delta, j \in F} (max \in X_i \wedge Q_a(max))$.

It is then an easy exercise to show that A accepts w iff $\widehat{w} \models \varphi_A$.

Some observations are in order. Note that the sentence above is of the form $\exists X_1 \ldots \exists X_m \psi$ where ψ is a *first order* sentence using the predicates $X_1, \ldots X_m$. Thus we need only existential second order quantification to express automata.

When the automaton has m states, we have used m set variables. For those who care about such things, it should be clear that this is wasteful, and that at most $\lceil log_2 m \rceil$ variables suffice.

We thus have that every recognizable language is MSO-definable. The converse assures us that the non-regular languages cannot be described by the logic either.

We are given an MSO sentence φ and seek an automaton that accepts exactly the models of φ. What is the proof strategy for this construction? The semantics of the formula is "global" but it is defined inductively, hence the

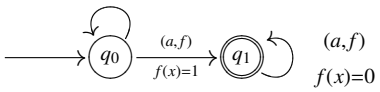


Figure 6: Automaton accepting the legal sequences

natural strategy is to proceed by induction on the structure of φ. Since recognizable languages are closed under complementation as well as union, it is clear that the boolean cases of the inductive step go through easily. Therefore, we need to think only in terms of atomic formulas and the cases of quantification.

For an inductive proof, it is clear that we cannot work with sentences, but that we need to work with formulas with free variables. For this, we extend the structure of words as follows: consider a word $w = a_1 \dots a_k$. Every first order variable is interpreted as some $i \in [k]$ and every set variable is a subset of $[k]$. Thus for each position $j \in [k]$ we can have an m-bit vector, specifying whether each of the m first order variables takes that value or not, and similarly an n-bit vector, specifying which of the sets j is in. This is formalized as follows.

Let φ be a formula with all its first order variables among $F_\varphi = \{x_1, \dots, x_m\}$ and all its set variables among $S_\varphi = \{X_1, \dots, X_n\}$. An interpretation for a word w such that $|w| = k$ is a pair $I_w = (I_F, I_S)$ where $I_F : F_\varphi \to [k]$ and $I_S : S_\varphi \to 2^{[k]}$.

Let $V_\varphi = F_\varphi \cup S_\varphi$. Let $\mathcal{F}_V = \{f \mid f : V_\varphi \to \{0,1\}\}$. Let $E_V = E \times \mathcal{F}_V$.

Thus, given a pair (w, I_w), where $w = a_1 \dots a_k$, the encoding $enc(w, I_w) = (a_1, f_1) \dots (a_k, f_k)$ is a word over E_V where $f_i(x_j) = 1$ iff $I_F(x_j) = i$ and $f_i(X_j) = 1$ iff $i \in I_S(X_j)$. On the other hand, given $u \in E_V^* = u_1 \dots u_k$, where $u_i = (a_i, f_i)$ we say u is legal if for all $x_j \in F_\varphi$, there exists a *unique* $i \in [k]$ such that $f_i(x_j) = 1$. Let Leg_V be the set of all legal words over E_V.

It is easily seen that *enc* gives a bijection between Leg_V and the set of pairs (w, I_w) of words with interpretations. Moreover we can check that Leg_V is recognizable: consider the automaton A_j below corresponding to $x_j \in F_\varphi$. Then $Leg_V = \bigcap_{j \in [m]} L(A_j)$. Since recognizable languages are closed under intersection, Leg_V is recognizable.

What we then need to prove is the following:

Lemma 4.2. *For every formula φ with all its first order variables among $F_\varphi = \{x_1, \dots, x_m\}$ and all its set variables among $S_\varphi = \{X_1, \dots, X_n\}$, there exists an automaton A_φ such that $L(A_\varphi) = \{enc(w, I_w) | (w, I_w) \models \varphi\}$.*

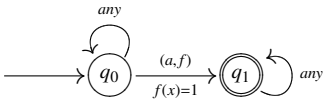


Figure 7: Automaton for $P_a(x)$

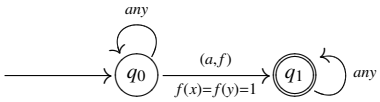


Figure 8: Automaton for $x = y$

Clearly, the theorem follows from the lemma, which is proved by induction on the structure of φ.

The base case is when φ is atomic, and is of the form $P_a(x), a \in E$, $x = y$, $x < y$, or $x \in X$. In each case we can construct an automaton as below, and take intersection with Leg_V.

The induction step is when φ is a negated formula, a disjunction, or a quantified formula.

$\varphi = \neg\varphi_1$: $L(\neg\varphi_1) = \overline{L(\varphi_1)} \cap Leg_V$. Since recognizable languages are closed under complementation and intersection, the required automaton exists.

$\varphi = \varphi_1 \vee \varphi_2$: Inductively consider A_{φ_1} and A_{φ_2} and "expand" their alphabets to E_V replacing every transition on (a, f) in A_{φ_1} by some (a, f') transition on the same states such that $f'(x) = f(x)$ for $x \in F_{\varphi_1}$ etc. Then we only need to appeal to closure of recognizable languages under union.

$\varphi = \exists x.\varphi_1$: Note that $L(\varphi)$ consists of those words $(a_1, f_1), \ldots, (a_k, f_k)$ such that for some word $(a_1, g_1), \ldots, (a_k, g_k)$ in $L(\varphi_1)$, each $g_i : (E_V \cup \{x\}) \to \{0, 1\}$ and $g_i\lceil E_V = f_i$. That is, f is the same as g, with the evaluation for x "discarded". Thus the construction is simple: consider the inductive automaton A_{φ_1}, and replace each (a, g) transition by (a, f) on the same

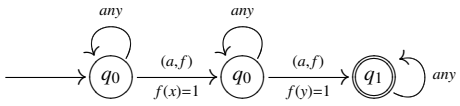


Figure 9: Automaton for $x < y$

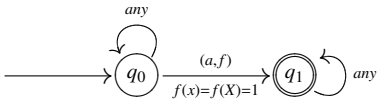


Figure 10: Automaton for $x \in X$

states, where $f = g\lceil E_V$.

$\varphi = \exists X.\varphi_1$: This case is proved exactly as above.

This completes the induction and proves the lemma. To prove the theorem, consider an MSO sentence φ and construct A_φ as above, then replace every (a, f) transition by a on the same states, and we are done.

4.1 Effectiveness

An EMSO sentence is an MSO formula of the form $\exists X_1 \ldots \exists X_m \psi$ where ψ is a *first order* sentence using the predicates $X_1, \ldots X_m$.

Corollary 4.3. *Every MSO sentence is equivalent to an EMSO sentence.*

This is a nontrivial normal form for the logic, obtained by a detour through automata theory. The fact that all set quantification can be pulled out to a prenex form is not obvious when we take a look at the semantics of the logic.

Theorem 4.1 offers us a construction of an MSO sentence corresponding to every automaton and vice versa. How effective is this construction?

We need to consider the cost of all the operations in the induction step. If we consider nondeterministic automata of size n inductively, union needs at most $2n + 1$ states, intersection at most n^2 states, projection at most n states, and complementation at most 2^n states. Does this mean that the construction is exponential? Sadly, it is much worse. Each projection may convert a deterministic automaton to a nondeterministic one, which means that the next complementation will involve the subset construction and hence cause an exponential blowup. Thus every quantifier alternation gives one exponent and we have a procedure that generates $2^{2^{\cdot^{\cdot^{2^n}}}}$ where the *tower* of exponents is of size k, where k is the number of logical operators in the sentence.

Can we do better? A theorem by Meyer and Stockmeyer, 1972 ([MS72]) asserts that such a *non-elementary* blowup is unavoidable.

4.2 Remarks

What we see here is a trade-off between two forms of specification, one logical, and the other in terms of automata. The former has a compositional semantics, and operations like complementation come for free. The latter has only local structure and combinational operators are hard to work with. On the other hand, there are easy and efficient algorithms to determine properties of automata, whereas algorithmic questions on logical formulas tend to be hard. Given an automaton, we can efficiently *minimize* it, so that we can work with the optimal one, whereas for formulas, it is hard to see such optimal equivalents.

4.3 First order

Theorem 4.1 establishes an equivalence between regular languages and MSO-definable languages. What can we say about FO-definable languages? Clearly this is a strict subclass of regular languages.

Theorem 4.4. Schützenberger, 1965 ([Sch65])*: FO-definable languages are exactly the star-free languages.*

The proof of this theorem involves use of algebraic machinery that is very interesting but beyond the scope of this elementary exposition. But one application of such machinery is the decidability of the following very interesting question: given an MSO sentence φ can we determine that it is already equivalent to some FO sentence (over the same vocabulary)? Yes, and an algorithm was presented by Hashiguchi in 1988 ([Has88]).

5 Generalizations

Having established a connection between definability in monadic second order logic and recognizability by finite state automata, one can ask whether this connection is of any significance beyond the relationship itself. Indeed, this connection is merely a starting point for journeys in many directions.

5.1 Arithmetic

Büchi used this in 1960 to prove the decidability of **WS1S**, the *weak* monadic second order theory of 1 successor: all $MSO(<)$ sentences true in the structure $(\omega, S, <)$ of natural numbers where the second order quantification is

contrained to be only over *finite* sets. From this, we can infer the decidability of Presburger arithmetic, the first order theory of addition on the structure $(\omega, +)$. The main idea is that a number, say 27, can be represented as a word 11011 which in turn can be represented by the finite set $\{0, 1, 3, 4\}$. Now we can write a sentence $\varphi(X_1, X_2, X_3)$ to assert that X_i represents the number x_i, and that $x_1 + x_2 = x_3$. Thus every sentence in $FO(+)$ can be translated into $WS1S$, establishing the decidability of Presburger arithmetic. Is there a translation possible in the other direction? It turns out ([MV96]) that we need to extend the first order theory of addition with an additional arithmetical predicate so that we can translate $WS1S$ into it: $V_2(m, n)$ if the highest power of 2 that divides m is n. There are many more interesting connections between definability in arithmetical theories and recognizability, see [MV96] for a nice survey.

5.2 Trees

We have all along considered only sequences as inputs to automata. A natural generalization is to consider automata over finite **trees** whose nodes are labelled by elements of the finite alphabet E. It is then convenient to present E as a tuple $(E_0, E_1, \ldots, E_m)$ where E_i is the set of elements of *arity* i: the idea is that when a tree node is labelled by $a \in E_i$, it has i children. The automaton can then proceed bottom-up: on reading each leaf, the automaton assumes a state. On reading any non-leaf node labelled $a \in E_k$, if after reading its k children, the automaton is in states $(q_1, \ldots, q_k)$, then the transition relation specifies what the next state can be, say q, which is the state that the automaton assumes. Proceeding this way, the automaton is in some state q_f when it reaches the root. If q_f is designated to be an accepting state, the tree is accepted, otherwise rejected. Thus we can associate a tree language with such a bottom-up tree automaton, and talk of recognizable tree languages. Correspondingly, we can define $MSO(S_1, S_2, \ldots S_m)$, the monadic second order logic with S_i representing the i-way generalization of the successor function. The details are interesting but not difficult, and we can prove a correspondence between recognizability of tree languages and MSO-definability. Many other techniques like determinization and minimization, and results like characterization of first order definability transfer neatly from words to trees.

Instead of a bottom-up processing tree automaton, we could also consider a top-down automaton that works its way downward, splitting into several states depending on the arity of the letter being read. Such automata are expressively equivalent to bottom-up automata, though the determinis-

tic subclass is strictly weaker in the top-down case.

Often, we wish to consider trees where the arities at nodes are not predetermined, as in the case of *XML documents* which are trees but of unbounded arity. Recognizability and definability of such unranked tree languages is a topic of much interest to contemporary researchers, as lifting the machinery from ranked to unranked trees proves to be difficult. A natural machine model to process such input is that of tree walking automata: from any node, the transition includes an instruction to move up, down, left, right etc thus enabling the automaton to walk all over the tree. Once again, the theory is difficult and still under development. The e-book [CDG+07] is a good source of information on automata over finite trees.

5.3 Infinite behaviours

The connection between finite state automata and monadic second order logic yields its most beautiful results in the context of infinite words and infinite trees. In 1962, Büchi showed that $S1S$, the monadic second order theory of 1 successor (that is, the set of sentences of $MSO(<)$ true in $(\omega, <)$) is decidable, by setting up a correspondence between definability in the logic and recognizability by a simple generalization of finite state automaton. Consider an automaton $A = (Q, \delta, I, F)$ over E. A run of A on an infinite word $w = a_0a_1\ldots$ is an infinite sequence $\rho = q_0q_1\ldots$, where for all $j \geq 0$, $(q_j, a_j, q_{j+1}) \in \delta$ and $q_0 \in I$. Since Q is finite, such a run must visit some states infinitely often. Let $Inf(\rho)$ denote this set. The run is accepting if $Inf(\rho) \cap F \neq \emptyset$. The word w is accepted by A if there is an accepting run on it.

Interestingly, almost all results and techniques transfer from finite to infinite words though with a great deal of work. For instance, the automata defined above are not determinizable. We need to enrich the acceptance condition to enumerate sets $F_1, \ldots, F_k$ such that for some i, $F_i = Inf(\rho)$ for an accepting run ρ. Such automata can then be determinized ([Saf88]) but the complexity is high.

In a celebrated theorem, Rabin ([Rab69]) showed the correspondence between automata on infinite trees (necessarily top-down) and $S2S$, the monadic second order theory of the infinite binary tree, thus proving the decidability of the latter. Not only is the proof intricate, but the core technique of complementing finite automata over infinite trees, turned out to be immensely useful for later developments. Today the decidability of many logical theories is proved by the technique of *interpretation* into $S2S$. The connection between this theory and that of infinite games has opened au-

tomata theory and logic to a range of new questions. The reader is referred to the excellent surveys in ([GTW02]) for an illuminating introduction to the theory.

Once we have moved from words to trees, the natural move to consider automata over **graphs**. Unfortunately, while the machine model is easy to define, it turns out to be too powerful. For instance, the problem of checking whether the automaton accepts some non-empty language at all, does not admit any algorithmic solution. Logics and automata on finite and infinite graphs is an arena of interesting current research ([Tho03]).

5.4 Infinite alphabets

We have all along been working only with finite alpabets. Considering that finite state machines have only bounded memory, it is *a priori* reasonable that their input alphabet is finite. If the input alphabet were infinite, it is hardly clear how such a machine could tell infinitely many elements apart. And yet, there are many good reasons to consider mechanisms that achieve precisely this. An important reason is data manipulation: when we need to work with data values such as integers, we implicitly need infinite state mecanisms, and the question is whether some techniques can at all be usefully lifted from the finite state experience to such situations.

Consider the set of all finite sequences of natural numbers (given in binary) separated by hashes. A word of this language, for example, is 100#11#1101#100. Now consider the subset L containing all sequences with some number repeating in it. It is easily seen that L is not regular. The problem with L has little to do with the representation of the input sequence. If we were given a bound on the numbers occurring in any sequence, we could easily build a finite state automaton recognizing L. The difficulty arises precisely because we don't have such a bound or because we have 'unbounded data'.

In the last decade, there have been interesting developments in the theory of automata over infinite alphabets, showing intriguing connections with variable restricted first order logics, as well as logics with several order relations. See [AR10] for a brief survey of these attempts.

6 Closing remarks

We started with a discussion of memory infrastructure required for reasoning and suggested that automata models offer abstractions of bounded

memory. This led us to the question of how we could describe such models in logic, closing the circle. Büchi's theorem shows an equivalence between definability in monadic second order logic over words and recognizability by finite state automata. The fact that this theorem can be generalized in many ways (to trees, infinite objects, and so on) attests to the basic nature of this equivalence. In formulaic terms, logical descriptions and memory models can be seen as being dual to each other, but identifying the precise connection between the two can be enriching.

In a sense, these computational models can be seen as an addition to the philosopher's toolkit, for finer analyses of expressiveness embodied in connectives and quantifiers. When we move to logics of agency and interaction, memory structures play an important role and such additions to the toolkit are indeed necessary. Refining this understanding of memory structures, and their associated logics offers a systematic way of addressing many questions of great complexity related to memory. This is of special interest when we consider not individual memory but collective and distributed memory. Logics of interaction can offer an abstract view of such memory and the reasoning involved. While there are many research attempts in this direction, these are early days and we need to await the clarity of insight comparable to Büchi's theorem from half a century ago.

On the other hand, the issue of how complex memory needs to be for effective reasoning, the space complexity of logics, has received little mathematical attention, and most natural questions remain unanswered. The interplay between these two strands, memory for logic and logics of memory, is likely to lead to a rich mathematical theory of relevance to logic and computation.

Acknowledgements

This presentation owes a great deal to Wolfgang Thomas and his survey ([Tho96]) in both content and style. I thank Johan van Benthem for encouragement and exhortation, and my co-workers Soumya Paul and Sunil Simon for help with figures. I thank the Netherlands Institute of Advanced Study (http://nias.knaw.nl) for a Lorentz Fellowship stay during which this article was written.

References

[ABSRW02] Michael Alekhnovich, Eli Ben-Sasson, Alexander A. Razborov, and Avi Wigderson. Space complexity in propo-

sitional calculus. *SIAM J. Comput.*, 31(4):1184–1211, 2002.

[AR10] M. Amaldev and R. Ramanujam. Automata on infinite alphabets. In *Modern Applications of Automata Theory*. World Scientific, 2010.

[BBLT06] Arnold Beckmann, Ulrich Berger, Benedikt Löwe, and John V. Tucker, editors. *Logical Approaches to Computational Barriers, Second Conference on Computability in Europe, CiE 2006, Swansea, UK, June 30-July 5, 2006, Proceedings*, volume 3988 of *Lecture Notes in Computer Science*. Springer, 2006.

[BDL08] Arnold Beckmann, Costas Dimitracopoulos, and Benedikt Löwe, editors. *Logic and Theory of Algorithms, 4th Conference on Computability in Europe, CiE 2008, Athens, Greece, June 15-20, 2008, Proceedings*, volume 5028 of *Lecture Notes in Computer Science*. Springer, 2008.

[CDG+07] H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi. Tree automata techniques and applications. Available on: `http://www.grappa.univ-lille3.fr/tata`, 2007. release October, 12th 2007.

[Dav06] Martin Davis. The church-turing thesis: Consensus and opposition. In Beckmann et al. [BBLT06], pages 125–132.

[FOC72] *13th Annual Symposium on Switching and Automata Theory, 25-27 October 1972, The University of Maryland, USA*. IEEE, 1972.

[GTW02] Erich Grädel, Wolfgang Thomas, and Thomas Wilke, editors. *Automata, Logics, and Infinite Games: A Guide to Current Research [outcome of a Dagstuhl seminar, February 2001]*, volume 2500 of *Lecture Notes in Computer Science*. Springer, 2002.

[Hal80] Maurice Halbwachs. *The collective memory*. Harper and Row Colophon Books, 1980.

[Has88] Kosaburo Hashiguchi. Algorithms for determining relative star height and star height. *Inf. Comput.*, 78(2):124–169, 1988.

[HU79] John E. Hopcroft and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979.

[J.R60] J.R. Büchi. Weak second order arithmetic and finite automata. *Z. Math. Logik Grundlagen Math.*, 6:66–92, 1960.

[KF94] Michael Kaminski and Nissim Francez. Finite-memory automata. *Theor. Comput. Sci.*, 134(2):329–363, 1994.

[Kol08] Antonina Kolokolova. Many facets of complexity in logic. In Beckmann et al. [BDL08], pages 316–325.

[Koz97] Dexter Kozen. *Automata Theory and Computability*. Springer, 1997.

[MS72] Albert R. Meyer and Larry J. Stockmeyer. The equivalence problem for regular expressions with squaring requires exponential space. In *FOCS* [FOC72], pages 125–129.

[MV96] Christian Michaux and Roger Villemaire. Presburger arithmetic and recognizability of sets of natural numbers by automata: New proofs of cobham's and semenov's theorems. *Ann. Pure Appl. Logic*, 77(3):251–277, 1996.

[Par71] Rohit Parikh. Existence and feasibility in arithmetic. *J. Symb. Log.*, 36(3):494–508, 1971.

[Rab69] Michael Rabin. Decidability of second order theories and automata on infinite trees. *Trans. Amer. Math. Soc.*, 141:1–35, 1969.

[RS59] Michael Rabin and Dana Scott. Finite automata and their decision problems. *IBM Journal of Research and Development*, 3(2):114–125, 1959.

[RV03] Branislav Rovan and Peter Vojtás, editors. *Mathematical Foundations of Computer Science 2003, 28th International Symposium, MFCS 2003, Bratislava, Slovakia, August 25-29, 2003, Proceedings*, volume 2747 of *Lecture Notes in Computer Science*. Springer, 2003.

[Saf88] S. Safra. On the complexity of ω-automata. In 29^{th} *IEEE FOCS*, pages 319–327, 1988.

[Sch65] Marcel Paul Schützenberger. On finite monoids having only trivial subgroups. *Information and Control*, 8(2):190–194, 1965.

[Sip05] Michael Sipser. *Introduction to the Theory of Computation (second edition)*. Course Technology, 2005.

[Tho90] Wolfgang Thomas. Automata on infinite objects. In *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics (B)*, pages 133–192. Springer, 1990.

[Tho96] Wolfgang Thomas. Languages, automata, and logic. In *Handbook of Formal Languages*, pages 389–455. Springer, 1996.

[Tho03] Wolfgang Thomas. Constructing infinite graphs with a decidable mso-theory. In Rovan and Vojtás [RV03], pages 113–124.

PART IV

Logics of Information and Agency

Logics of Informational Attitudes and Informative Actions

ERIC PACUIT*

1 Introduction

There is an extensive literature focused on using logical methods to reason about communities of agents engaged in some form of social interaction. Much of the work builds upon existing logical frameworks developed by philosophers and computer scientists incorporating insights and ideas from philosophy (especially epistemology and philosophy of action), game theory, decision theory and social choice theory. The result is a web of logical systems each addressing different aspects of rational agency and social interaction. This paper focuses one one aspect of this broad area: logical systems designed to model the agents' *informational attitudes* (eg., knowledge, belief, certainty) in social interactive situations. This includes notions of *group* knowledge and *informative action*. Indeed, a key challenge for the logician is to account for the many dynamic processes that govern the agents' (social) interactions over time. Inference, observation and communication are all examples of such processes that are the focus of current logics of informational update and belief revision (see, for example, the papers, [15, 31, 77][1]). This paper will introduce these epistemic and doxastic logics as models of "rational interaction" and provide pointers to some current literature.

The point of departure for modern epistemic and doxastic logic is Jaakko Hintikka's seminal text *Knowledge and Belief: An Introduction to the Logic of the Two Notions* [57][2]. In fact, Hintikka was not the first to recognize that discourse about knowledge and belief could be the subject of a logical analysis. Indeed, Hintikka cites G.H. Von Wright's *An Essay in Modal*

*A Resident Fellow and Assistant Professor, Dept. of Philosophy, Univ. of Tilburg

[1]Of course, one may argue that (logical) *inference* is the central topic of *any* logic. What we have in mind here is reasoning *about* agents that make inferences.

[2]This important book has recently been re-issued and extended with some of Hintikka's latest papers on epistemic logic [58].

Logic [94] as the starting point for his logical analysis. A comprehensive history of epistemic and doxastic logic is beyond the scope of this paper; however, the interested reader can consult the following three sources for relevant historical details:

1. Paul Gochet and Pascal Gribomont's article in the *Handbook of the History of Logic* [39] has an extensive discussion of the main highlights in the technical development of epistemic logic;

2. Robert Goldblatt's article in the *Handbook of the History of Logic* [40] has a nearly complete history of the mathematical development of modal logic in the 20th century; and

3. Vincent Hendricks and John Symons [55, Section 2] describe some key developments in modal logic that led to Hintikka's book.

While Hintikka's project sparked some discussion among mainstream epistemologists (especially regarding the "KK Principle": does knowing something imply that one knows that one knows it?[3], much of the work on epistemic and doxastic logic was taken over by Game Theorists [1] and Computer Scientists [36] in the 1990s. Recently, focus is shifting back to Philosophy with a growing interest in "bridging the gap between formal and mainstream epistemology": witness the collection of articles in [54] and the book *Mainstream and Formal Epistemology* by Vincent Hendricks (cf. the paper [53]).

Thus, the field of Epistemic Logic has developed into an interdisciplinary area no longer immersed *only* in the traditional questions of mainstream epistemology. Much recent work focuses on explicating epistemic issues in, for example, game theory [24] and economics [83], computer security [49, 80], distributed systems [47], and *social software* [73][4]. The situation is nicely summarized in a recent article by Robert Stalnaker who suggests that a logical analysis can

> "...bring out contrasting features of some alternative conceptions of knowledge, conceptions that may not provide plausible analyses of knowledge generally, but that may provide interesting models of knowledge that are appropriate for particular applications, and that may illuminate in an idealized way, one or another of the dimensions of the complex epistemological terrain." [91, pg. 170]

[3]Timothy Williamson [93, Chapter 5] has a well-known and persuasive argument against this principle (cf. [32] for a discussion of interesting issues for epistemic logic deriving from Williamson's argument).

[4]See also Parikh's contribution to this volume.

In this survey, the modeling of informational attitudes of a group of (rational) agents engaged in some form of social interaction (eg. having a conversation or playing a card game) takes center stage.

Many logical systems today focus on (individual and group) informational attitudes often with a special focus on how the agents' information changes over time. Sometimes differences between "competing" logical systems are technical in nature reflecting different conventions used by different research communities. And so, with a certain amount of technical work, such frameworks are seen to be equivalent up to model transformations (cf. [43, 69, 71, 17]). Other differences point to key conceptual issues about rational interaction.

The main objective of this paper is to not only to introduce important logical frameworks but also help the reader navigate the extensive literature on (dynamic) epistemic and doxastic logic. Needless to say, we will not be able to do justice to all of this extensive literature. This would require a textbook presentation. Fortunately, there are a number of excellent textbooks on this material (see [36, 31, 15]). The article will be self-contained, though familiarity with basic concepts in modal logic may be helpful[5].

2 Informational Attitudes

Contemporary epistemology provides us with a rich typology of informational attitudes. There are numerous notions of knowledge around: the pre-Gettier "justified true belief" view, reliability accounts [41], counterfactual accounts [70], and *active* vs. *passive* knowledge [90, pg. 299], to name just a few (cf. [87] for a survey). Similarly, beliefs come in many forms: graded or flat-out [50], conditional and lexicographic [24], safe and strong [4]. On top of all this, beliefs seem to be just one example in a large variety of "acceptance-like" attitudes [85]. In this paper, we concentrate on a general distinction between attitudes of *hard* and *soft* information [13, 4] without taking a stance on which of these attitudes, if any, should be seen as primary, either for epistemology in general or for specific applications.

Hard information, and its companion attitude, is information that is *veridical* and *not revisable*. This notion is intended to capture what the agents are fully and correctly certain of in a given social situation. So, if an agent has hard information that some fact φ is true, then φ really is true. In absence of better terminology and following common usage in the literature, we use the term *knowledge* to describe this very strong type of informational

[5] See [16] for a modern textbook introduction to modal logic and [20] for an overview of some of the more advanced topics.

attitude. However, we make no claim as to whether this notion captures one of the many notions of knowledge just mentioned (in fact, it probably does not) and simply note that "hard information" shares *some* of the characteristics that have been attributed to knowledge in the epistemological literature such as veridicality. *Soft information* is, roughly speaking, anything that is not "hard": it is not necessarily veridical and/or highly revisable in the presence of new information. As such, it comes much closer to *beliefs* or more generally attitudes that can be described as "regarding something as true" [84].

Thus, we identify *revisability* as a key distinguishing feature. Typically, discussions of epistemic logic focus instead on the epistemic capabilities of the *agents* such as *introspection* or *logical omniscience*. For example, it is typically assumed that if an agent has the (hard or soft) information that φ is true, then this fact is fully transparent (to the agent). In order keep the presentation manageable, we do not go into details about these interesting issues (cf. [36] for extensive discussions).

Before going into details, a few comments about the general approach to modeling are in order. The formal models introduced below can be broadly described as "possible worlds models" familiar in much of the philosophical logic literature. These models assume and underlying set of *states of nature* describing the (ground) facts about the situation being modeled that do not depend on the agents' uncertainties. Typically, these facts are represented by sentences in some propositional (or first-order) language. Each agent is assumed to entertain a number of *possibilities*, called *possible worlds* or simply *(epistemic) states*. These "possibilities" are intended to represent "the current state of the world". So each possibility is associated with a *unique* state of nature (i.e., there is a function from possible worlds to sets of sentences "true" at that world, but this function need not be 1-1 or even onto). Crucial for the epistemic-logic analysis is the assumption that there may be *different* possible worlds associated with the same state of nature. Such possible worlds are important for representing higher-order information (eg., information about the other agents' information). One final common feature is that the agents' informational attitudes are directed towards *propositions*, also called *events* in the game-theory literature, represented as sets of possible worlds. These basic modeling choices are not uncontroversial, but such issues are beyond the scope of this paper[6] and so we opt for mathematical precision in favor of philosophical carefulness.

[6]The interested reader can consult [75] for a discussion.

2.1 Models of Hard Information

Let Agt be a non-empty set of agents and At a (countable or finite) set of atomic sentences. Elements $p \in \mathsf{At}$ are intended to describe ground facts, for example, "it is raining" or "the red card is on the table", in the situation being modeled. A non-empty set W, called *possible worlds* or *states*, are intended to represent the different ways the situation being modeled may evolve. Rather than *directly* representing the agents' *hard information*, the models given below describe the "implicit consequences" of this information in terms of "*epistemic indistinguishability relations*"[7]. The idea is that each agent has some "hard information" about the situation being modeled and agents cannot distinguish between states that agree on this information. In basic epistemic models, this "epistemic indistinguishability" is represented by *equivalence relations* on W:

Definition 2.1 (Epistemic Model). An **epistemic model** (based on the set of agents Agt and set of atomic propositions At) is a tuple $\langle W, \{\sim_i\}_{i\in\mathsf{Agt}}, V\rangle$ where W is a non-empty set; for each $i \in \mathsf{Agt}$, $\sim_i \subseteq W \times W$ is reflexive, transitive and symmetric; and $V : \mathsf{At} \to \wp(W)$ is a valuation function. ◁

A simple propositional modal language will be used to describe properties of these structures. Formally, let $\mathcal{L}_{EL}$ be the (smallest) set of sentences generated by the following grammar:

$$\varphi \;:=\; p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi$$

where $p \in \mathsf{At}$ and $i \in \mathsf{Agt}$. The additional propositional connectives $(\rightarrow, \leftrightarrow, \vee)$ are defined as usual and the dual of K_i, denoted L_i, is defined as follows: $L_i\varphi \;:=\; \neg K_i \neg\varphi$. The intended interpretation of $K_i\varphi$ is "according to agent i's current (hard) information, φ is true" (following standard notation we can also say "agent i knows that φ is true"). Given a story or situation we are interested in modeling, each state $w \in W$ of an epistemic model $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathsf{Agt}}, V\rangle$ represents a possible scenario which can be described in the formal language given above: if $\varphi \in \mathcal{L}_{EL}$, $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathsf{Agt}}, V\rangle$ and $w \in W$, we write $\mathcal{M}, w \models \varphi$ if φ is a correct description of some aspect of the situation represented by w. This can be made precise as follows:

[7] The phrasing "epistemic indistinguishability", although common in the epistemic logic literature, is misleading since, as a relation, "indistinguishability" is *not* transitive. A standard example is: a cup of coffee with n grains of sugar is indistinguishable from a cup with $n+1$ grains; however, transitivity would imply that a cup with 0 grains of sugar is indistinguishable from a cup with 1000 grains of sugar. In this context, two states are "epistemicly indistinguishable" for an agent if the agent has the "same information" in both states. This is indeed an equivalence relation.

Definition 2.2 (Truth). Let $\mathcal{M} = \langle W, \{\sim_i\}_{i \in \mathsf{Agt}}, V\rangle$ be an epistemic model. For each $w \in W$, φ **is true at state** w, denoted $\mathcal{M}, w \models \varphi$, is defined by induction on the structure of φ:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_i\varphi$ iff for all $v \in W$, if $w \sim_i v$ then $\mathcal{M}, v \models \varphi$

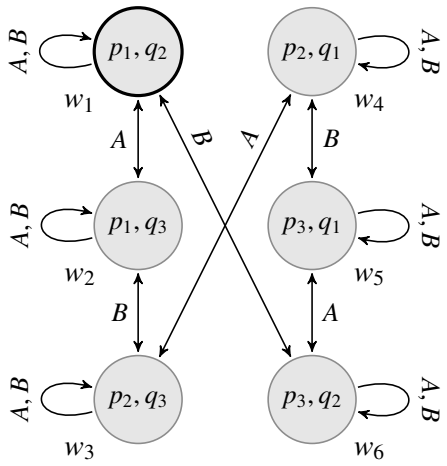
We say φ is **satisfiable** if there is an epistemic model $\mathcal{M} = \langle W, \{\sim_i\}_{i \in \mathsf{Agt}}, V\rangle$ and state $w \in W$ such that $\mathcal{M}, w \models \varphi$; and φ is **valid in** $\mathcal{M}$, denoted $\mathcal{M} \models \varphi$ if $\mathcal{M}, w \models \varphi$ for all $w \in W$. ◁

Given the definition of the dual of K_i, it is easy to see that

$$\mathcal{M}, w \models L_i\varphi \text{ iff there is a } v \in W \text{ such that } \mathcal{M}, v \models \varphi.$$

Thus an interpretation of $L_i\varphi$ is "φ is consistent with agent i's current (hard) information". The following example will illustrate the above definitions.

Suppose there are two agents, Ann (A) and Bob (B), and three cards labeled with the numbers 1, 2 and 3. Consider the following scenario: Ann is dealt one of the cards, Bob is given one of the cards and the third card is put face down on a table. What are the relevant possible worlds for this scenario? The answer to this question depends, in part, on the level of detail in the description of the situation being modeled. For example, relevant details may include whether Ann is holding the card in her right hand or left hand, the color of the cards or whether it is raining outside. The level of detail is fixed by the choice of atomic propositions. For example, suppose that $\mathsf{At} = \{p_1, p_2, p_3, q_1, q_2, q_3\}$ where p_i is intended to mean that "Ann has card i" and q_i is intended to mean "Bob has card i". Since each agent is given precisely one of the three possible cards, there are 6 relevant possible worlds, $W = \{w_1, w_2, w_3, w_4, w_5, w_6\}$, one for each way the cards could be distributed. What about the agents' information? Some of the aspects about the situation being modeled can be classified as "informative" for the agents. For example, since the third card is placed face down on the table, neither agent "knows" the number written on the other agent's card. The complete epistemic state of the agents is described by the epistemic model pictured below (in this picture, an i-labeled arrow from state w to state v means $w \sim_i v$ and each state is labeled with the atomic propositions true at that state):



- $\mathcal{M}, w_1 \models K_A p_1 \wedge K_A \neg q_1$
- $\mathcal{M}, w_1 \models K_A(q_2 \vee q_3)$
- $\mathcal{M}, w_1 \models K_A(K_B p_2 \vee K_B \neg p_2)$

The reader is invited to check that the formulas to the right are indeed true at state w_1 according to Definition 2.2. The intuitive interpretation of these formulas describe (part of) the hard information that Ann has in the above situation. For example, Ann knows that she has card 1 (i.e., it is assumed that Ann is looking at her card); Ann knows that Bob does not have card 1 (because, for example, Ann has background knowledge that there are only three cards with no duplicates); and Ann knows that Bob either has card 2 or card 3 and she knows that Bob knows *whether* he has card 2 (this can also be derived from her background knowledge).

Notice that the set of states that Ann considers possible at w_1 is $\{w_1, w_2\}$. This set is the truth set of the formula p_1 (i.e., $\{x \mid \mathcal{M}, x \models p_1\} = \{w_1, w_2\}$); and so, we can say that *all Ann knows about the situation is that she has card 1*. The other propositions that Ann knows are (non-monotonic) *consequences* of this proposition (given her background knowledge about the situation). This suggests that it may be useful to include an operator "for all agent i knows". In fact, this notion was introduced by Hector [64] and, although the logical analysis turned out to be a challenge cf. [62, 45, 46, 33], has proven useful in the epistemic analysis of certain solution concepts (cf. the paper[48]).

The above epistemic models are intended to represent the agents' *hard information* about the situation being modeled. In fact, we can be much more precise about the sense in which these models "represent" the agents' hard information by using standard techniques from the mathematical theory of modal logic [20]. In particular, *modal correspondence theory* rigorously relates properties of the the relation in an epistemic model with

modal formulas cf. Chapter 3 [20][8]. The following table lists some key formulas in the language $\mathcal{L}_{EL}$ with their corresponding (first-order) property and the relevant underlying assumption.

Assumption	Formula	Property
Logical Omniscience	$K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$	—
Veridical	$K_i\varphi \to \varphi$	Reflexive
Positive Introspection	$K_i\varphi \to K_iK_i\varphi$	Transitive
Negative Introspection	$\neg K_i\varphi \to K_i\neg K_i\varphi$	Euclidean

Viewed as a description, even an idealized one, of *knowledge*, the above properties have raised many criticisms. While the logical omniscience assumption (which is valid on all models regardless of the properties of the accessibility relation) generated the most extensive criticisms [88] and responses (cf. [36, Chapter 9]), the two introspection principles have also been the object of intense discussion (cf. [93, 32][9]). These discussions are fundamental for the theory of knowledge and its formalization, but here we choose to bracket them, and instead take epistemic models for what they are: models of hard information, in the sense introduced above.

2.2 Varieties of Soft Information

A small modification of the above epistemic models allows us to model a softer informational attitude. Indeed, by simply replacing the assumption of reflexivity of the relation $\sim_i$ with seriality (for each state w there is a

[8]To be more precise, the key notion here is *frame definability*: a frame is a pair $\langle W, R\rangle$ where W is a nonempty set and R a relation on W. A modal formula is valid on a frame if it is valid in every model (cf. Definition 2.1) based on that frame. It can be shown that some modal formulas have first-order *correspondents* P where for any frame $\langle W, R\rangle$, the relation R has property P iff φ is valid on $\langle W, R\rangle$. A highlight of this theory is *Sahlqvist's Theorem* which provides and algorithm for finding first-order correspondents for certain modal formulas. See Sections 3.5 - 3.7 [20] for an extended discussion.

[9]In fact, Hintikka explicitly rejects negative introspection: "The consequences of this principle, however, are obviously wrong. By its means (together with certain intuitively acceptable principles) we could, for example, show that the following sentence is self sustaining $p \to K_iL_ip$" [57, pg. 54]. Hintikka regards this last formula as counter-intuitive since it means that if it possible that an agent knows some fact p then that fact must be true. However, it seems plausible that an agent can justifiably believe that she knows some fact p but p is in fact false. Other authors have pointed out difficulties with this principle in modal systems with both knowledge and belief modalities: see, in particular, [91] and [86, Section 13.7].

state v such that $w \sim_i v$), but keeping the other aspects of the model the same, we can capture what epistemic logicians have called "*beliefs*". Formally, a **doxastic model** is a tuple $\langle W, \{R_i\}_{i\in \mathsf{Agt}}, V\rangle$ where W is a nonempty set of states, R_i is a transitive, Euclidean and serial relation on W and V is a valuation function (cf. Definition 2.1). Truth is defined precisely as in Definition 2.2, replacing $\sim_i$ with R_i. This notion of belief is very close to the above hard informational attitude and, in fact, shares all the properties of K_i listed above except *Veracity* (this is replaced with a weaker assumption that agents are "consistent" and so cannot believe contradictions). This points to a logical analysis of both informational attitudes with various "bridge principles" relating knowledge and belief (such as knowing something implies believing it or if an agent believes φ then the agent knows that he believes it). However, we do not discuss this line of research here since these models are not our preferred ways of representing the agents' soft information (see, for example, [42, 91]).

A key aspect of beliefs which is not yet represented in the above models is that they are *revisable* in the presence of new information. While there is an extensive literature on the theory of belief revision (see the article by Booth and Meyer in this collection for a discussion), the focus here is how to extend the above models with a representation of softer, revisable informational attitudes. The standard approach is to include a *plausibility ordering* for each agent: a preorder (reflexive and transitive) denoted $\preceq_i \subseteq W \times W$. If $w \preceq_i v$ we say "player i considers v at least as plausible as w." For $X \subseteq W$, let

$$Min_{\preceq_i}(X) = \{v \in W \mid v \preceq_i w \text{ for all } w \in X\}$$

denote the set of minimal elements of X according to $\preceq_i$. Thus while the $\sim_i$ partitions the set of possible worlds according to the hard information the agents are assumed to have about the situation, the plausibility ordering $\preceq_i$ represents which of the possible worlds the agent considers more likely (i.e., it represents the players soft information). Models representing both the agents' hard and soft information have been used not only by logicians [12, 30, 4] but also by game theorists [21] and computer scientists [23, 63]:

Definition 2.3 (Epistemic-Doxastic Models). Suppose Agt is a set of agents and At a set of atomic propositions, an **epistemic doxastic model** is a tuple $\langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, \{\preceq_i\}_{i\in \mathsf{Agt}}, V\rangle$ where $\langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, V\rangle$ is an epistemic model and for each $i \in \mathsf{Agt}$, $\preceq_i$ is a well-founded[10], reflexive and transitive relation on W satisfying the following properties, for all $w, v \in W$

[10]Well-foundedness is only needed to ensure that for any set X, $Min_{\preceq_i}(X)$ is nonempty. This is important only when W is infinite.

1. *plausibility implies possibility*: if $w \preceq_i v$ then $w \sim_i v$.

2. *locally-connected*: if $w \sim_i v$ then either $w \preceq_i v$ or $v \preceq_i w$. ◁

Remark 2.4. *Note that if $w \prec_i v$ then, since $\sim_i$ is symmetric, we also have $v \prec_i w$, and so by property 1, $w \not\preceq_i v$ and $v \not\preceq_i w$. Thus, we have the following equivalence: $w \sim_i v$ iff $w \preceq_i v$ or $v \preceq_i w$.*

Let $[w]_i$ be the equivalence class of w under $\sim_i$. Then local connectedness implies that $\preceq_i$ totally orders $[w]_i$ and well-foundedness implies that $Min_{\preceq_i}([w]_i)$ is nonempty. This richer model allows us to formally define a variety of (soft) informational attitudes. We first need some additional notation: the plausibility relation $\preceq_i$ can be lifted to subsets of W as follows[11]

$$X \preceq_i Y \text{ iff } x \preceq_i y \text{ for all } x \in X \text{ and } y \in Y$$

Suppose $\mathcal{M} = \langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, \{\preceq_i\}_{i\in \mathsf{Agt}}, V\rangle$ is an epistemic-doxastic model with $w \in W$, consider the following extensions to the language $\mathcal{L}_{EL}$

- *Belief*: $\mathcal{M}, w \models B_i\varphi$ iff for all $v \in Min_{\preceq_i}([w]_i)$, $\mathcal{M}, v \models \varphi$. This is the usual notion of belief which satisfies the standard properties discussed above (eg., positive and negative introspection).

- *Safe Belief*: $\mathcal{M}, w \models \Box_i\varphi$ iff for all v, if $v \preceq_i w$ then $\mathcal{M}, v \models \varphi$.Thus, φ is safely believed if φ is true in *all* states the agent considers more plausible. This stronger notion of belief has also been called *certainty* by some authors (cf. [86, Section 13.7]).

- *Strong Belief*: $\mathcal{M}, w \models B_i^s\varphi$ iff there is a v such that $w \sim_i v$ and $\mathcal{M}, v \models \varphi$ and $\{x \mid \mathcal{M}, x \models \varphi\} \cap [w]_i \preceq_i \{x \mid \mathcal{M}, x \models \neg\varphi\} \cap [w]_i$.So φ is strongly believed provided it is epistemically possible and agent i considers *any* state satisfying φ more plausible than *any* state satisfying $\neg\varphi$. This notion has also been studied by Stalnaker [89] and Battigalli and Siniscalchi [10].

The logic of these notions has been extensively studied by Alexandru Baltag and Sonja Smets in a series of articles [4, 6, 8]. We conclude this section with a few remarks about the relationship between these different notions. For example, it is not hard to see that if agent i knows that φ then i (safely, strongly) believes that φ. However, much more can be said about the logical relationship between these different notions (cf. [8]).

[11] This is only one of many possible choices here, but it is the most natural in this setting (cf. [68, Chapter 4]).

As noted above, a crucial feature of these informational attitudes is that they are *defeasible* in light of new evidence. In fact, we can characterize these attitudes in terms of the type of evidence which can prompt the agent to adjust her beliefs. To make this precise, we introduce the notion of a *conditional belief*: suppose $\mathcal{M} = \langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, \{\preceq_i\}_{i\in \mathsf{Agt}}, V\rangle$ is an epistemic-doxastic and φ and ψ are formulas, then we say *i believes* φ *given* ψ, denoted $B_i^{\psi}\varphi$, provided

$$\mathcal{M}, w \models B_i^{\psi}\varphi \text{ iff for all } v \in Min_{\preceq_i}([\![\psi]\!]_{\mathcal{M}} \cap [w]_i),\ \mathcal{M}, v \models \varphi$$

where $[\![\varphi]\!]_{\mathcal{M}} = \{w \mid \mathcal{M}, w \models \varphi\}$ is the *truth set* of φ. So, 'B_i^{ψ}' encodes what agent i will believe upon receiving (possibly misleading) evidence that ψ is *true*. Two observations are immediate. First of all, we can now define belief $B_i\varphi$ as $B_i^{\top}\varphi$ (belief in φ given a tautology). Second, unlike beliefs, conditional beliefs may be inconsistent (i.e., $B^{\psi}\bot$ may be true at some state). In such a case, agent i cannot (on pain of inconsistency) revise by ψ, but this will only happen if the agent has hard information that ψ is false. Indeed, $K\neg\varphi$ is logically equivalent to $B_i^{\varphi}\bot$ (over the class of epistemic-doxastic models). This suggests the following (dynamic) characterization of an agents' hard information as unrevisable beleifs:

$$\mathcal{M}, w \models K_i\varphi \text{ iff } \mathcal{M}, w \models B_i^{\psi}\varphi \text{ for all } \psi$$

Safe belief and strong belief can be similarly characterized by restricting the admissible evidence:

- $\mathcal{M}, w \models \Box_i\varphi$ iff $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ with $\mathcal{M}, w \models \psi$. That is, i safely beliefs φ iff i continues to believe φ given any true formula.
- $\mathcal{M}, w \models B_i^s\varphi$ iff $\mathcal{M}, w \models B_i\varphi$ and $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ with $\mathcal{M}, w \models \neg K_i(\psi \rightarrow \neg\varphi)$. That is, agent i strongly believes φ iff i believes φ and continues to believe φ given any evidence (truthful or not) that is not known to contradict φ.

Baltag and Smets [8] provide an elegant logical characterization of the above notions by adding the safe belief modality ($\Box_i$) to the epistemic language $\mathcal{L}_{EL}$ (denote the new language $\mathcal{L}_{EDL}$). First of all, note that conditional belief (and hence belief) and strong belief are *definable* in this language:

- $B_i^{\varphi}\psi \ :=\ L_i\varphi \rightarrow L_i(\varphi \wedge \Box_i(\varphi \rightarrow \psi))$
- $B_i^s\varphi \ :=\ B_i\varphi \wedge K_i(\varphi \rightarrow \Box_i\varphi)$

All that remains is to characterize properties of an epistemic-doxastic model (Definition 2.3). As discussed in the previous Section, K_i satisfies logical omnsicience, veracity and both positive and negative introspection. Safe belief, $\Box_i$, shares all of these properties except negative introspection. Modal correspondence theory can again be used to characterize the remaining properties:

- Knowledge implies safe belief: $K_i\varphi \rightarrow \Box_i\varphi$ (*Definition 2.3, property 1*)
- Locally connected: $K_i(\varphi \vee \Box\psi) \wedge K_i(\psi \vee \Box\varphi) \rightarrow K_i\varphi \vee K_i\psi$ (*Definition 2.3, property 2*)

Remark 2.5. *The above models use a "crisp" notion of uncertainty, i.e., for each agent and state w, any other state $v \in W$ is either is or is not possible/more plausible than w. However, there is an extensive body of literature developing* graded*, or* quantitative*, models of uncertainty [44]. For instance, in the Game Theory literature it is standard to represent the players'* beliefs *by probabilities [2, 51]. The idea here is to use probability distributions in place of the above plausibility orderings. Formally, a epistemic-proabalistic model is a tuple $\mathcal{M} = \langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, \{P_i\}_{i\in \mathsf{Agt}}, V\rangle$ where $\langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, V\rangle$ is an epistemic model and $P_i : W \rightarrow \Delta(W)$ ($\Delta(W) = \{p : W \rightarrow [0, 1] \mid p$ is a probability measure $\}$) assigns to each state a probability measure over W. Write p_i^w for the i's probability measure at state w. We make two natural assumptions (cf. Definition 2.3):*

1. *For all $v \in W$, if $p_i^w(v) > 0$ then $p_i^w = p_i^v$ (i.e., if i assigns a nonzero probability to state v at state w then the agent uses the same probability measure at both states)*
2. *For all v, if $w \nsim_i v$ then $p_i^w(v) = 0$ (i.e., assign nonzero probability only to the states in i's (hard) information set, compare this with Definition 2.3 item 1).*

Many different formal languages have been used to describe these rich structures. Examples range from '$\Box_i\varphi$' with the intended meaning "φ is more probable than $\neg\varphi$ for agent i" [56] to more expressive languages containing operators of the form $B_i^q\varphi$ (with q a rational number) and interpreted as follows:

$$\mathcal{M}, w \models B_i^q(\varphi) \text{ iff } p_i^w(\{v \mid \mathcal{M}, v \models \varphi\}) \geq q.$$

These models have also been the subject of sophisticated logical analyses [35, 34, 52] complementing the logical frameworks introduced in this paper [5].

2.3 Group Attitudes

Suppose there are two friends Ann and Bob on a bus separated by a crowd. Before the bus comes to the next stop a mutual friend from outside the bus yells "get off at the next stop to get a drink?". Say Ann is standing near the front door and Bob near the back door. When the bus comes to a stop, will they get off? Of course, this depends, in part, on Ann and Bob's preferences. Suppose that both Ann and Bob want to have a drink with their mutual friend, but *only if both are there for the drink*. So Ann will only get off the bus if she "knows" (justifiable believes) that Bob will also get off (similarly for Bob). But this does not seem to be enough (after all, she needs some assurance that Bob is thinking along the same lines). In particular, she needs to "know" (justifiably believe) that Bob "knows" (justifiably believes) that she is going to get off at the next stop. Is this state of knowledge sufficient for Ann and Bob to coordinate their actions? Both Lewis [65] and Clark and Marshall [28] argue that a condition of *common knowledge* is necessary for such coordinated actions. In fact, a seminal result by Halpern and Moses [47] shows that, without synchronized clocks, such coordinated action is impossible. Michael Chwe [27] has a number of examples that point out the everyday importance of the notion of common knowledge.

Both the game theory community and the epistemic logic community have extensively studied formal models of common knowledge and belief. Barwise [9] highlights three main approaches to formalize common knowledge: (i) the iterated view, (ii) the fixed-point view and (iii) the shared situation view. Here we will focus only on the first two approaches (cf. [18] for a rigorous comparison between (i) and (ii)). Vanderschraaf and Sillari [92] provide an extensive discussion of the literature (see also [36] for a general discussion).

Consider the statement "everyone in group G knows φ". If there are only finitely many agents, this notion can be easily defined in the basic epistemic language $\mathcal{L}_{EL}$:

$$E_G\varphi \;:=\; \bigwedge_{i\in G} K_i\varphi$$

where $G \subseteq \mathsf{Agt}$. Following Lewis [65][12], the intended interpretation of "it is common knowledge in G that φ" (denoted $C_G\varphi$) is the infinite conjunc-

[12] Although see [29] for an alternative reconstruction of Lewis' notion of common knowl-

tion:

$$\varphi \wedge E_G\varphi \wedge E_GE_G\varphi \wedge E_GE_GE_G\varphi \wedge \cdots$$

However, this involves an *infinite* conjunction, so cannot be a formula in the language of epistemic logic. This suggests that common knowledge is not definable in the language of multi-agent epistemic logic[13]. Thus we need to add a new symbol to the language $C_G\varphi$ whose intended interpretation is "it is common knowledge in the group G that φ". Let $\mathcal{L}_{EL}^C$ be the smallest set generated by the following grammar:

$$p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid C_G\varphi$$

with $p \in \mathsf{At}$ and $G \subseteq \mathsf{Agt}$.

Before giving semantics to $C_G\varphi$, we consider $E_GE_GE_G\varphi$. This formula says that "everyone from group G knows that everyone from group G knows that everyone from group G knows that φ". When will this be true at a state w in an epistemic model? First some notation: a **path on length** n **for** G in an epistemic model is a sequence of states $(w_0, w_2, \ldots, w_n)$ where for each $l = 0, \ldots, n-1$, we have $w_l \sim_i w_{l+1}$ for some $i \in G$ (for example $w_0 \sim_1 w_1 \sim_2 w_2 \sim_1 w_3$ is a path of length 3 for $\{1, 2\}$). Thus, $E_GE_GE_G\varphi$ is true at state w iff every path of length 3 for G starting at w leads to a state where φ is true. This suggests the following definition:

Definition 2.6 (Interpretation of C). Let $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, V\rangle$ be an epistemic model and $w \in W$. The truth of formulas of the form $C\varphi$ is:

$$\mathcal{M}, w \models C_G\varphi \text{ iff } \text{ for all } v \in W, \text{ if } wR_G^*v \text{ then } \mathcal{M}, v \models \varphi$$

where $R_G^* := (\bigcup_{i\in G} \sim_i)^*$ is the reflexive transitive closure of $\bigcup_{i\in G} \sim_i$. ◁

Sometimes it is useful to work with this equivalent characterization:

$\mathcal{M}, w \models C_G\varphi$ iff every finite path for G from w ends with a state satisfying the formula φ.

The logical analysis is much more complicated in languages with a common knowledge operator; however, the following two axioms can be said to characterize[14] common knowledge:

edge. A more precise statement is "the infinite conjunction is a necessary but not a sufficient condition for common knowledge".

[13] In fact, one can prove this using standard methods in modal logic.

[14] Techniques similar to the previously mentioned *correspondence theory* can be applied here to make this precise: see [14] for a discussion.

- Fixed-Point: $C_G\varphi \rightarrow E_G C_G \varphi$
- Induction: $\varphi \wedge C_G(\varphi \rightarrow E_G\varphi) \rightarrow C_G\varphi$

The first formula captures the "self-evident" nature of common knowledge: if some fact is common knowledge in some group G then everyone in G not only knows the fact but also that it is common knowledge. Robert Aumann [1] uses this as an alternative characterization of common knowledge:

> Suppose you are told "Ann and Bob are going together,"' and respond "sure, that's common knowledge." What you mean is not only that everyone knows this, but also that the announcement is pointless, occasions no surprise, reveals nothing new; in effect, that the situation after the announcement does not differ from that before. ...the event "Ann and Bob are going together" — call it E — is common knowledge if and only if some event — call it F — happened that entails E and also entails all players' knowing F (like all players met Ann and Bob at an intimate party). [1, pg. 271]

Remark 2.7. *In this section we have focused only on the notion of common* knowledge *(hard information). What about notions of common (safe, strong) belief? The general approach outlined above also works for these informational attitudes: for example, suppose* $wR_i^B v$ *iff* $v \in Min_{\preceq_i}([w]_i)$ *and define* R_G^B *to be the transitive closure of* $\cup_{i\in G} R_i^B$*. Of course, this does suggest interesting technical and conceptual issues, but these are beyond the scope of this paper (cf. [22, 66, 67]).*

While it is true that coordinated actions do happen, the analysis of many social situations suggests that other "levels of knowledge", short of the above infinite-common knowledge level are also relevant. Such levels can arise in certain pragmatic situations:

Example 2.8. *Suppose that Ann would like Bob to attend her talk; however, she only wants Bob to attend if he is interested in the subject of her talk, not because he is just being polite. There is a very simple procedure to solve Ann's problem: Have a (trusted) friend tell Bob the time and subject of her talk.*

Taking a cue from computer science, perhaps we can prove *that this simple procedure correctly solves Ann's problem. However, it is not so clear how to define a correct solution to Ann's problem. If Bob is actually present*

during Ann's talk, can we conclude that Ann's procedure succeeded? Not really. Bob may have figured out that Ann wanted him to attend, and so is there only out of politeness. Thus for Ann's procedure to succeed, she must achieve a certain "level of knowledge" (cf. [74]) between her and Bob. Besides both Ann and Bob knowing about the talk and Ann knowing that Bob knows, we have

Bob does not know *that Ann knows about the talk.*

This last point is important, since, if Bob knows that Ann knows that he knows about the talk, he may feel social pressure to attend[15]*. Thus, the procedure* to have a friend tell Bob about the talk, but not reveal that it is at Ann's suggestion*, will satisfy all the conditions. Telling Bob directly will satisfy the first three, but not the essential last condition.*

We conclude this section by briefly discussing another notion of "group knowledge": *distributed knowledge*. Intuitively, φ is distributed knowledge among a group of agents if φ would be known if all the agents in the group put all their information together. Formally, given an epistemic model (beliefs do not play a role here) $\mathcal{M} = \langle W, \{\sim_i\}_{i \in \mathsf{Agt}}, V\rangle$, let $R_G^D = \cap_{i \in G} \sim_i$, then define

$$\mathcal{M}, w \models D_G\varphi \text{ iff for all } v \in W, \text{ if } wR_G^D v \text{ then } \mathcal{M}, v \models \varphi.$$

Note that $D_G\varphi$ is *not* simply equivalent to $\bigwedge_{i \in G} K_i\varphi$ (the reader is invited to prove this well-known fact). Indeed, the logical analysis has raised a number of interesting technical and conceptual issues (cf. [47, 60, 81, 15]).

3 Informative Actions

The logical models and languages introduced in the previous Section provide *static* descriptions of the situation being modeled. However, many of the situations we are interested in concern agents interacting over time, and this *dynamics* also calls for a logical analysis. Indeed, an important challenge for the logician is to account for the many dynamic processes that govern the agents' social interactions. Inference, observation and communication are all examples of such processes that are the focus of current logics of informational update and belief revision (see, for example, [15, 31, 77][16]). In this Section, we discuss some key issues that appear when shifting from a static to a dynamic perspective.

[15] Of course, this is not meant to be a complete analysis of "social politeness".

[16] Of course, one may argue that (logical) *inference* is the central topic of *any* logic. What we have in mind here is reasoning *about* agents that make inferences.

The main issue is how to incorporate *new* information into an epistemic-doxastic model. At a fixed moment in time the agents are in some *epistemic state* (which may be described by an epistemic(-doxastic) model). The question is how does (the model of) this epistemic state change during the course of some social interaction? The first step towards answering this question is identifying (and formally describing) the *informative* events that shape a particular social interaction. Typical examples include showing one's hand in a card game, make a public or private announcement or sending an email message. However, this step is not always straightforward since the information conveyed by a particular event may depend on many factors which need to be specified. Even the *absence* of an event can trigger a change in an agent's informational state: Recall the famous observation of Sherlock Holmes in *Silver Blaze*: "Is there any point to which you would wish to draw my attention?" "To the curious incident of the dog in the night-time." "The dog did nothing in the night-time." "That was the curious incident," remarked Sherlock Holmes.

Current dynamic epistemic(-doxastic) logics focus on three key issues:

1. The agents' *observational* powers. Agents may perceive the same event differently and this can be described in terms of what agents do or do not observe. Examples range from *public announcements* where everyone witnesses the same event to private communications between two or more agents with the other agents not even being aware that an event took place.

2. The *type* of change triggered by the event. Agents may differ in precisely how they incorporate new information into their epistemic states. These differences are based, in part, on the agents' perception of the *source* of the information. For example, an agent may consider a particular source of information *infallible* (not allowing for the possibility that the source is mistaken) or merely *trustworthy* (accepting the information as reliable though allowing for the possibility of a mistake).

3. The underlying *protocol* specifying which events (observations, messages, actions) are available (or permitted) at any given moment. This is intended to represent the rules or conventions that govern many of our social interactions. For example, in a conversation, it is typically not polite to "blurt everything out at the beginning", as we must speak in small chunks. Other natural conversational protocol rules include "do not repeat yourself", "let others speak in turn", and "be honest". Imposing such rules *restricts* the legitimate sequences of possible statements or events.

A comprehensive theory of rational interaction focuses on the sometimes subtle interplay between these three aspects (cf. [15]).

The most basic type of informational change is a so-called *public announcement* [79, 38]. This is the event where some proposition φ (in the

language of $\mathcal{L}_{EL}$) is made *publicly* available. That is, it is completely open and all agents not only observe the event but also observe everyone else observing the event, and so on *ad infinitum* (cf. item 1 above). Furthermore, all agents treat the source as *infallible* (cf. item 2 above). Thus the effect of such an event on an epistemic(-doxastic) model should be clear: *remove* all states that do not satisfy φ. Formally,

Definition 3.1 (Public Announcement). Suppose $\mathcal{M} = \langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, \{\preceq_i\}_{i\in \mathsf{Agt}}, V\rangle$ is an epistemic-doxastic model and φ is a formula (in $\mathcal{L}_{EDL}$). The model updated by the **public announcement of** φ is the structure $\mathcal{M}^{\varphi} = \langle W^{\varphi}, \{\sim_i^{\varphi}\}_{i\in \mathsf{Agt}}, \{\preceq_i^{\varphi}\}_{i\in \mathsf{Agt}}, V^{\varphi}\rangle$ where $W^{\varphi} = \{w \in W \mid \mathcal{M}, w \models \varphi\}$, for each $i \in \mathsf{Agt}$, $\sim_i^{\varphi} = \sim_i \cap W^{\varphi} \times W^{\varphi}$, $\preceq_i^{\varphi} = \preceq_i \cap W^{\varphi} \times W^{\varphi}$, and for all atomic proposition p, $V^{\varphi}(p) = V(p) \cap W^{\varphi}$. ◁

It is not hard to see that if $\mathcal{M}$ is an epistemic-doxastic model then so is $\mathcal{M}^{\varphi}$. So, the models $\mathcal{M}$ and $\mathcal{M}^{\varphi}$ describe two different moments in time with $\mathcal{M}$ describing the current or initial information state of the agents and $\mathcal{M}^{\varphi}$ the information state *after* the information that φ is true has been incorporated in $\mathcal{M}$. This temporal dimension needs to also be represented in our logical language: let $\mathcal{L}_{PAL}$ extend $\mathcal{L}_{EDL}$ with expressions of the form $[\varphi]\psi$ with $\varphi \in \mathcal{L}_{EDL}$. The intended interpretation of $[\varphi]\psi$ is "ψ is true after the public announcement of φ" and truth is defined as $\mathcal{M}, w \models [\varphi]\psi$ iff if $\mathcal{M}, w \models \varphi$ then $\mathcal{M}^{\varphi}, w \models \psi$.

For the moment, focus only on the agents' hard information and consider the formula $\neg K_i\psi \wedge [\varphi]K_i\psi$: this says that "agent i (currently) does not know ψ but after the announcement of φ, agent i knows ψ". So, the language of $\mathcal{L}_{PAL}$ describes what is true both before and after the announcement. A fundamental insight is that there is a strong logical relationship between what is true before and after an announcement in the form of so-called *reduction axioms*:

$[\varphi]p$	$\leftrightarrow$	$\varphi \rightarrow p$, where $p \in \mathsf{At}$
$[\varphi]\neg\psi$	$\leftrightarrow$	$\varphi \rightarrow \neg[\varphi]\psi$
$[\varphi](\psi \wedge \chi)$	$\leftrightarrow$	$[\varphi]\psi \wedge [\varphi]\chi$
$[\varphi][\psi]\chi$	$\leftrightarrow$	$[\varphi \wedge [\varphi]\psi]\chi$
$[\varphi]K_i\varphi$	$\leftrightarrow$	$\varphi \rightarrow K_i(\varphi \rightarrow [\varphi]\psi)$

These are reduction axioms in the sense that going from left to right either the number of announcement operators is reduced or the complexity of the formulas within the scope of announcement operators is reduced. These reductions axioms provide an insightful syntactic analysis of announcements which complements the semantic analysis. In a sense, the

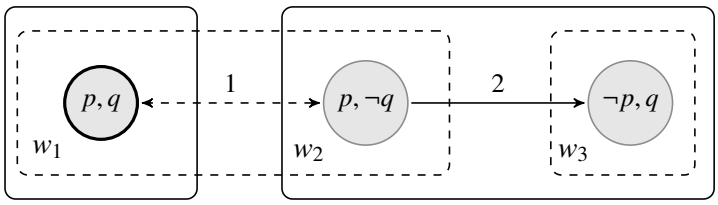
reduction axioms describe the effect of an announcement in terms of what is true before the announcement. By relating pre- and postconditions for each logical operator, the reduction axioms completely characterize the announcement operator.

The above reductions axioms also illustrate the mixture of factual and *procedural* truth that drives conversations or processes of observation (cf. item 3 above). To be more explicit about this point, consider the formula $\langle\varphi\rangle\top$ (with $\langle\varphi\rangle\psi = \neg[\varphi]\neg\psi$ the dual of $[\varphi]$) which means "φ is *announceable*". It is not hard to see that $\langle\varphi\rangle\top \leftrightarrow \varphi$ is derivable using standard modal reasoning and the above reduction axioms. The left-to-right direction represents a semantic fact about public announcements (only true facts can be announced), but the right-to-left direction represents specific *procedural information*: every true formula is available for announcement. But this is only one of many different protocols and different assumptions about the protocol is reflected in a logical analysis. Consider the following variations of the knowledge reduction axiom (cf. [17, Section 4]):

1. $\langle\varphi\rangle K_i\psi \leftrightarrow \varphi \wedge K_i\langle\varphi\rangle\psi$
2. $\langle\varphi\rangle K_i\psi \leftrightarrow \langle\varphi\rangle\top \wedge K_i(\varphi \rightarrow \langle\varphi\rangle\psi)$
3. $\langle\varphi\rangle K_i\psi \leftrightarrow \langle\varphi\rangle\top \wedge K_i(\langle\varphi\rangle\top \rightarrow \langle\varphi\rangle\psi)$

Each of these axioms represent a different assumption about the underlying protocol and how that affects the agents' knowledge. The first is the above reduction axiom (in the dual form) and assumes a specific protocol (which is common knowledge) where all true formulas are always available for announcement. The second (weaker) axiom is valid when there is a fixed protocol that is common knowledge. Finally, the third adds a requirement that the agents must know which formulas are currently available for announcement. Of course, the above three formulas are all *equivalent* given our definition of truth in an epistemic(-doxastic) model (Definition 2.2) and public announcement (Definition 3.1). In order to see a difference, the *protocol information* must be explicitly represented in the model (cf. Section 3.1 and [17]).

We end this introductory Section with a few comments about the effect of a public announcement on the agents' soft information. In particular, it is natural to wonder about the precise relationship between $B_i^{\varphi}\psi$ and $[\varphi]B_i\psi$. *Prima Facie*, the two statements seem to express the same thing; and, in fact, they are equivalent provided ψ is a *ground formula* (i.e., does not contain any modal operators). However, consider state w_1 in the following epistemic-doxastic model:



In this model, the solid lines represent agent 2's hard and soft information (the box is 2's hard information $\sim_2$ and the arrow represent 2's soft information $\preceq_2$) while the dashed lines represent 1's hard and soft information. (Reflexive arrows are not drawn to keep down the clutter in the picture.) Note that at state w_1, agent 2 *knows* p and q (eg., $w_1 \models K_2(p \wedge q)$), and agent 1 believes p but not q ($w_1 \models B_1 p \wedge \neg B_1 q$). Now, although agent 1 does not *know* that agent 2 knows p, agent 1 does believe that agent 2 believes q ($w_1 \models B_1 B_2 q$). Furthermore, agent 1 maintains this belief *conditional on* p: $w_1 \models B_1^p B_2 q$. However, public announcing the true fact p, removes state w_3 and so we have $w_1 \models [p]\neg B_1 B_2 q$. Thus a belief in ψ conditional on φ is *not* the same as a belief in ψ *after* the public announcement of φ. This point is worth reiterating: the reader is invited to check that $B_i^p(p \wedge \neg K_i p)$ is satisfiable but $[!p]B_i(p \wedge \neg K_i p)$ is not satisfiable. The situation is nicely summarized as follows: "$B_i^{\psi}\varphi$ says that if agent i would learn φ then she would come to believe that ψ was the case (before the learning)...$[!\varphi]B_i\psi$ says that after learning φ, agent i would come to believe that ψ is the case (in the worlds after the learning)." [7, pg. 2]. While a public announcement increases the agents' knowledge about the state of the world by reducing the total number of possibilities, it also reveals inaccuracies agents may have about the *other* agents' information. The example above is also interesting because the announcement of a *true* fact misleads agent 1 by forcing her to drop her belief that agent 2 believes q (cf. [15, pg. 182]). Nonetheless, we do have a reduction axiom for conditional beliefs:

$$[\varphi]B_i^{\psi}\chi \leftrightarrow (\varphi \rightarrow B_i^{\varphi \wedge [\varphi]\psi}[\varphi]\chi)$$

What about languages that include group knowledge operators (note that $w_1 \models [p]C_{\{1,2\}}p$)? The situation is much more complex in languages with common knowledge/belief operators. Baltag et al. [3] proved that the extension of $\mathcal{L}_{EL}$ with common knowledge and public announcement operators is strictly more expressive than with common knowledge alone. Therefore a reduction axiom for formulas of the form $[\varphi]C_G\psi$ does not exist. Nonetheless, a reduction axiom-style analysis is still possible, though the details are beyond the scope of this paper (see [19]).

3.1 Two Models of Informational Dynamics

Many different logical systems today describe the dynamics of information over time in a social situation. However, two main approaches can be singled out. The first is exemplified by *epistemic temporal logic* (ETL, [36, 76]) which uses linear or branching time models with added epistemic structure induced by the agents' different capabilities for observing events. These models provide a "grand stage" where histories of some social interaction unfold constrained by a *protocol* (cf., item 3. in the previous Section). The other approach is exemplified by *dynamic epistemic logic* (DEL, [3, 31]) which describes social interactions in terms of epistemic **event models** (which may occur inside modalities of the language). Similar to the way epistemic models are used to capture the (hard) information the agents' have about a *fixed* social situation, an **event model** describes the agents' information about which actual events are currently taking place (cf. item 1 in the previous Section). The temporal evolution of the situation is then computed from some initial epistemic model through a process of successive "product updates". In this Section, we demonstrate each approach by formalizing Example 2.8.

Epistemic Temporal Logic. Fix a finite set of agents $\mathcal{A}$ and a (possibly infinite) set of events[17] Σ. A **history** is a finite sequence of events[18] from Σ. We write Σ^* for the set of histories built from elements of Σ. For a history h, we write he for the history h followed by the event e. Given $h, h' \in \Sigma^*$, we write $h \preceq h'$ if h is a prefix of h', and $h \prec_e h'$ if $h' = he$ for some event e.

For example, consider the social interaction described in Example 2.8. There are three participants: Ann (A), Bob (B) and Ann's friend (call him Charles (C)). What are the relevant primitive events? To keep things simple, assume that Ann's talk is either at 2PM or 3PM and initially none of the agents know this. Say, that Ann receives a message stating that her talk is at 2PM (denote this event — Ann receiving a private message saying that

[17]There is a large literature addressing the many subtleties surrounding the very notion of an *event* and when one event *causes* another event (see, for example, [26]). However, for this paper we take the notion of event as primitive. What is needed is that if an event takes place at some time t, then the fact that the event took place can be observed by a relevant set of agents at t. Compare this with the notion of an event from probability theory. If we assume that at each clock tick a coin is flipped exactly once, then "the coin landed heads" is a possible event. However, "the coin landed head more than tails" would not be an event, since it cannot be observed at any one moment. As we will see, the second statement will be considered a *property* of histories, or sequences of events.

[18]To be precise, elements of Σ should, perhaps, be thought of as event *types* whereas elements of a history are event *tokens*.

her talk is at 2PM — by e_A^{2PM}). Now, after Ann receives the message that the talk is at 2PM, she proceeds to tell her trusted friend Charles that the talk is at 2PM (and that she wants him to inform Bob of the time of the talk without acknowledging that the information can from her — call this event e_C^A), then Charles tells Bob this information (call this event e_B^C). Thus, the history

$$e_A^{2PM}\, e_C^A\, e_B^C$$

represents the sequence of events where "Ann receives a (private) message stating that the talk is at 2PM, Ann tells Charles the talk is at 2PM, then Charles tells Bob the talk is at 2PM". Of course, there are other events that are also relevant to this situation. For one thing, Ann could have received a message stating that her talk is at 3PM (denote this event by e_A^{3PM}). This will be important to capture Bob's uncertainty about whether Ann knows that he knows about the talk. Furthermore, Charles may learn about the time of the talk independently of Ann (denote these two events by e_C^{2PM}, e_C^{3PM}). So, for example, the history

$$e_A^{2PM}\, e_C^{2PM}\, e_B^C$$

represents the situation where Charles independently learns about the time of the talk and informs Bob.

There are a number of simplifying assumptions that we adopt in this section. They are not crucial for the analysis of Example 2.8, but do simplify the some of the formal details. Since, histories are sequences of (discrete) events, we assume the existence of a global discrete clock (whether the agents have access to this clock is another issue that will be discussed shortly). The length of the history then represents the amount of time that has passed. Note that this implies that we are assuming a finite past with a possibly infinite future. Furthermore, we assume that at each clock tick, or moment, *some* event takes place (which need not be an event that any agent directly observes). Thus, we can include an event e_t (for 'clock tick') which can represent that "Charles does *not* tell Bob that the talk is at 2PM." So the history

$$e_A^{2PM}\, e_C^A\, e_t$$

describes the sequence of events where, after learning about the time of the talk, Ann informs Charles, but Charles does *not* go on to tell Bob that the talk is at 2PM. Once a set of events Σ is fixed, the temporal evolution and moment-by-moment uncertainty of the agents can be described.

Definition 3.2 (ETL Models). Let Σ be a set of events and At a set of atomic propositions. A **protocol** is a set $H \subseteq \Sigma^*$ closed under non-empty

prefixes. An **ETL model** is a tuple $\langle \Sigma, \mathsf{H}, \{\sim_i\}_{i\in\mathcal{A}}, V\rangle$ with H a protocol, for each $i \in \mathcal{A}$, an equivalence relation $\sim_i$ on H and V a valuation function ($V : \mathsf{At} \to 2^{\mathsf{H}}$). ◁

An ETL model describes how the agents' *hard* information evolves over time in some social situation. The protocol describes (among other things) the temporal structure, with h' such that $h \prec_e h'$ representing the point in time after e has happened in h. The relations $\sim_i$ represent the uncertainty of the agents about how the current history has evolved. Thus, $h \sim_i h'$ means that from agent i's point of view, the history h' looks the same as the history h.

A protocol in an ETL model captures not only the temporal structure of the social situation but also assumptions about the nature of the participants. Typically, a protocol does not include all *possible* ways a social situation could evolve. This allows us to account for the for the *motivation* of the agents. For example in Example 2.8, the history

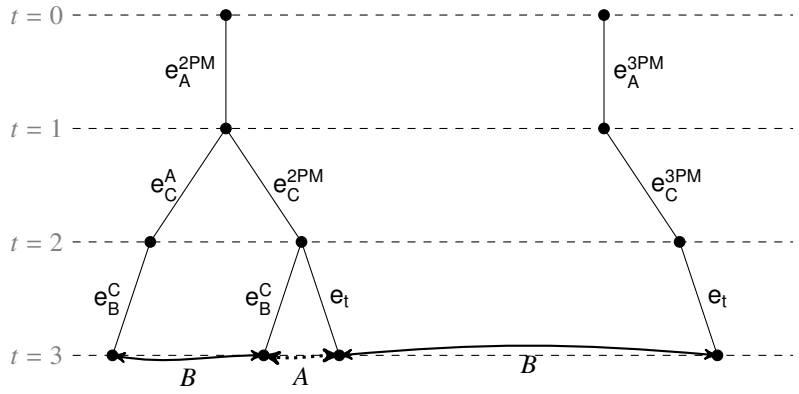
$$\mathsf{e}_{\mathsf{A}}^{\mathsf{3PM}}\ \mathsf{e}_{\mathsf{C}}^{\mathsf{A}}\ \mathsf{e}_{\mathsf{B}}^{\mathsf{C}}$$

describes the sequence of events where Ann learns the talk is at 3PM but tells Charles (who goes on to inform Bob) that the talk is at 2PM. Of course, given that Ann *wants* Bob to attend her talk, this should not be part of (Ann's) protocol. Similarly, since we assume Charles is trustworthy, we should not include any histories where e_{t} follows the event $\mathsf{e}_{\mathsf{C}}^{\mathsf{A}}$. Taking into account these underlying assumptions about the motivations (eg. Ann wants Bob to attend the talk) and dispositions (eg. Charles tells the truth and lives up to his promises) of the agents we can drop a number of histories from the protocol shown above. Note that we keep the history

$$\mathsf{e}_{\mathsf{A}}^{\mathsf{2PM}}\ \mathsf{e}_{\mathsf{C}}^{\mathsf{2PM}}\ \mathsf{e}_{\mathsf{t}}$$

in the protocol, since if Charles learns independently about the time of the talk, then he is under no obligation to inform Bob. In the picture below, we also add some of the uncertainty relations for Ann and Bob (to keep the picture simple, we do not draw the full ETL model). The solid line represents Bob's uncertainty while the dashed line represents Ann's uncertainty. The main assumption is that Bob can only observe the event ($\mathsf{e}_{\mathsf{B}}^{\mathsf{C}}$). So, for example, the histories $h = \mathsf{e}_{\mathsf{A}}^{\mathsf{2PM}}\ \mathsf{e}_{\mathsf{C}}^{\mathsf{A}}\ \mathsf{e}_{\mathsf{B}}^{\mathsf{C}}$ and $h' = \mathsf{e}_{\mathsf{A}}^{\mathsf{2PM}}\ \mathsf{e}_{\mathsf{C}}^{\mathsf{2PM}}\ \mathsf{e}_{\mathsf{B}}^{\mathsf{C}}$ look the same to Bob (i.e., $h \sim_B h'$).[19]

[19] Again we do not include any reflexive arrows in the picture to keep things simple.



Assumptions about the underlying protocol in an ETL model corresponds to "fixing the playground" where the agents will interact. As we have seen, the protocol not only describes the temporal structure of the situation being modeled, but also any *causal* relationships between events (eg., sending a message must always proceed receiving that message) plus the motivations and dispositions of the participants (eg., liars send messages that they *know* — or believe — to be false). Thus the "knowledge" of agent i at a history h in some ETL model is derived from both i's observational powers (via the $\sim_i$ relation) and i's information about the (fixed) protocol.

We give the bare necessities to facilitate a comparison between ETL and DEL. Different modal languages describe ETL models (see, for example, [59, 36]), with 'branching' or 'linear' variants. Let At be a countable set of atomic propositions. The language $\mathcal{L}_{ETL}$ extends the epistemic language $\mathcal{L}_{EL}$ with "event" modalities:

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid \langle e\rangle\varphi$$

where $i \in \mathcal{A}$, $e \in \Sigma$ and $p \in \mathsf{At}$. The boolean connectives ($\vee, \rightarrow, \leftrightarrow$) and the dual modal operators (L_i, $[e]$) are defined as usual. The intended interpretation of '$\langle e\rangle\varphi$' is "after event e (does) take place, φ is true." Formulas are interpreted at histories: Let $\mathcal{H} = \langle \Sigma, \mathsf{H}, \{\sim_i\}_{i\in\mathcal{A}}, V\rangle$ be an ETL model, φ a formula and $h \in \mathsf{H}$, we define $\mathcal{H}, h \models \varphi$ inductively as follows (we only give the modal definitions)

1. $\mathcal{H}, h \models K_i\varphi$ iff for each $h' \in \mathsf{H}$, if $h \sim_i h'$ then $\mathcal{H}, h' \models \varphi$
2. $\mathcal{H}, h \models \langle e\rangle\varphi$ iff there exists $h' \in \mathsf{H}$ such that $h \prec_e h'$ and $\mathcal{H}, h' \models \varphi$

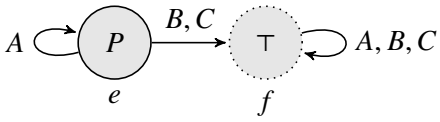
Natural extensions of the $\mathcal{L}_{ETL}$ include group operators (cf. Section 2.3) and more expressive temporal operators (e.g., arbitrary future or past modalities).

Dynamic Epistemic Logic. An alternative account of interactive dynamics was elaborated by [3, 11, 19] and others. From an initial epistemic model, temporal structure evolves, explicitly triggered by complex informative events.

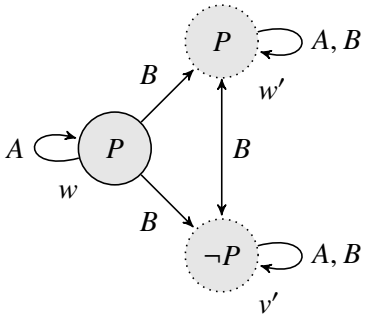
Returning to our running example (Example 2.8), initially we assume that none of the agents knows the time of Ann's talk. Let P be the atomic proposition "Ann's talk is at 2PM." Whereas an ETL model describes the agents' information at all moments, **event models** are used to build new epistemic models as needed.

Definition 3.3 (Event Model). An **event model** is a tuple $\langle S, \{\longrightarrow_i\}_{i\in\mathcal{A}}, \mathsf{pre}\rangle$, where S is a nonempty set of **primitive events**, for each $i \in \mathcal{A}$, $\longrightarrow_i \subseteq S \times S$ and $\mathsf{pre} : S \to \mathcal{L}_{EL}$ is the **pre-condition function**. ◁

Given two primitive events e and f, the intuitive meaning of $e \longrightarrow_i f$ is "if event e takes place then agent i *thinks* it is event f" Event models then describe an "epistemic event". In Example 2.8 the first event is Ann receiving a private message that the talk is at 2PM. This can be described by a simple event model with two primitive events e (with precondition P) and f (with precondition $\top$: f is the "skip" event),



Thus, initially Ann observes the actual event e (and so, learning that P is true) while Bob an Charles observe a skip event (and so, their information does not change). What is the effect of this event on the initial situation (where no one knows the time of the talk)? Intuitively, it is not hard to see that after this event, Ann knows that P while Bob and Charles are still ignorant of *P and the fact that Ann knows P*. That is, incorporating this event into the initial epistemic model should yield (for simplicity we only draw Ann and Bob's uncertainty relations):

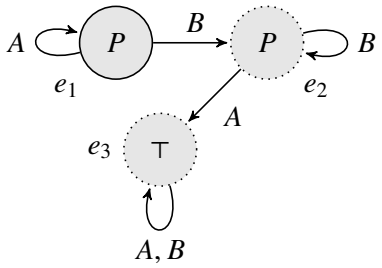


The following definition gives a general procedure for constructing a new epistemic model from a given epistemic model and an event model.

Definition 3.4 (Product Update). The **product update** $\mathcal{M} \otimes \mathcal{E}$ of an epistemic model $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathcal{A}}, V \rangle$ and event model $\mathcal{E} = \langle S, \{\longrightarrow_i\}_{i \in \mathcal{A}}, \mathsf{pre} \rangle$ is the epistemic model $\langle W', R_i', V' \rangle$ with

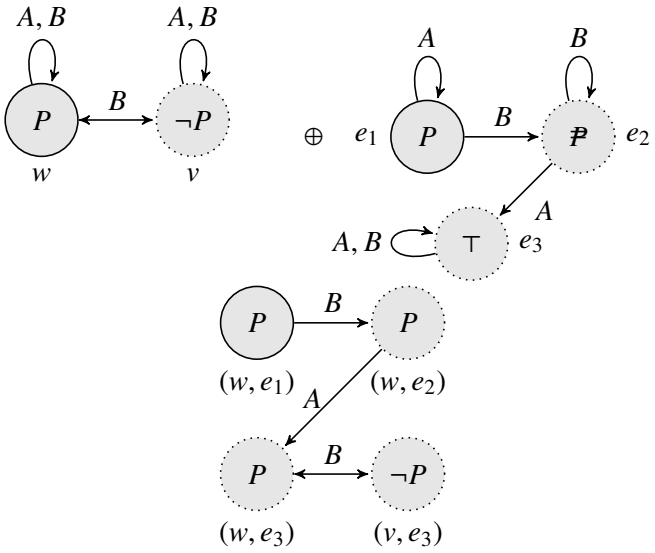
1. $W' = \{(w, e) \mid w \in W, e \in S \text{ and } \mathcal{M}, w \models \mathsf{pre}(e)\}$,
2. $(w, e) R_i' (w', e')$ iff $w R_i w'$ in $\mathcal{M}$ and $e \longrightarrow_i e'$ in $\mathcal{E}$, and
3. For all $P \in \mathsf{At}$, $(s, e) \in V'(P)$ iff $s \in V(P)$ ◁

We illustrate this construction using our running example. The main event in Example 2.8 is "Charles telling Bob (without Ann present) that Ann's talk is at 2PM". This can be described using the following event model (again only the Ann and Bob relations will be drawn): Ann is aware of the actual event taking place while Bob thinks the event is a private message to himself.



As in the previous section, there are implicit assumptions here about the motivations and dispositions of the agents. Thus, even though Ann is not present during the actual event[20], she *trusts* that Charles will honestly tell Bob that the talk is at 2PM (without revealing he received the information from her). This explains why in the above event model, $e_1 \longrightarrow_A e_1$. Starting from a slightly modified epistemic model from the one given above (where Bob now knows that Ann knows *whether* the talk is at 2PM), using Definition 3.4, we can *calculate* the effect of the above event model as follows:

[20] Of course, we must assume that she knows precisely *when* Charles will meet with Bob.



Again, for simplicity, not all the reflexive arrows are drawn.

Finally, a few comments about syntactic issues. The language $\mathcal{L}_{DEL}$ extends $\mathcal{L}_{EL}$ with operators $\langle \mathcal{E}, e \rangle$ for each pair of event models $\mathcal{E}$ and event e in the domain of $\mathcal{E}$. Truth is defined as usual: We only give the typical DEL modalities:

$$\mathcal{M}, w \models \langle \mathcal{E}, e \rangle \varphi \text{ iff } \mathcal{M}, w \models \mathsf{pre}(e) \text{ and } \mathcal{M} \otimes \mathcal{E}, (w, e) \models \varphi$$

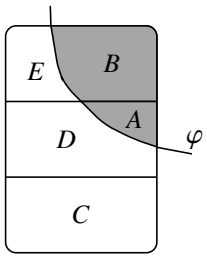
Remark 3.5. *We conclude by noting that the public announcement of the previous Section is a special case of Definition 3.3. Given a formula $\varphi \in \mathcal{L}_{EL}$, the public announcement is the event model $\mathcal{E}_\varphi = \langle \{e\}, \{\longrightarrow_i\}_{i \in \mathcal{A}}, \mathsf{pre} \rangle$ where for each $i \in \mathcal{A}$, $e \longrightarrow_i e$ and $\mathsf{pre}(e) = \varphi$. As the reader is invited to verify, the product update of an epistemic model $\mathcal{M}$ with a public announcement event $\mathcal{E}_\varphi$ ($\mathcal{M} \otimes \mathcal{E}_\varphi$) is (isomorphic) to the model $\mathcal{M}^\varphi$ of Definition 3.1.*

3.2 Varieties of Informational Change

The dynamic models discussed in the previous Section focus on the agents' observational powers and procedural information. The assumption is that precisely how an agent incorporates new information depends on only two factors: what the agent has observed and the underlying protocol (which is typically assumed to be common knowledge). To what degree the agent *trusts* the source of the information is not taken into account (cf. item 2 from Section 3). In this section, we show how to extend our logical analysis

with this information. We only have the space here for some introductory remarks: see [15, Chapter 7] and [8] for more extensive discussions.

The general problem we focus on in the Section is how to incorporate the evidence that φ is true into an epistemic-doxastic model $\mathcal{M}$. The approach taken so far is to *eliminate* all worlds inconsistent with (each agent's observation of) the evidence that φ is true. (This may reveal *more* than φ is true given an underlying protocol). However, not all sources of evidence are 100% reliable opening the door to the possibility that later evidence may contradict earlier evidence. Consider the situation from agent i's point-of-view: Abstractly, the problem is how to define a *new ordering* over i's (hard) information cell given i's current soft information (represented as a total ordering over the set of states that i considers possible) and the incoming information represented as the *truth* set of some formula φ:



Rather than *removing* the states inconsistent with φ (in the above case, this would be the states in the set $C \cup D \cup E$), the goal is to *rearrange* the states in such a way that φ is believed. In the above example, this means that at least the set A should become the new minimal set. But there is a variety ways to fill in the rest of the order with each way corresponding to a different "policy" the agent takes towards the incoming information [82]. We only have space here to discuss two of these policies (both have been widely discussed in the literature, see for example, [15, Chapter 7]). The first captures the situation where the agent only tentatively accepts the incoming information φ by making the best φ the new minimal set and keeping the rest of the ordering the same. Before formally defining the policy we need some notation: given an epistemic-doxastic model $\mathcal{M}$, let $best_i(\varphi, w) = Min_{\leq_i}([w]_i \cap \{x \mid \mathcal{M}, x \models \varphi\})$ denote the best φ worlds at state w.

Definition 3.6 (Conservative Upgrade). Given an epistemic-doxastic model $\mathcal{M} = \langle W, \{\sim_i\}_{i\in \mathsf{Agt}}, \{\leq_i\}_{i\in \mathsf{Agt}}, V\rangle$ be an epistemic-doxastic model and a formula φ, the *conservative* upgrade of $\mathcal{M}$ with φ is the model $\mathcal{M}^{\uparrow\varphi} = \langle W^{\uparrow\varphi}, \{\sim_i^{\uparrow\varphi}\}_{i\in \mathsf{Agt}}, \{\leq_i^{\uparrow\varphi}\}_{i\in \mathsf{Agt}}, V^{\uparrow\varphi}\rangle$ with $W^{\uparrow\varphi} = W$, for each i, $\sim_i^{\uparrow\varphi} = \sim_i$, $V^{\uparrow\varphi} = V$ and for all $i \in \mathsf{Agt}$ and $w \in W^{\uparrow\varphi}$ we have:

1. If $v \in best_i(\varphi, w)$ then $v \prec_i^{\uparrow\varphi} x$ for all $x \in [w]_i$, and
2. for all $x, y \in [w]_i - best_i(\varphi, w)$, $x \preceq_i^{\uparrow\varphi} y$ iff $x \preceq_i y$. ◁

In the above picture a conservative upgrade with φ results in the new ordering $A \prec_i C \prec_i D \prec_i B \cup E$. A logical analysis of this type of information change includes formulas of the form $[\uparrow_i\varphi]\psi$ intended to mean "after i's conservative upgrade of φ, ψ is true" and interpreted as follows: $\mathcal{M}, w \models [\uparrow_i\varphi]\psi$ iff $\mathcal{M}^{\uparrow_i\varphi}, w \models \psi$. We also have reduction axioms for conditional beliefs:

$$[\uparrow\varphi]B^{\psi}\chi \leftrightarrow (B^{\varphi}\neg[\uparrow\varphi]\psi \wedge B^{[\uparrow\varphi]\psi}[\uparrow\varphi]\chi) \vee (\neg B^{\varphi}\neg[\uparrow\varphi]\psi \wedge B^{\varphi\wedge[\uparrow\varphi]\psi}[\uparrow\varphi]\chi)$$

(We leave out the i subscripts to make the formula easier to read). The reader is invited to check the validity of the above axiom. The second policy we introduce here models a more "radical" change to the agent's plausibility ordering: *all* φ worlds are moved ahead of all other worlds. Thus, rather than focusing on only the best φ worlds, the agent shifts *all* φ worlds consistent with i's current information: let $[\![\varphi]\!]_i^w = \{x \mid \mathcal{M}, x \models \varphi\} \cap [w]_i$ denote this set of φ worlds:

Definition 3.7 (Radical Upgrade). Given an epistemic-doxastic model $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathsf{Agt}}, \{\preceq_i\}_{i\in\mathsf{Agt}}, V\rangle$ be an epistemic-doxastic model and a formula φ, the *conservative* upgrade of $\mathcal{M}$ with φ is the model $\mathcal{M}^{\Uparrow\varphi} = \langle W^{\Uparrow\varphi}, \{\sim_i^{\Uparrow\varphi}\}_{i\in\mathsf{Agt}}, \{\preceq_i^{\Uparrow\varphi}\}_{i\in\mathsf{Agt}}, V^{\Uparrow\varphi}\rangle$ with $W^{\Uparrow\varphi} = W$, for each i, $\sim_i^{\Uparrow\varphi} = \sim_i$, $V^{\Uparrow\varphi} = V$ and for all $i \in \mathsf{Agt}$ and $w \in W^{\Uparrow\varphi}$ we have:

1. for all $x \in [\![\varphi]\!]_i^w$ and $y \in [\![\neg\varphi]\!]_i^w$, set $x \prec_i^{\Uparrow\varphi} y$,
2. for all $x, y \in [\![\varphi]\!]_i^w$, set $x \preceq_i^{\Uparrow\varphi} y$ iff $x \preceq_i y$, and
3. for all $x, y \in [\![\neg\varphi]\!]_i^w$, set $x \preceq_i^{\Uparrow\varphi} y$ iff $x \preceq_i y$. ◁

In the above picture a conservative upgrade with φ results in the new ordering $A \prec_i B \prec_i C \prec_i D \prec_i E$. A logical analysis of this type of information change includes formulas of the form $[\Uparrow_i\varphi]\psi$ intended to mean "after i's radical upgrade of φ, ψ is true" and interpreted as follows: $\mathcal{M}, w \models [\Uparrow_i\varphi]\psi$ iff $\mathcal{M}^{\Uparrow_i\varphi}, w \models \psi$. As the reader is invited to check, the conservative upgrade is a special case of this radical upgrade: the conservative upgrade of φ at w is the radical upgrade of $best_i(\varphi, w)$. In fact, both of these operations can be seen as instances of a more general *lexicographic update* (cf. [15, Chapter 7]). In fact, the above reduction axiom for conservative upgrade can be *derived* from the following reduction axiom for radical upgrade: (again, we leave out the i subscripts to make the formula easier to read)

$$[\Uparrow\varphi]B^{\psi}\chi \leftrightarrow (L(\varphi \wedge [\Uparrow\varphi]\psi) \wedge B^{\varphi\wedge[\Uparrow\varphi]\psi}[\Uparrow\varphi]\chi) \vee (\neg L(\varphi \wedge [\Uparrow\varphi]\psi) \wedge B^{[\Uparrow\varphi]\psi}[\Uparrow\varphi]\chi)$$

4 Conclusions

Agents are faced with many diverse tasks as they interact with the environment and one another. At certain moments, they must *react* to their (perhaps surprising) observations while at other moments they must be *proactive* and choose to perform a specific action. One central underlying assumption is that "rational" agents obtain what they want via the implementation of (successful) *plans* (cf. [25]). And this implementation often requires, among other things, representation of various informational attitudes of the other agents involved in the social interaction. In social situations there are many (sometimes competing) *sources* for these attitudes: for example, the type of "communicatory event" (public announcement, private announcement), the disposition of the other participants (liars, truth-tellers) and other implicit assumptions about procedural information (reducing the number of possible histories). This naturally leads to different notions of "knowledge" and "belief" that drive social interaction.

An overarching theme in is this paper is that during a social interaction, the agents' "knowledge" and "beliefs" both influence *and* are shaped by the *social* events. The following example taken from [72] illustrates this point. Suppose that Uma is a physician whose neighbour Sam is ill and consider the following cases

Case 1: . Uma does not know and has not been informed. Uma has no obligation (as yet) to treat her neighbour.

Case 2: The neighbour's daughter Ann comes to Uma's house and tells her. Now Uma does have an obligation to treat Sam, or perhaps call in an ambulance or a specialist.

In both of these cases, the issue of an obligation arises. This obligation is circumstantial in the sense that in other situations, the obligation might not apply. If Sam is ill, Uma needs to know that he is ill, and the nature of the illness, but not where Sam went to school. Thus an agent's obligations are often dependent on what the agent knows, and indeed one cannot reasonably be expected to respond to a problem if one is not aware of its existence. This, in turn, creates a secondary obligation on Ann to inform Uma that her father is ill.

Based on the logical framework discussed in Section 3.1 and [61], Pacuit et al. [72] develop a logical framework that formalizes the reasoning of Uma and Ann in the above example. It is argued that this reasoning is shaped by the assumption that Uma and Ann's preferences are aligned (i.e., both want Sam to get better). For example, Ann will not be under any

obligation to tell Uma that her father is ill, if Ann justifiably believes that Uma would not treat her father even if she knew of his illness. Thus, in order for Ann to *know* that she has an obligation to tell Uma about her father's illness, Ann must *know* that "Uma will, in fact, treat her father (in a reasonable amount of time) upon learning of his illness". More formally, in all the histories that Ann currently considers possible, the event where her father is treated for his illness is always preceded by the event where she tells Uma about his illness. That is, the histories where Uma learns of Sam's illness but does not treat him are not part of the protocol. Similar reasoning is needed for Uma to derive that she has an obligation to treat Sam. Obviously, if Uma has a good reason to believe that Ann always lies about her father being ill, then she is under no obligation to treat Sam. See [72] for a formal treatment of these examples.

This paper surveyed a number of logical systems that model the reasoning and dynamic processes that govern many of our social interactions. This is a well-developed area attempting to balance sophisticated logical analysis with philosophical insight. Furthermore, the logical systems discussed in this paper have been successfully used to sharpen the analysis of key epistemic issues in a variety of disciplines. However, they represent only one component of a logical analysis of *rational interaction*. Indeed, as the above example illustrates, a comprehensive account of rational interaction cannot always be isolated from other aspects of rational agency and social interaction – such as the motivational attitudes of the agent or her social obligations.

References

[1] Aumann, R. Interactive epistemology I: Knowledge. *International Journal of Game Theory 28* (1999), 263–300.

[2] Aumann, R. Interactive epistemology II: Probability. *International Journal of Game Theory 28* (1999), 301 – 314.

[3] Baltag, A., Moss, L., and Solecki, S. The logic of common knowledge, public announcements and private suspicions. In *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)* (1998), I. Gilboa, Ed., pp. 43 – 56.

[4] Baltag, A., and Smets, S. Conditional doxastic models: A qualitative approach to dynamic belief revision. In *Proceedings of WOLLIC 2006, Electronic Notes in Theoretical Computer Science* (2006), G. Mints and R. de Queiroz, Eds., vol. 165, pp. 5 – 21.

[5] Baltag, A., and Smets, S. From conditional probability to the logic of doxastic actions. In *TARK '07: Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge* (2007), ACM, pp. 52–61.

[6] Baltag, A., and Smets, S. The logic of conditional doxastic actions. In *New Perspectives on Games and Interaction*, R. van Rooij and K. Apt, Eds. Texts in Logic and Games, Amsterdam University Press, 2008.

[7] Baltag, A., and Smets, S. A qualitative theory of dynamic interactive belief revision. In *Logic and the Foundation of Game and Decision Theory (LOFT7)* (2008), G. Bonanno, W. van der Hoek, and M. Wooldridge, Eds., vol. 3 of *Texts in Logic and Games*, Amsterdam University Press, pp. 13–60.

[8] Baltag, A., and Smets, S. ESSLLI 2009 course: Dynamic logics for interactive belief revision. Slides available online at `http://alexandru.tiddlyspot.com/#%5B%5BESSLLI09%20COURSE%5D%5D`, 2009.

[9] Barwise, J. Three views of common knowledge. In *TARK '88: Proceedings of the 2nd conference on Theoretical aspects of reasoning about knowledge* (San Francisco, CA, USA, 1988), Morgan Kaufmann Publishers Inc., pp. 365–379.

[10] Battigalli, P., and Siniscalchi, M. Strong belief and forward induction reasoning. *Journal of Economic Theory 105* (2002), 356 – 391.

[11] van Benthem, J. 'One is a lonely number': on the logic of communication. In *Logic Colloquium '02* (2002), Z. Chatzidakis, P. Koepke, and W. Pohlers, Eds., ASL and A. K. Peters, pp. 96 – 129. Available at http://staff.science.uva.nl/~johan/Muenster.pdf.

[12] van Benthem, J. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics 14*, 2 (2004), 129 – 155.

[13] van Benthem, J. Rational animals: What is 'KRA'? invited lecture Malaga ESSLLI Summer School 2006, 2005.

[14] van Benthem, J. Modal frame correspondences and fixed-points. *Studia Logica 83*, 1-3 (2006), 133–155.

[15] van Benthem, J. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.

[16] van Benthem, J. *Modal Logic for Open Minds*. CSLI Publications, 2010.

[17] van Benthem, J., Gerbrandy, J., Hoshi, T., and Pacuit, E. Merging frameworks of interaction. *Journal of Philosophical Logic 38*, 5 (2009), 491 – 526.

[18] van Benthem, J., and Sarenac, D. The geometry of knowledge. In *Aspects of Universal Logic,* (2004), vol. 17, pp. 1–31.

[19] van Benthem, J., van Eijck, J., and Kooi, B. Logics of communication and change. *Information and Computation 204*, 11 (2006), 1620 – 1662.

[20] Blackburn, P., de Rijke, M., and Venema, Y. *Modal Logic*. Campbridge University Press, 2002.

[21] Board, O. Dynamic interactive epistemology. *Games and Economic Behavior 49* (2004), 49 – 80.

[22] Bonanno, G. On the logic of common belief. *Mathematical Logical Quarterly 42* (1996), 305 – 311.

[23] Boutilier, C. *Conditional Logics for Default Reasoning and Belief Revision*. PhD thesis, University of Toronto, 1992.

[24] Brandenburger, A. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory 35* (2007), 465–492.

[25] Bratman, M. *Intention, Plans and Practical Reason*. Harvard University Press, London, 1987.

[26] Cartwright, N. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, 2007.

[27] Chwe, M. S.-Y. *Rational Ritual*. Princeton University Press, 2001.

[28] Clark, H., and Marshall, C. R. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*, Joshi, Webber, and Sag, Eds. Cambridge University Press, 1981.

[29] Cubitt, R. P., and Sugden, R. Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory. *Economics and Philosophy 19*, 2 (2003), 175–210.

[30] van Ditmarsch, H. Prolegomena to dynamic logic for belief revision. *Synthese: Knowledge, Rationality, and Action 147* (2005), 229 – 275.

[31] van Ditmarsch, H., van der Hoek, W., and Kooi, B. *Dynamic Epistemic Logic*. Springer, 2007.

[32] Egré, P., and Bonnay, D. Inexact knowledge with introspection. *Journal of Philosophical Logic 38*, 2 (2009), 179 – 228.

[33] Engelfriet, J., and Venema, Y. A modal logic of information change. In *Proceedings of TARK 1998* (1998), pp. 125–131.

[34] Fagin, R., and Halpern, J. Reasoning about knowledge and probability. *Journal of the ACM 41*, 2 (1994), 340 – 367.

[35] Fagin, R., Halpern, J., and Megiddo, N. A logic for reasoning about probabilities. *Information and Computation 87*, 1 (1990), 78 – 128.

[36] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. *Reasoning about Knowledge*. The MIT Press, 1995.

[37] Gabbay, D. and Woods, J. (Editors) *The Handbook of the History of Logic: Logic and Modalities in the Twentieth Century*. Volume 7, Elsevier, 2006.

[38] Gerbrandy, J. *Bisimulations on Planet Kripke*. PhD thesis, Institute for Logic, Language and Computation (DS-1999-01), 1999.

[39] Gochet, P., and Gribomont, P. Epistemic logic. In *[37]*. Elsevier, 2006.

[40] Goldblatt, R. Mathematical modal logic: A view of its evolution. In *[37]*. Elsevier, 2006.

[41] Goldman, A. What is justified belief? In *Justification and Kowledge*, G. Pappas, Ed. D. Reidel, 1976.

[42] Halpern, J. Should knowledge entail belief? *Journal of Philosophical Logic 25*, 5 (1996), 483 – 494.

[43] Halpern, J. Set-theoretic completeness for epistemic and conditional logic. *Annals of Mathematics and Artificial Intelligence 26* (1999), 1 – 27.

[44] Halpern, J. *Reasoning about Uncertainty*. The MIT Press, 2003.

[45] Halpern, J., and Lakemeyer, G. Levesque's axiomatization of only knowing is incomplete. *Artificial Intelligence 74*, 2 (1995), 381 – 387.

[46] Halpern, J., and Lakemeyer, G. Multi-agent only knowing. *Journal of Logic and Computation 11* (2001), 41 – 70.

[47] Halpern, J., and Moses, Y. Knowledge and common knowledge in a distributed environment. *ACM-PODC* (1983), 50 – 61.

[48] Halpern, J., and Pass, R. A logical characterization of iterated admissibility. In *TARK '09: Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge* (New York, NY, USA, 2009), ACM, pp. 146–155.

[49] Halpern, J., and Pucella, R. Modeling adversaries in a logic for security protocol analysis. In *Formal Aspects of Security* (2003).

[50] Harman, G. *Change in View*. MIT Press, 1986.

[51] Harsanyi, J. C. Games with incompletet information played by bayesian players parts I-III. *Management Sciences 14* (1967).

[52] Heifetz, A., and Mongin, P. Probability logic for type spaces. *Games and Economic Behavior 35* (2001), 31 – 53.

[53] Hendricks, V. *Mainstream and Formal Epistemology*. Cambridge University Press, 2005.

[54] Hendricks, V. Editor, special issue: "8 bridges between formal and mainstream epistemology". *Philosophical Studies 128*, 1 (2006), 1–227.

[55] Hendricks, V., and Symons, J. Where's the bridge? epistemology and epistemic logic. *Philosophical Studies 128* (2006), 137 – 167.

[56] Herzig, A. Modal probability, belief, and actions. *Fundam. Inf. 57*, 2-4 (2003), 323–344.

[57] Hintikka, J. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.

[58] Hintikka, J. *Knowledge and Belief: An Introduction to the Logic of the Two Notions (with an Introduction by V. Hendricks and J. Symons)*. King's College Publications, 2005.

[59] Hodkinson, I., and Reynolds, M. Temporal logic. In *Handbook of Modal Logic*. forthcoming.

[60] van der Hoek, W., van Linder, B., and Meyer, J.-J. Group knowledge is not alwasy distributed (neither is it always implicity). *Mathematical Social Sciences 38* (1999), 215 – 240.

[61] Horty, J. *Agency and Deontic Logic*. Oxford University Press, 2001.

[62] Humberstone, L. The modal logic of 'all and only'. *Notre Dame Journal of Formal Logic 28* (1987), 177 – 188.

[63] Lamarre, P., and Shoham, Y. Knowledge, certainty, belief and conditionalisation. In *Proceedings of the International Conference on Knoweldge Representation and Reasoning* (1994), pp. 415 – 424.

[64] Levesque, H. J. All I know: a study in autoepistemic logic. *Artificial Intelligence 42*, 3 (1990), 263 – 309.

[65] Lewis, D. *Convention*. Harvard University Press, 1969.

[66] Lismont, L., and Mongin, P. On the logic of common belief and common knowledge. *Theory and Decision 37*, 1 (1994), 75 – 106.

[67] Lismont, L., and Mongin, P. Strong Completeness Theorems for Weak Logics of Common Belief. *Journal of Philosophical Logic 32*, 2 (2003), 115 – 137.

[68] Liu, F. *Changing for the better: Preference dynamics and agent diversity.* PhD thesis, Institute for logic, lanugage and computation (ILLC), 2008.

[69] Lomuscio, A., and Ryan, M. On the relation between interpreted systems and kripke models. In *Proceedings of the AI97 Workshop on Theoretical and Practical Foundation of Intelligent Agents and Agent-Oriented Systems* (1997), vol. LNCS 1441.

[70] Nozick, R. Knowledge and skepticism. In *Philosophical Investigations*. The MIT Press, 1981, pp. 172 – 185.

[71] Pacuit, E. Some comments on history based structures. *Journal of Applied Logic Applied Logic 5*, 4 (2007), 613–624.

[72] Pacuit, E., Parikh, R., and Cogan, E. The logic of knowledge based obligation. *Knowledge, Rationality and Action: A Subjournal of Synthese 149*, 2 (2006), 311 – 341.

[73] Parikh, R. Social software. *Synthese 132* (2002), 187–211.

[74] Parikh, R. Levels of knowledge, games, and group action. *Research in Economics 57* (2003), 267 — 281.

[75] Parikh, R. Sentences belief and logical omniscience or what does deduction tell us? *Review of Symbolic Logic 1*, 4 (2008), 514 – 529.

[76] Parikh, R., and Ramanujam, R. Distributed processes and the logic of knowledge. In *Logic of Programs* (1985), vol. 193 of *Lecture Notes in Computer Science*, Springer, pp. 256 – 268.

[77] Parikh, R., and Ramanujam, R. A knowledge based semantics of messages. *Journal of Logic, Language and Information 12* (2003), 453 – 467.

[78] Plaza, J. Logics of Public Communications. *Synthese: Knowledge, Rationality and Action 158(2)*, (2007), 165 - 179.

[79] Plaza, J. Logics of public communications. In *Proceedings, 4th International Symposium on Methodologies for Intelligent Systems* (1989), M. L. Emrich, M. S. Pfeifer, M. Hadzikadic, and Z. Ras, Eds., pp. 201–216 (republished as [78]).

[80] Ramanujam, R., and Suresh, S. Deciding knowledge properties of security protocols. In *Proceedings of Theoretical Aspects of Rationality and Knowledge* (2005), pp. 219–235.

[81] Roelofsen, F. Distributed knowledge. *Journal of Applied Non-Classical Logics 17*, 2 (2007), 255 – 273.

[82] Rott, H. Shifting priorities: Simple representations for 27 itereated theory change operators. In *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg* (2006), H. Lagerlund, S. Lindström, and R. Sliwinski, Eds., vol. 53 of *Uppsala Philosophical Studies*, pp. 359 – 384.

[83] Samuelson, L. Modeling knowledge in economic analysis. *Journal of Economic Literature 57* (2004), 367 – 403.

[84] Schwitzgebel, E. Belief. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., fall 2008 ed. 2008.

[85] Shah, N., and Velleman, J. Doxastic deliberation. *The Philosophical Review 114*, 4 (2005), 497 – 534.

[86] Shoham, Y., and Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[87] Sosa, E., Kim, J., Fantl, J., and McGrath, M., Eds. *Epistemology: An Anthology*. Wily-Blackwell, 2008.

[88] Stalnaker, R. The problem of logical omniscience I. *Synthese 89* (1991), 425 – 440.

[89] Stalnaker, R. On the evaluation of solution concepts. *Theory and Decision 37*, 42 (1994).

[90] Stalnaker, R. Extensive and strategic forms: Games and models for games. *Research in Economics 53* (1999), 293 – 319.

[91] Stalnaker, R. On logics of knowledge and belief. *Philosophical Studies 128* (2006), 169 – 199.

[92] Vanderschraaf, P., and Sillari, G. Common knowledge. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., spring 2009 ed. 2009.

[93] Williamson, T. *Knowledge and its Limits*. Oxford University Press, 2000.

[94] von Wright, G. H. *An Essay in Modal Logic*. North-Holland, Amsterdam, 1951.

Belief Change

RICHARD BOOTH* and THOMAS MEYER†

Abstract

In this paper we present a brief overview of *belief change*, a research area concerned with the question of how a rational agent ought to change its mind in the face of new, possibly conflicting, information. We limit ourselves to *logic-based* belief change, with a particular emphasis on classical propositional logic as the underlying logic in which beliefs are to be represented. Our intention is to provide the reader with a basic introduction to the work done in this area over the past 30 years. In doing so we hope to sketch the main historical results, provide appropriate pointers to further references, and discuss some current developments. We trust that this will spur on the interested reader to learn more about the topic, and perhaps to join us in the further development of this exciting field of research.

1 Introduction

Consider the situation in which an agent has just encountered a bird: let's call it Tweety. Part of the agent's beliefs about the world is that birds fly. Being a logical agent, it therefore believes that Tweety flies. On closer inspection, though, the agent learns that Tweety is an ostrich. Since the agent also believes that ostriches don't fly, it is now faced with a dilemma: Can Tweety fly, or can't it?

The simple scenario above aptly illustrates the central topic of this paper—that a rational intelligent agent is sometimes forced to adjust its current beliefs in some appropriate fashion when confronted with new information. The investigation of the reasoning patterns involved in such a task is known as the study of *belief change*.

The approach to the problem of belief change that we discuss in this paper is logic-based. Both the beliefs of an agent and new information presented to it will be represented in a logic language with a strong emphasis on the case where the underlying logic is a classical propositional

*Univ. of Luxembourg Luxembourg

†Meraka Institute, CSIR and School of Computer Science, Univ. of Kwazulu-Natal, South Africa

logic. Although much of the early work on belief change has a somewhat weaker assumption about the underlying logic, requiring simply that it be a logic equipped with a Tarskian consequence relation and satisfying Compactness, the usual assumption in practice was to use a propositional logic. However, as we shall soon see, propositional logic on its own is not enough to obtain unique answers to the problems of belief change. The primary principle we shall use to guide us is known as the *Principle of Minimal Change*. The idea is simple and intuitive. Information is hard to come by and if an agent has gone to the trouble of incorporating a piece of information into its set of beliefs, it has presumably done so for a good reason. It should therefore give up any beliefs it has only if it is forced to do so. That is, any changes to its current stock of beliefs should be minimal.

Traditionally, approaches to belief change have followed one of two trajectories, with the differences centred around the question of whether beliefs should be represented as *belief bases* (arbitrary sets of sentences) or logically closed *theories*. The AGM approach to belief change [1, 28] (named after its originators Alchourrón, Gärdenfors and Makinson), perhaps the most influential voice within this field of research, is based on the assumption that beliefs need to be represented as theories. The idea here is that we are interested in belief change on the *knowledge level* and that the particular syntactic formulation that we choose for representing the beliefs of an agent is largely irrelevant. On the other hand, the case made for the use of belief bases, which originated with the work of Sven Ove Hansson [34], is that the sentences chosen to represent the beliefs of an agent are somehow more basic than those that merely follow logically from these basic sentences. Although these two approaches start off with different, seemingly conflicting basic assumptions, we shall see that they actually have much in common. In fact, one of the assumptions underlying both approaches is the necessity of introducing additional structure to the representation of beliefs in order to obtain unique results to specific problems in belief change. In much of the work on this topic the additional structure is not represented in the underlying logic itself, but is viewed as meta-information of some kind, and our work here strongly emphasises that approach. Having said that, it is important to note that there is a growing body of work in which information such as this is incorporated into the logic itself—a topic which we touch on in Section 8.

Our intention in this paper is to provide the reader with a basic introduction to the work done in the area of belief change over the past 30 years. In doing so we hope to sketch the main historical results, provide appropriate pointers to further references, and discuss some current developments. Of course, it is impossible to present a truly comprehensive account of a

research area in a paper such as this and our perspective on matters will invariably be subjective, to some extent. The reader is urged to keep this in mind when going through the paper.

On to more concrete matters, then. We commence with a discussion of the formal preliminaries needed to digest the rest of the paper in Section 2. This is followed in Sections 3 and 4 by accounts of the two basic operators investigated in belief change: *belief contraction* and *belief revision*. In Section 5 we take a closer look at the semantic methods for constructing belief change operators before we use this approach to consider *iterated belief revision* in Section 6. Section 7 discusses the links between belief change and the area of *nonmonotonic reasoning*, while Section 8 considers approaches to belief change using *epistemic logics*. Finally, Section 9 takes a brief look at recent developments in belief change before we conclude in Section 10.

2 Preliminaries

First the logical framework. We start with a quite abstract formulation (L, Cn), where we just have a set L whose elements are the *sentences*, together with a consequence operator Cn which takes sets of sentences $B \subseteq L$ to sets of sentences $Cn(B)$ which intuitively represents all the sentences which are *entailed* by B. Cn is assumed to be a compact *Tarskian* consequence operator (after Alfred Tarski), i.e., it satisfies the following four properties for all $B, B_1, B_2 \subseteq L$:

- $B \subseteq Cn(B)$ **(Reflexivity)**
- $B_1 \subseteq B_2$ implies $Cn(B_1) \subseteq Cn(B_2)$ **(Monotony)**
- $Cn(Cn(B)) = Cn(B)$ **(Idempotence)**
- If $\varphi \in Cn(B)$ then $\varphi \in Cn(B')$ for some finite $B' \subseteq B$ **(Compactness)**

We call any arbitrary set $B \subseteq L$ a *belief base*, but if $B = Cn(B)$, i.e., B is *closed* under Cn, then we call B a *theory.* Following the tradition of the AGM approach, we will use K rather than B to denote theories. $\alpha \in L$ is a tautology iff $\alpha \in Cn(\emptyset)$. From Cn we can define a notion of *consistency*. A set B of sentences is Cn-consistent iff $Cn(B) \neq L$. We will just say consistent if the consequence operator is clear from the context.

Definition 1. *An* abstract deduction system *is a pair* (L, Cn) *as above.*

This logical setup, although very general, is surprisingly already rich enough to explore many interesting issues in belief change. However, traditionally researchers (including AGM) have worked within more specific background logical systems. In particular the machinery of propositional logic is usually taken as minimum. We may take $L = L_P$, consisting of all sentences built up from some set of propositional variables using the connectives $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$. The classical propositional consequence operator is denoted by Cn_0. We call a *supraclassical deduction system* a pair (L_P, Cn) where in addition to the four properties mentioned above, Cn is assumed to satisfy

- $Cn_0(B) \subseteq Cn(B)$ **(Supraclassicality)**
- $\varphi \in Cn(B \cup \{\theta\})$ iff $(\theta \rightarrow \varphi) \in Cn(B)$ **(Deduction)**

For supraclassical deduction systems we have $B \cup \{\alpha\}$ is consistent iff $\neg\alpha \notin Cn(B)$. In particular α is consistent iff $\neg\alpha \notin Cn(\emptyset)$.

2.1 The problem formalised

We are now ready to state formally the problem of belief revision:

> Assume some fixed abstract deduction system (L, Cn) as background. Then given an initial belief base $B \subseteq L$ and some new information represented as a sentence $\alpha \in L$, find a new belief base $B * \alpha$ which includes α and is consistent.

The requirement that $B * \alpha$ be consistent is crucial here. Without it we might as well just add α set-theoretically to B and stop there. But if $B \cup \{\alpha\}$ is inconsistent then entailment is trivialised, since by definition of inconsistency then *all* sentences are entailed by $B \cup \{\alpha\}$, rendering it useless. So how should we approach this problem?

One influential idea, which comes from Isaac Levi [48] is to decompose the operation into two main steps. First, B is altered if necessary so as to "make room" for, i.e., become consistent with, the incoming sentence α. This is achieved by making B deductively weaker. This is known as *contraction*. Here we should adhere to the principle of minimal change, according to which this weakening should be made as "small" as possible. (See [62] for a discussion of this principle.) In the second, trivial step, the new formula is then simply joined on to the result (this is known as *expansion*). Clearly the difficult step is the first one. So in order to answer the problem of revision we first need to address the problem of contraction. We turn to this in the next section.

3 Belief contraction

Note that for the purposes of revision we just need to make B consistent with α. In the case where the background deduction system is supraclassical this is the same as ensuring $\neg\alpha \notin Cn(B)$. But in general (for instance if negation is not available in the language) these two things will be different. So in general there are two kinds of contraction operator: the first is *inconsistency-based* and the second is *entailment-based*.[1] We will focus on entailment-based contraction here. We denote the result of contracting B so that it no longer entails a given sentence α by $B - \alpha$. We will first deal with the case where B is an arbitrary belief base. Later we will look at the special case where it is a theory.

3.1 Partial meet base contraction

One of the best-known approaches to contraction is *partial meet contraction* [1, 37]. Here the idea is to calculate contraction in three steps:

1. Focus for the first step on those subsets of B which do not entail α and which are *maximal* with this property. We denote this set by $B \perp \alpha$.
2. Then, a certain number of the elements of this set are somehow selected as the "best" or "most preferred" by means of a selection function γ: $\gamma(B \perp \alpha) \subseteq B \perp \alpha$.
3. Finally, the intersection of these best elements is taken: $\bigcap \gamma(B \perp \alpha)$

Let's formalise all this, starting with the set $B \perp \alpha$ in step 1.

Definition 2. *Let $B \subseteq L$ and $\alpha \in L$. Then $B \perp \alpha$ is the set of subsets $X \subseteq L$ such that $X \in B \perp \alpha$ iff* (i). *$X \subseteq B$,* (ii). *$\alpha \notin Cn(X)$,* (iii). *For all $X' \subseteq B$, if $X \subset X'$ then $\alpha \in Cn(X')$. We call $B \perp \alpha$ the set of* α-remainders of B

If α is a tautology then $B \perp \alpha = \emptyset$, but this is the only case for which $B \perp \alpha = \emptyset$. This is a result of the following fact, the proof of which requires **Monotony** and **Compactness** of Cn as well as Zorn's Lemma.

Fact 3 ([2]). *If $\alpha \notin Cn(Y)$ and $Y \subseteq B$ then there exists $X \in B \perp \alpha$ such that $Y \subseteq X$.*

In other words, every non-α-implying subset of B may be extended to a maximal non-α-implying subset of B. Next comes the definition of selection function.

[1] See also Section 9.

Definition 4. *Let $B \subseteq L$. A* selection function for B *is a function γ such that for all $\alpha \in L$,* (i) *if $B \perp \alpha \neq \emptyset$ then $\emptyset \neq \gamma(B \perp \alpha) \subseteq B \perp \alpha$, and* (ii) *if $B \perp \alpha = \emptyset$ then $\gamma(B \perp \alpha) = \{B\}$.*

Finally we can use a selection function for B to define a contraction operator $-_\gamma$ for B:

$$B -_\gamma \alpha = \bigcap \gamma(B \perp \alpha).$$

Definition 5. *If $-$ can be defined via some selection function γ for B as above then $-$ is a* partial meet base contraction *operator (for B).*

Two special cases of partial meet contraction deserve mention. If the selection function picks a single element of $B \perp \alpha$, it is called a *maxichoice contraction*. If it picks the whole of $B \perp \alpha$, it is called a *full meet contraction*. Observe that full meet contraction is unique, whereas there are many different maxichoice contractions: one for each element of $B \perp \alpha$.

Partial meet base contraction may be characterised as follows.

Theorem 6 ([36]). *$-$ is a partial meet base contraction operator for B iff it satisfies the following properties:*

- *If $\alpha \notin Cn(\emptyset)$ then $\alpha \notin Cn(B - \alpha)$* ***(Success)***
- *$B - \alpha \subseteq B$* ***(Inclusion)***
- *If $\beta \in B \setminus B - \alpha$ then there exists B' such that $B - \alpha \subseteq B' \subseteq B$, $\alpha \notin Cn(B')$ and $\alpha \in Cn(B' \cup \{\beta\})$* ***(Relevance)***
- *If for all $B' \subseteq B$ we have $\alpha \in Cn(B')$ iff $\beta \in Cn(B')$ then $B - \alpha = B - \beta$* ***(Uniformity)***

The above properties may be explained as follows. **Success** says the sentence to be removed *is* actually removed, i.e., is no longer a consequence of the base[2] and **Inclusion** states that no new beliefs may be added in the course of removing α.[3] **Relevance** seeks to avoid unnecessary loss of information. It says that a sentence β should be given up only if it contributes to the fact that B, and not $B - \alpha$, entails α. **Uniformity** states that if two sentences are indistinguishable from the viewpoint of B, in that every subset of B which implies one also implies the other, then the results of contracting by them should be the same. As was noted in [40], the only properties of Cn which are actually used in the proof of Theorem 6 are **Monotony** and **Compactness**.

The following two reasonable properties can be shown to follow from those in Theorem 6 (see [34]), and thus are satisfied by any partial meet base contraction operator.

[2] But see [23] for a discussion on why this is not always desirable.

[3] See [10] for an argument against this postulate.

- If $\alpha \notin Cn(B)$ then $B - \alpha = B$ ***(Vacuity)***
- If $Cn(\alpha) = Cn(\beta)$ then $B - \alpha = B - \beta$ ***(Preservation)***

Vacuity says that if α is not entailed by B to begin with, then nothing needs to be changed. It can be shown to be a consequence of **Relevance** and **Inclusion**. **Preservation** says that if two sentences are equivalent under logical consequence then contracting by them should give the same results. It can be shown to follow from **Uniformity**, while in the special case when B is a theory, it is actually equivalent to **Uniformity** in the presence of **Vacuity**.

3.2 Partial meet theory contraction

The preceding construction works equally well when B is taken to be a theory K. But in this case, since the input to contraction is a theory, we should expect the output to be a theory too. This is ensured because in this case the elements of $K \perp \alpha$ are themselves theories, and the intersection of any family of theories is again a theory. When applied to a theory K we will refer to the above construction as *partial meet theory contraction*.

In this case we obtain a different representation theorem, which was one of the main results of AGM.[4]

Theorem 7 ([1]). *Assume we work with a supraclassical deduction system* (L_P, Cn)*, and let* K *be a theory. Then* $-$ *is a partial meet theory contraction operator for* K *iff it satisfies* ***Success****,* ***Inclusion****,* ***Vacuity****,* ***Preservation*** *and the following properties:*

- $K - \alpha = Cn(K - \alpha)$ ***(Closure)***
- $K \subseteq Cn((K - \alpha) \cup \{\alpha\})$ ***(Recovery)***

Note that this result requires the assumption of a supraclassical deduction system as background. It may not hold for general abstract deduction systems (see [24] for discussion on this).

The postulates listed in the above theorem are collectively known as the *basic AGM contraction postulates*. **Closure** says that the result of theory contraction is another theory, while **Recovery** says that if one removes α and then simply adds it again (and then closes under logical consequence) then one should get back all the initial beliefs K. **Recovery** has been by far the most controversial of the AGM contraction postulates, with many authors calling it into question (see [34, pp. 72-74]). It should be noted that

[4]Note that, historically, partial meet theory contraction actually pre-dates the more general version for arbitrary bases given above.

this postulate is specific to the *theory* version of partial meet contraction, i.e., it does not hold in general for partial meet *base* contraction, where B is allowed to be an arbitrary base. For supraclassical deduction systems, in the presence of **Closure**, **Inclusion** and **Vacuity** it is equivalent to **Relevance** [25].

3.3 The supplementary postulates

In partial meet contraction, when selecting the remainder sets via γ, what we have is an instance of a *choice situation.* We have a number of alternatives up for selection, namely $B \perp \alpha$, and some of them are singled out as being in some sense more preferred. So we make some crossover into the realm of *rational choice.* How can this choice be made? We can assume it is made on the basis of a binary preference relation $\sqsubseteq$ over the set of all possible remainder sets for B, i.e., the set $\{X \mid X \in B \perp \alpha \text{ for some } \alpha \in L\}$. For any two possible remainder sets X, Y, we write $X \sqsubseteq Y$ to mean that Y is at least as preferred as X, and use $\sqsubset$ to denote the strict part of $\sqsubseteq$, i.e., $X \sqsubset Y$ iff both $X \sqsubseteq Y$ and $Y \not\sqsubseteq X$. Then $\sqsubseteq$ can be used as the basis for a selection function $\gamma_{\sqsubseteq}$ by setting, in the principal case in which $B \perp \alpha \neq \emptyset$,

$$\gamma_{\sqsubseteq}(B \perp \alpha) = \{X \in B \perp \alpha \mid Y \sqsubseteq X \text{ for all } Y \in B \perp \alpha\}.$$

That is, $\gamma_{\sqsubseteq}(B \perp \alpha)$ consists of those elements of $B \perp \alpha$ which are at least as preferred as *all* other elements of $B \perp \alpha$. If a selection function γ for B is generated from some relation $\sqsubseteq$ in this way then we say γ is a *relational* selection function, and a partial meet contraction operator $-_\gamma$ which can be generated from some relational selection function γ will be called a *relational partial meet contraction* operator.

By putting some mild constraints on the relation $\sqsubseteq$, we can constrain the behaviour of the resulting relational partial meet contraction operator in interesting ways. Consider the following two properties:

- $X \sqsubseteq Y$ and $Y \sqsubseteq Z$ implies $X \sqsubseteq Z$ **(Transitivity)**
- If $X \subset Y$ then $X \sqsubset Y$ **(Maximising)**

The first property is a standard requirement for a relation of preference. The second is motivated by minimal change considerations: when contracting B it is always preferable to retain as much of B as possible, so a given subset Y of B should always be strictly more preferred to any of its strict subsets. If γ is generated from some $\sqsubseteq$ satisfying **Transitivity** then we say $-_\gamma$ is a *transitively relational* partial meet contraction operator, while if it generated from some $\sqsubseteq$ satisfying, in addition, **Maximising** then $-_\gamma$ is a *transitively, maximisingly relational* partial meet contraction.

Note that, if B is a theory, then these two collapse into the same thing, in the sense that $-_\gamma$ is transitively relational iff it is transitively, maximisingly relational [33].

We obtain the following results, both of which assume we work in a supraclassical deduction system:

Theorem 8 ([37]). *Assume a supraclassical deduction system as background. Let B be a belief base and suppose $-$ is a transitively, maximisingly relational partial meet base contraction operator for B. Then $-$ satisfies the following property:*

- $(B - \alpha) \cap (B - \beta) \subseteq B - (\alpha \wedge \beta)$ ***(Conjunctive Overlap)***

For the case of relational partial meet *theory* contraction we can say more:

Theorem 9 ([1]). *Assume a supraclassical deduction system as background. Let K be a theory and $-$ an operator for K. Then $-$ is a transitively relational partial meet theory contraction operator for K iff it satisfies all the basic AGM contraction postulates (see Theorem 7) plus* ***Conjunctive Overlap*** *and the following property:*

- *If $\alpha \notin K - (\alpha \wedge \beta)$ then*

  $K - (\alpha \wedge \beta) \subseteq K - \alpha$ ***(Conjunctive Inclusion)***

The postulates **Conjunctive Overlap** and **Conjunctive Inclusion** are known as the *AGM supplementary contraction postulates*. They go a step beyond the basic postulates, in that they relate the results of contracting by conjunctions $\alpha \wedge \beta$ with the result of contracting by the individual conjuncts. We refer the interested reader to [1, 34] for discussion on these postulates.

3.4 Kernel contraction

The partial meet approach to contraction focusses on *maximal* subsets of B which do not imply the sentence α to be removed. Another approach is instead to single out the *minimal* subsets which *do* entail α, and then to make sure at least one sentence is removed from each. This is the idea behind Hansson's operation of *kernel contraction* [38], which is a generalisation of the *safe contraction* of Alchourrón and Makinson [3].

Definition 10. *Let B be a belief base and $\alpha \in L$. Then $B \perp\!\!\!\perp \alpha$ is the set of sets X such that* (i). *$X \subseteq B$,* (ii). *$\alpha \in Cn(X)$,* (iii). *If $X' \subset X$ then $\alpha \notin Cn(X')$. We call the elements of $B \perp\!\!\!\perp \alpha$ the* α-kernels of B.

To remove α, it is necessary and sufficient to remove at least one sentence from every α-kernel. To this end, we assume the existence of an function which makes an "incision" into every such set, returning the sentences which must be discarded.

Definition 11. *σ is an* incision function for *B if* (i). $\sigma(B \perp\!\!\!\perp \alpha) \subseteq \bigcup B \perp\!\!\!\perp \alpha$, *and* (ii). $X \in B \perp\!\!\!\perp \alpha$ *implies* $\sigma(B \perp\!\!\!\perp \alpha) \cap X \neq \emptyset$.

Every incision function σ then yields a contraction operator by setting $B -_\sigma \alpha = B \setminus \sigma(B \perp\!\!\!\perp \alpha)$.

Definition 12. *Let $-$ be an operator for B. If $-$ equals $-_\sigma$ for some incision function σ for B then it is called a* kernel base contraction operator *(for B).*

Kernel base contraction may be characterised in the following way:

Theorem 13 ([38]). *$-$ is a kernel base contraction operator for B iff it satisfies **Success**, **Inclusion**, **Uniformity** and the following property:*

- *If $\beta \in B \setminus B - \alpha$ then there exists B' such that $B' \subseteq B$, $\alpha \notin Cn(B')$ and $\alpha \in Cn(B' \cup \{\beta\})$* ***(Core retainment)***

As with Theorem 6, it was noted in [40] that the only properties required to prove this result are **Monotony** and **Compactness**. Note that **Core retainment** is weaker than **Relevance** and so we see that every partial meet base contraction operator is a kernel base contraction operator. The converse, however, does not hold, i.e., there exist kernel base contraction operators which are not partial meet base contraction operators (see [34, p91]) for a counterexample, and [22] for more on the relation between partial meet and kernel base contraction).

The above discussion was about kernel *base* contraction. It is also possible to employ the construction in the case when B is a theory except, since $K -_\sigma \alpha$ is not guaranteed to be a theory (even when K is), it is necessary to add a post-processing step of closing under Cn. That is, a kernel theory contraction operator for K is any operator of the form $K \approx_\sigma \alpha = Cn(K -_\sigma \alpha)$, where σ is an incision function for K and $-_\sigma$ is defined from σ as for kernel base contraction. However, in this case (at least for supraclassical deduction systems), the distinction between kernel theory contraction and partial meet theory contraction disappears, in that every kernel theory contraction operator is a partial meet theory contraction operator, and vice versa [38].

4 Belief Revision

As stated earlier, once we have a contraction operation for a belief base B, we can use it to define a revision operator via the Levi Identity. The Levi Identity comes in two flavours, according to whether we want the result of revision to be a theory or not. In the former case we take $B * \alpha = (B - \neg\alpha) \cup \{\alpha\}$, in the latter case we take $K * \alpha = Cn((K - \neg\alpha) \cup \{\alpha\})$. We call the former the non-closing Levi Identity, and the latter the closing Levi Identity. Whenever we talk of the Levi Identity in connection with an arbitrary belief base B we shall implicitly assume it is the non-closing version we are using, while if we use it in connection with a theory K, we shall assume the closing version.[5] Throughout this section we shall assume, for simplicity, that we work in a supraclassical deduction system.

First we deal with arbitrary belief bases, where the result is not expected to be logically closed.

Definition 14. *Let B be a belief base. If $*$ can be defined from some partial meet base contraction operator for B using the (non-closing) Levi Identity then it is a* partial meet base revision *operator for B.*

Partial meet base revision may be characterised as follows:

Theorem 15 ([37]). *$*$ is an operation of partial meet base revision for B iff it satisfies the following properties:*

- $\alpha \in B * \alpha$ ***(Success)***
- *If α is consistent then $B * \alpha$ is consistent* ***(Consistency)***
- $B * \alpha \subseteq B \cup \{\alpha\}$ ***(Inclusion)***
- *If $\beta \in B \setminus B * \alpha$ then there exists B' such that $B * \alpha \subseteq B' \subseteq B \cup \{\alpha\}$, B' is consistent and $B' \cup \{\beta\}$ is inconsistent.* ***(Relevance)***
- *If, for all $B' \subseteq B$, we have $B' \cup \{\alpha\}$ is consistent iff $B' \cup \{\beta\}$ is consistent, then $B \cap (B * \alpha) = B \cap (B * \beta)$* ***(Uniformity)***

Success and **Consistency** are taken as fundamental requirements here.[6] **Inclusion** places an upper-bound on the result of revision. It says the result should not contain any sentence not included in B, apart from the new

[5] The Levi Identity breaks revision by α into two steps: contraction (by $\neg\alpha$) and expansion (by α), in **that** order. Another possibiity is to reverse this order and do the expansion (by α) first, followed by the contraction (by $\neg\alpha$). This possibility is explored in [37].

[6] Although **Success** is not beyond controversy, since one can certainly imagine situations in which new information is not accepted. See [39, 35] for studies of revision operators which don't satisfy it.

information α. **Relevance** and **Uniformity** are similar to their namesakes in the list of base contraction postulates.

Let us move on to theory revision.

Definition 16. *Let K be a theory. If $*$ can be defined from some partial meet theory contraction operator for K via the (closing) Levi Identity, then $*$ is a* partial meet theory revision *operator for K.*

The following result is the AGM characterisation of partial meet theory revision.

Theorem 17 ([1]). *$*$ is a partial meet theory revision operator for a theory K iff it satisfies **Success**, **Consistency** and the following basic AGM postulates for theory revision:*

- $K * \alpha = Cn(K * \alpha)$ ***(Closure)***
- $K * \alpha \subseteq Cn(K \cup \{\alpha\})$ ***(Inclusion)***
- *If $K \cup \{\alpha\}$ is consistent then $K * \alpha = Cn(K \cup \{\alpha\})$* ***(Vacuity)***
- *If $Cn(\alpha) = Cn(\beta)$ then $K * \alpha = K * \beta$* ***(Preservation)***

Furthermore, $*$ is a transitively relational partial meet theory revision operator (i.e, is defined via the Levi Identity from some transitively relational partial meet contraction operator) iff it satisfies, in addition, the following two supplementary AGM revision postulates:

- $K * (\alpha \wedge \beta) \subseteq Cn((K * \alpha) \cup \{\beta\})$ ***(Subexpansion)***
- *If $(K * \alpha) \cup \{\beta\}$ is consistent then $Cn((K * \alpha) \cup \{\beta\}) \subseteq K * (\alpha \wedge \beta)$* ***(Superexpansion)***

Observe that, for the remainder of this paper we will take *AGM revision* to mean transitively relational partial meet revision.

It is also possible to use the Levi Identity to define revision from kernel contraction, leading to *kernel revision* operators. Axiomatic characterisations are given in [40], and we refer the reader to that paper for details.

Finally in this section, while the Levi Identity deals with how to define revision in terms of contraction, it is also possible to go the other way and define contraction in terms of revision by using the *Harper Identity* [41]:

$$B - \alpha = B \cap (B * \neg\alpha).$$

The Levi and Harper identities can be thought of as inverses to each other. They ensure a very tight connection between contraction and revision.

5 On the semantic side

The previous sections have been developed against the background of some given, fixed abstract (sometimes supraclassical) deduction system (L, Cn), which represents the background logic we work in. These systems can be said to be *syntactical*, in the sense that they simply declare (via Cn) which sentences are entailed by which sentences. There is, of course, usually another side to logic which is the *semantical* side. It is the semantics of a logic which tells us what are the objects, or *possible worlds*, or *models* which the sentences in L are actually *talking about*. In this section we investigate belief change from a more semantical viewpoint. The ideas behind this approach originate in a famous paper by Adam Grove [31]. For this and the next section we make a number of simplifying assumptions: *(i)* we assume that we are working in a supraclassical deduction system (L_{P}, Cn), *(ii)* we furthermore assume L_{P} is generated by only finitely many propositional atoms, and *(iii)* we will focus only on *theory* revision and contraction.

What are the models in a supraclassical deduction system? One way to define them is as the set of *maximally consistent theories of* L_{P}.

$$\mathcal{W} \stackrel{\text{def}}{=} \{M \subseteq L_{\mathrm{P}} \mid M \text{ is a } Cn\text{-consistent theory and for no } Cn\text{-consistent theory } M' \subseteq L \text{ do we have } M \subset M'\}.$$

Given $M \in \mathcal{W}$ and $B \subseteq L_{\mathrm{P}}$, we say M is *a model of* B iff $B \subseteq M$. Then the set of models of B is denoted by $[B]$.

The set $\mathcal{W}$ defines a consequence relation $Cn_{\mathcal{W}}$ by setting, for any $B \subseteq L_{\mathrm{P}}$, $Cn_{\mathcal{W}}(B) = \bigcap[B]$, i.e., a sentence is entailed by B iff it is contained in all models of B. Then $\mathcal{W}$ provides a semantics which is sound and complete with respect to (L_{P}, Cn), in the sense that, for any $B \subseteq L_{\mathrm{P}}$, the identity $Cn_{\mathcal{W}}(B) = Cn(B)$ holds.

Now suppose we have a theory K representing our initial beliefs. This corresponds to the belief that the actual "true" world is one of the worlds in $[K]$. It turns out that performing transitively relational partial meet theory contraction on K is equivalent to choosing, on the basis of some *total preorder* over the set $\mathcal{W}$, some countermodels of (i,e, models of the negation of) the sentence to be contracted, and adding them to $[K]$. To be more precise, let $\leq$ be a total preorder[7], or *tpo* for short, over $\mathcal{W}$. For $M_1, M_2 \in \mathcal{W}$, $M_1 \leq M_2$ may be informally read as "M_1 is at least as plausible (as a candidate to be the real world) as M_2". Given any subset

[7] A binary relation $\leq$ over a set S is a *total preorder* iff it is *(i)* reflexive, i.e., $s \leq s$ for all $s \in S$, *(ii)* transitive, and *(iii)* connected, i.e., either $s \leq t$ or $t \leq s$ for all $s, t \in S$.

$T \subseteq \mathcal{W}$, we denote by $\min_{\leq}(T)$ the minimal elements of T under $\leq$, i.e., $\min_{\leq}(T) = \{t \in T \mid t \leq t' \text{ for all } t' \in T\}$. We assume $\leq$ is *anchored on* $[K]$, i.e., $[K] = \min_{\leq}(\mathcal{W})$. Then we may use $\leq$ to define a contraction operator for K as follows:

$$K -_{\leq} \alpha = \begin{cases} K & \text{if } \alpha \in Cn(\emptyset) \\ K \cap \bigcap \min_{\leq}([\neg\alpha]) & \text{otherwise.} \end{cases}$$

In other words, the models of the new theory are obtained by taking the minimal models of $\neg\alpha$ and adding them to the models of K.

Theorem 18 ([31]). *Let K be a theory. Then $-$ is a transitively relational partial meet theory contraction operator for K iff $-$ equals $-_{\leq}$ for some tpo over $\mathcal{W}$ which is anchored on $[K]$.*

This is not the only way we could use a plausibility order to define a contraction operator. Rott and Pagnucco introduced and axiomatically characterised the operation of *severe withdrawal* [64].

$$K-\alpha = \begin{cases} K & \text{if } \alpha \in Cn(\emptyset) \\ \bigcap \{M \in \mathcal{W} \mid M \leq M' \text{ for some } M' \in \min_{\leq}([\neg\alpha])\} & \text{otherwise.} \end{cases}$$

Here, the models of the new belief set are obtained by taking *all* models which are at least as plausible as the $\leq$-minimal $\neg\alpha$-models. This operation was independently proposed, using a different construction, by Isaac Levi under the name *mild contraction* [49]. Unlike partial meet theory contraction, severe withdrawal does not satisfy **Recovery**. Yet another possibility was explored by Meyer et al. [50]. *Systematic withdrawal* is just like severe withdrawal except we add to $[K]$ not only the most plausible $\neg\alpha$-models, but all models which are *strictly* more plausible than them.

$$K - \alpha = \begin{cases} K & \text{if } \alpha \in Cn(\emptyset) \\ K \cap \bigcap \{M \in \mathcal{W} \mid M < M' \text{ for some } M' \in \min_{\leq}(\neg\alpha)\} \cap \bigcap \min_{\leq}([\neg\alpha]) & \text{otherwise.} \end{cases}$$

What about defining revision from a plausibility order $\leq$? We may just apply the Levi Identity to each of the three families of contraction operator above. It turns out we get the same revision operator in each case, viz.

$$K *_{\leq} \alpha = \begin{cases} L & \text{if } \neg\alpha \in Cn(\emptyset) \\ \bigcap \min_{\leq}([\alpha]) & \text{otherwise} \end{cases}$$

In other words the models of the new theory which results from revising by α are exactly $\leq$-minimal α-models.

Theorem 19 ([31]). *Let K be a theory. Then $*$ is a transitively relational partial meet theory revision operator for K iff $*$ equals $*_{\leq}$ for some tpo over $\mathcal{W}$ which is anchored on $[K]$.*

In the theory case, we have that the AGM postulates are equivalent to total preorders over the set of models. There is a third way of characterising AGM theory contraction, namely as an ordering of entrenchment among the sentences of the language [54, 55]. The best-known version of such entrenchment orderings is the *epistemic entrenchment orderings* put forward by Gärdenfors and Makinson [28, 26]. We do not these in details here, but rather refer the interested reader to the references provided.

6 Iterated theory revision as revising tpos

Everything in the preceding sections has been about "one-shot" belief change. There is an initial theory, there is some new input and then there is a new theory. However, in realistic settings, a rational agent does not simply "shut down" after incorporating this input, but must be ready to receive the *next* input, followed by further inputs after that. That is to say, belief change is an iterative process, and any theory of belief change worthy of the name should be able to account for this. The question is, then, does the theory sketched in the previous sections adequately handle iterated changes? The answer, as researchers began to realise in the mid 1990s, is "no".

What does the theory described until now have to say about iterated belief change? Notice that the extra structure required to carry out revision, be it incision functions, selection functions, or total preorders over models is always defined *relative to the theory which is being changed*. Thus, for example, when using the tpo construction, there is a fixed total preorder $\leq_K$ associated to each different theory K. So, to revise a theory K by a sentence φ, we can use the total preorder $\leq_K$ associated to K to compute the result $K * \varphi$. If we then want to further revise this new theory by ψ, then we use the tpo $\leq_{K*\varphi}$ associated to it to compute $(K * \varphi) * \psi$. There are three, interrelated problems with this:

1. There need be hardly any relation between the successive tpos $\leq_K$ and $\leq_{K*\varphi}$, where intuitively we might expect some.
2. Some intuitively plausible properties of iterated revision may be violated (see below).
3. This method totally disregards the role that "revision history" may play in determining results of belief change.

What researchers realised in the mid 1990s is that, to address these shortcomings, the theory of belief change should be widened so that it deals not only with change on the level of theories, but that it should address change in the very structure used to change those theories. A contraction or revision operator should tell us not only what the new theory should be, but should also provide us with a new selection function/incision function/tpo over models which is then the target for the next input. In fact most the best-known approaches to iterated change deal with tpos rather than the other ways of modelling the extra-structure. Furthermore the focus in this area tends to be more on revision than contraction (but see [19, 17, 18, 57, 42]) so in the following we focus on iterated theory revision as a problem of revising tpos.

6.1 Revising total preorders

So given K and a total preorder $\leq$ associated to K, the result of revision should be a new theory $K * \varphi$ *together with* a new associated tpo $\leq_{K*\varphi}$. However we can simplify a bit, since the tpo associated to any theory contains enough information to recapture the theory anyway (since $[K] = \min_{\leq}(\mathcal{W})$). So, our new revision problem may be formulated as follows:

> Given an initial tpo $\leq$ over $\mathcal{W}$, and revision input α, determine a new tpo $\leq^*_\alpha$ over $\mathcal{W}$.

The theory should extend the foregoing theory of single-step revision, which means the new belief set $K(\leq^*_\alpha)$ should be derived from the initial tpo and α using the partial meet revision recipe from Theorem 19. This means that the new lowest level $\min_{\leq^*_\alpha}(\mathcal{W})$ in the new tpo is determined already - it is equal to $\min_{\leq}([\alpha])$. But what about the rest of the ordering? The most obvious thing to do, if we want to be motivated by the principle of minimal change, is to simply leave the rest of the ordering untouched, and sure enough, this was one of the first proposals for tpo revision. Boutilier called it *Natural Revision* [15, 16], though the idea dates back to [69]. Formally it is defined as follows:

$$M_1 \leq^{*_B}_\alpha M_2 \text{ iff } \begin{cases} \textit{either} & M_1 \in \min_{\leq}([\alpha]) \\ \textit{or} & M_1, M_2 \notin \min_{\leq}([\alpha]) \text{ and } M_1 \leq M_2. \end{cases}$$

The problem with natural revision is that it makes *too few* changes. This was recognised by Darwiche and Pearl, who proposed four postulates for regulating tpo revision [20]:

(CR1) If $M_1, M_2 \in [\alpha]$ then $M_1 \leq^*_\alpha M_2$ iff $M_1 \leq M_2$

(CR2) If $M_1, M_2 \in [\neg\alpha]$ then $M_1 \leq^*_\alpha M_2$ iff $M_1 \leq M_2$

(CR3) If $M_1 \in [\alpha]$ and $M_2 \in [\neg\alpha]$ and $M_1 \leq M_2$ then $M_1 \leq^*_\alpha M_2$

(CR4) If $M_1 \in [\alpha]$ and $M_2 \in [\neg\alpha]$ and $M_1 < M_2$ then $M_1 <^*_\alpha M_2$

(CR1) and **(CR2)** say that, when revising $\leq$ by α, the relative ordering of models within $[\alpha]$, respectively within $[\neg\alpha]$, should remain unchanged. **(CR3)** and **(CR4)** say that if a given α-model was judged to be at least as (respectively strictly more) plausible as a given $\neg\alpha$-model before revising by α, then that relation should be preserved after the revision. Essentially revising by α should not cause any degradation in plausibility of any α-model with respect to the $\neg\alpha$-models.

As noted by Darwiche and Pearl themselves, the above postulates do not rule out natural revision as a sensible approach to tpo revision, because $*_B$ satisfies all these postulates. However $*_B$ does not satisfy the following strengthening of **(CR3)** and **(CR4)**, which was suggested independently in [11, 45]:

(CR5) If $M_1 \in [\alpha]$ and $M_2 \in [\neg\alpha]$ and $M_1 \leq M_2$ then $M_1 <^*_\alpha M_2$

(CR5) forces there to be a *strict* increase in plausibility of the α-models in relation to the $\neg\alpha$-models which were not deemed more plausible to begin with.

The above postulates can be repackaged as postulates constraining the theory following a double revision:

(C1) If $\alpha \in Cn(\beta)$ then $K\left((\leq^*_\alpha)^*_\beta\right) = K\left(\leq^*_\beta\right)$

(C2) If $\neg\alpha \in Cn(\beta)$ then $K\left((\leq^*_\alpha)^*_\beta\right) = K\left(\leq^*_\beta\right)$

(C3) If $\alpha \in K\left(\leq^*_\beta\right)$ then $\alpha \in K\left((\leq^*_\alpha)^*_\beta\right)$

(C4) If $\neg\alpha \notin K\left(\leq^*_\beta\right)$ then $\neg\alpha \notin K\left((\leq^*_\alpha)^*_\beta\right)$

(C5) If $\neg\alpha \notin K\left(\leq^*_\beta\right)$ then $\alpha \in K\left((\leq^*_\alpha)^*_\beta\right)$

(C1) says if two inputs arrive, the second entailing the first, then the first can be ignored when calculating the resulting theory. **(C2)** says if two contradictory inputs arrive, then the effects of the first are cancelled out. **(C3)** and **(C4)** say that if α would be believed, resp. not rejected, after receiving β alone, then this should not change if β were to be preceded by an input α. Finally **(C5)** postulates a condition under which belief in an input α is guaranteed to survive the arrival of a subsequent input β.

Theorem 20 ([11, 20, 45]). *Let $*$ be a tpo revision operator such that always $K(\leq^*_\alpha) = \bigcap \min_\leq([\alpha])$. Then, for each $i = 1, 2, 3, 4, 5$, $*$ satisfies **(CR**i**)** iff it satisfies **(C**i**).***

A few concrete tpo revision operators have been proposed which satisfy all of the above postulates. For example in *lexicographic revision* [56, 69] the new tpo following input α is determined by placing *all* α-models strictly below all $\neg\alpha$-models while leaving the relative ordering within the sets $[\alpha]$ and $[\neg\alpha]$ unchanged. This is a most radical form of tpo revision, where the new information α is given total priority over the initial ordering $\leq$. At the opposite end of the spectrum is *restrained revision* [11], in which the strict part of the initial ordering is preserved (apart from the minimal α-models, which become strictly more plausible than all the other models), with α-models being promoted only ahead of the $\neg\alpha$-models which were on the same plausibility "level" (see also [63]).

7 Belief revision and nonmonotonic reasoning

In this section we discuss the connections between belief revision and the work done in the nonmonotonic reasoning community. A logic is said to be *nonmonotonic* if its associated entailment relation $\mid\!\sim$ need not satisfy the following monotonicity property: if $A \mid\!\sim \beta$ then $A \cup \{\alpha\} \mid\!\sim \beta$. With $\mid\!\sim$ seen as a relation of plausible consequence, there are many examples to show that monotonicity is an undesirable property. Perhaps the one most deeply entrenched in the nonmonotonic reasoning literature is the Tweety example (the example we used in the introduction). Given that Tweety is a bird, it seems plausible to infer that Tweety can fly. But given the additional evidence that Tweety is an ostrich, we should abandon our conclusion about Tweety's flying capabilities.

While there are many approaches to nonmonotonic reasoning (see e.g., [61, 53]), we consider here the influential framework proposed by Kraus, Lehmann, and Magidor [47] and show that it has a strong connection with AGM belief revision. Formally, Kraus et al. take $\mid\!\sim$ to be a binary relation on sentences of a propositional logic where $\alpha \mid\!\sim \beta$ is to be read as "β follows plausibly from α". For example, if we represent the information that Tweety is a bird by the atom b, and that Tweety can fly by the atom f, the statement $b \mid\!\sim f$ is to be read as "from the fact that Tweety is a bird it follows plausibly that Tweety can fly". Kraus et al. define $\mid\!\sim$ as a *rational consequence relation* iff it satisfies the properties Ref, LLE, RW, And, Or, CM, and RM given below.

(Ref) $\alpha \mid\!\sim \alpha$ **(Reflexivity)**

(LLE) If $Cn(\alpha) = Cn(\beta)$ and $\alpha \mid\!\sim \gamma$ then $\beta \mid\!\sim \gamma$ **(Left Logical Equivalence)**

(RW) If $\gamma \in Cn(\beta)$ and $\alpha \mid\!\sim \beta$ then $\alpha \mid\!\sim \gamma$ **(Right Weakening)**

(And) If $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$ then $\alpha \mid\!\sim \beta \wedge \gamma$

(Or) If $\alpha \mid\!\sim \gamma$ and $\beta \mid\!\sim \gamma$ then $\alpha \vee \beta \mid\!\sim \gamma$

(CM) If $\alpha \mid\!\sim \beta$ and $\alpha \mid\!\sim \gamma$ then $\alpha \wedge \beta \mid\!\sim \gamma$ **(Cautious Monotonicity)**

(RM) If $\alpha \mid\!\sim \gamma$ then either $\alpha \wedge \beta \mid\!\sim \gamma$ or $\alpha \mid\!\sim \neg\beta$ **(Rational Monotonicity)**

We do not discuss these properties in detail here. Instead, the interested reader is referred to the paper of Kraus et al. [47]. To make the connection with AGM belief revision, we need to go one step further. Gärdenfors and Makinson [27] define $\mid\!\sim$ as an *expectation based consequence relation* iff it is a rational consequence relation which also satisfies the property CP given below.

(CP) If $\alpha \mid\!\sim \bot$ then α is Cn-inconsistent **(Consistency Preservation)**

(where $\bot$ is any truth-functional contradiction, e.g., $p \wedge \neg p$). The underlying intuition provided by Gärdenfors and Makinson is that the reasoning of an agent is guided by its *expectations*. Every expectation based consequence relation $\mid\!\sim$ is based on a set of expectations E, playing a role that is analogous to that of a belief set K in theory change. Intuitively, E is the "current" set of expectations of the agent, and the plausible consequences of a sentence α are those sentences β for which $\alpha \mid\!\sim \beta$ holds. The set of expectations E is not explicitly mentioned in the definition of an expectation based consequence relation $\mid\!\sim$, but a suitable E can be recovered from $\mid\!\sim$ as follows: $E = \{\alpha \mid \top \mid\!\sim \alpha\}$. That is, E is taken as the set of plausible consequences of a tautology.

This places us in a position to define a method for translating between belief revision and expectation based consequence relations. Given a consequence relation $\mid\!\sim$, we take the set of expectations E associated with $\mid\!\sim$ as the theory K to be revised, and we define $K * \alpha$ as $\{\beta \mid \alpha \mid\!\sim \beta\}$. Conversely, given a theory K and a revision operator $*$, we define a nonmonotonic consequence relation $\mid\!\sim$ as follows: $\alpha \mid\!\sim \beta$ iff $\beta \in K * \alpha$. The main result, linking belief revision to nonmonotonic reasoning is the following theorem by Gärdenfors and Makinson [27] proving that these definitions allow us to show that AGM revision and expectation based nonmonotonic consequence coincide:

Theorem 21. *Let $\mid\!\sim$ be an expectation based consequence relation and let $E = \{\beta \mid \top \mid\!\sim \beta\}$. Then $E = Cn(E)$ (i.e. E is a theory). Furthermore, the*

revision operator $*$ for E, defined in terms of $\vdash\sim$ as follows: $E * \alpha = \{\beta \mid \alpha \vdash\sim \beta\}$, is an AGM revision operator. Conversely, consider a theory K, and let $*$ be an AGM revision operator for K. Then the consequence relation $\vdash\sim$ defined as follows: $\alpha \vdash\sim \beta$ iff $\beta \in K * \alpha$, is an expectation based consequence relation.

8 Information change in epistemic logic

In the work on belief change we have described thus far we have avoided discussion on its most fundamental notion: that of belief itself. Since the seminal book of Hintikka [43] this question is traditionally explored in the context of modal logic within *doxastic and epistemic logics*.[8] In these logics belief is studied *within* the object language as a modal operator. Although the initial work in this area was concerned purely with *static* notions of belief and knowledge, in more recent times there has been much interest in bringing dynamics into the picture, and studying how beliefs change in response to various learning events.

In a language for epistemic logic the sentences are built up from propositional variables using the usual propositional connectives, but now also an explicit modality for belief, so that whenever α is a sentence then so is $B\alpha$. The latter sentence has the intuitive reading that the agent believes α. Semantics is provided by a (single-agent) *epistemic model* $\mathcal{M} = (S, R, v)$, where S is a set of *states*, R is a binary *accessibility relation* over S, and v is a function which assigns a truth-value to every propositional variable at every state. For each $s \in S$, the set $R(s)$ of all states t which are accessible from s, i.e., such that $R(s, t)$ holds, intuitively represents those states which are consistent with the agent's information at state s. Evaluation of sentences is made with respect to a model-state pair $(\mathcal{M}, s)$, where $s \in S$, with the crucial clause for $B\alpha$ being as follows:

$$(\mathcal{M}, s) \models B\alpha \text{ iff } (\mathcal{M}, t) \models \alpha \text{ for all } t \in S \text{ s.t. } R(s, t).$$

By putting various restrictions on the accessibility relation R we can obtain different properties for the B-operator. For example by assuming that R is serial, transitive and Euclidean[9] we obtain the most common modal logic of belief, known as **KD45**, which is the axiomatic system with inference

[8] Strictly speaking, since we deal here with belief rather than knowledge, the adjective *doxastic* (rather than *epistemic*) is the appropriate one. However, since the use of the latter is widespread we shall use it in the rest of this section.

[9] R is serial iff for every s there exists some t such that $R(s, t)$. It is Euclidean iff $R(t, u)$ whenever both $R(s, t)$ and $R(s, u)$.

rules Modus Ponens (if α and $\alpha \rightarrow \beta$ are theorems then so is β) and Necessitation (if α is a theorem then so is $B\alpha$) and which has as axioms all instances of propositional tautologies together with the following:

K $B(\alpha \rightarrow \beta) \rightarrow (B\alpha \rightarrow B\beta)$ **(Distribution)**

D $B\alpha \rightarrow \neg B\neg\alpha$ **(Consistency)**

4 $B\alpha \rightarrow BB\alpha$ **(Positive Introspection)**

5 $\neg B\alpha \rightarrow B\neg B\alpha$ **(Negative Introspection)**

As can be seen in axioms **4** and **5** above, epistemic logics afford us the possibility to express *higher-order* beliefs, or *beliefs about beliefs*, directly in the object language. Moreover, in *multi-agent* extensions of epistemic logic, in which we have a number of different agents $\{1, \ldots, n\}$, to each of which we assign its own accessibility relation R_i, we can also have sentences of the form $B_iB_j\alpha$, expressing what one agent i believes about the beliefs of another agent j.

So far our description of epistemic logic only deals with static belief. In order to build dynamics into this framework one may introduce *dynamic modalities*. This is what is done in *Dynamic Epistemic Logic* (DEL) [73]. Full DEL comes with a rich typology of different belief-changing events, up to (different varieties of) private announcements and complex forms of epistemic action in which different agents can have different perspectives on the learning event. We will just mention here the prototypical kind of event, namely *public announcement* [29, 58] in which all agents in the scenario being modelled simultaneously learn that φ is true. In public announcement logic (PAL) we introduce new dynamic modalities $[!\varphi]$, where φ can be any sentence (including one containing belief modalities). The crucial semantic clause is as follows:

$$(\mathcal{M}, s) \vDash [!\varphi]\alpha \text{ iff if } (\mathcal{M}, s) \vDash \varphi \text{ then } (\mathcal{M}|\varphi, t) \vDash \alpha,$$

where $\mathcal{M} \mid \varphi$ is the epistemic model obtained from $\mathcal{M}$ by eliminating all states s' for which $(\mathcal{M}, s') \nvDash \varphi$, with the accessibility relations and valuation functions restricted accordingly. Note that a public announcement of φ represents *hard information* that φ is true. As such it is more in the spirit of belief *expansion* than revision (but see [4, 72]). Some interesting things happen if we try to reformulate the AGM revision postulates in terms of PAL. For one thing, the natural translation of the **Success** postulate does not hold, i.e., the sentence $[!\varphi]B\varphi$ need not be valid for all choices of φ. The best known counterexample is if we take φ to be a Moore-type sentence such as $p \wedge \neg Bp$, where p is a propositional variable ("p is true, but I don't believe it"). While it may well be the case that $(\mathcal{M}, s) \vDash p \wedge \neg Bp$, the

sentence $B(p \wedge \neg Bp)$ is inconsistent (i.e., true in no state). Thus although a Moore sentence may be truthfully announced, it can *never* be believed *after* the announcement.

The above considerations give rise to a distinction in the epistemic logic literature between *static* and *dynamic* belief revision. In static belief revision the result of revision expresses what an agent comes to believe about what *was* the case *before* the actual learning event took place. Dynamic revision deals with what the agent now comes to believe *after* the learning. The distinction only comes into effect when revising by higher-order beliefs such as with the Moore sentence above. When dealing with *factual* beliefs as with AGM the two notions coincide.

One way to model static belief revision which has been explored is via *doxastic conditionals*. This involves allowing sentences of the form $B^{\alpha}\varphi$ into the language, expressing the hypothetical belief that if the agent learned α then he would believe that φ was true before the learning. Roughly-speaking, the semantical structures for this language are obtained by replacing an agent's set of accessible states from each state by binary plausibility relations (total preorders) over states [7, 9]. Then the doxastic conditional $B^{\alpha}\varphi$ evaluates to true iff $(\mathcal{M}, s') \models \varphi$ for all the minimal states s' in the ordering such that $(\mathcal{M}, s') \models \alpha$. Within this framework one can define dynamic modalities for learning events of *soft* information, unlike the hard information of public announcement. These events have the effect of modifying the agents' plausibility relations rather than eliminating states from the picture completely. Van Benthem [71] studies modalities for two such events: *lexicographic upgrade* $[\Uparrow \varphi]$ and *conservative upgrade* $[\uparrow \varphi]$, which essentially correspond respectively to the lexicographic tpo revision method and to Natural revision described in Section 6.1 of the present paper.

For more details on belief change in epistemic logics we refer the interested reader to the survey articles [30] and [8]. Section 7 of the latter also includes a detailed comparison with AGM belief revision.

9 Current developments: belief change for other logics

From the work discussed so far it is clear that belief change has come a long way in the past 30 years. However, a look back at the work done over this period reveals an interesting tendency. Although the original aims were phrased in terms of a broad class of logic—all those with Tarskian consequence relation and satisfying **Compactness**—most of the work done in

the area is actually based on the assumption of an underlying *propositional logic*, whether finitely or infinitely generated. In this section we consider a departure from this trend, and discuss recent developments in belief change expressed in two logics other than full propositional logic: *propositional Horn logic* and *description logics*.

9.1 Propositional Horn contraction

One of the main reasons for considering belief change for Horn logic is that it has found extensive use in artificial intelligence and database theory, in areas where belief change is an issue to consider, such as logic programming, truth maintenance systems, and deductive databases. Delgrande [21] was the first to point this out and to investigate the contraction of theories for propositional Horn logic.

A *Horn clause* is a sentence of the form $p_1 \wedge p_2 \wedge \ldots \wedge p_n \rightarrow p_{n+1}$ where $n \geq 0$, and where the $p_i s$ are propositional atoms or one of $\bot$ or $\top$. A *Horn sentence* is a conjunction of Horn clauses. A *Horn set* is a set of Horn sentences. Given a propositional language L_{P}, the Horn language L_H generated from L_{P} is simply the Horn sentences occurring in L_{P}. The Horn logic obtained from L_H has the same semantics as the propositional logic obtained from L_{P}, but just restricted to Horn sentences. A *Horn theory* is a Horn set closed under logical consequence, but containing only Horn sentences. We denote Horn consequence by $Cn_H(.)$.

Delgrande's main contributions were threefold. Firstly, he showed that the move to Horn logic leads to two different types of contraction which coincide in the full propositional case. Given a Horn theory H, the *entailment-based* contraction, or *e*-contraction, of a sentence α should result in a new Horn belief $H -_e \alpha$ of which α is not a logical consequence: $H -_e \alpha \nvdash \alpha$. On the other hand, the *inconsistency-based* contraction, or *i*-contraction, of a sentence α should result in a new Horn belief $H -_i \alpha$ which is such that adding α to it does not result in an inconsistency: $H -_i \alpha \cup \{\alpha\} \nvdash \bot$. In full propositional logic, a way to express *i*-contraction in terms of *e*-contraction would be to require that $H -_e \neg\alpha \cup \{\alpha\} \nvdash \bot$. This cannot be expressed in Horn logic, though, because it is not possible to express the negation of the Horn sentence α (see also Section 3). Below we consider only *e*-contraction. Similar results have been obtained for *i*-contraction as well.

Delgrande's second contribution was to show that *e*-contraction for Horn theories should not satisfy the controversial **Recovery** postulate. As an example of the failure of **Recovery** for *e*-contraction, take $H = Cn_H(\{p \rightarrow r\})$ and let $\alpha = p \wedge q \rightarrow r$. Then any reasonable version of *e*-contraction will

yield $H -_e \alpha = Cn_H(\emptyset)$. So $Cn_H(H -_e \alpha \cup \{\alpha\}) = Cn_H(\{p \wedge q \rightarrow r\})$ and therefore $H \not\subseteq Cn_H(H \cup \{\alpha\})$.

Delgrande's third contribution was to base the construction of Horn contraction operators on partial meet contraction. The definitions of remainder sets, selection functions, and partial meet contraction, as well as maxichoice and full meet contraction all carry over directly to *e*-contraction and we will not repeat them here. We refer to these as *e*-remainder sets (denoted by $H\perp_e\alpha$), *e*-selection functions, partial meet *e*-contraction, maxichoice *e*-contraction and full meet *e*-contraction respectively. As in the full propositional case, it is easy to verify that all *e*-remainder sets are also Horn theories, and that all partial meet *e*-contractions (and therefore the maxichoice *e*-contractions, as well as full meet *e*-contraction) produce Horn theories.

In two subsequent papers, Booth et al. [13, 14] extended Delgrande's work in a number of interesting ways. They show that while Delgrande's partial meet constructions are all appropriate choices for *e*-contraction in Horn logic, they do not constitute *all* the appropriate forms of *e*-contraction. For example, let $H = Cn_H(\{p \rightarrow q, q \rightarrow r\})$. It can be verified that, for the *e*-contraction of $p \rightarrow r$, maxichoice yields either $H^1_{mc} = Cn_H(\{p \rightarrow q\})$ or $H^2_{mc} = Cn_H(\{q \rightarrow r, p \wedge r \rightarrow q\})$, that full meet yields $H_{fm} = Cn_H$ $(\{p \wedge r \rightarrow q\})$, and that these are the only three partial meet *e*-contractions. Now consider the Horn theory $H' = Cn_H(\{p \wedge q \rightarrow r, p \wedge r \rightarrow q\})$. It is clear that $H_{fm} \subseteq H' \subseteq H^2_{mc}$. But observe that H' is not a partial meet *e*-contraction. Booth et al. argue that H' ought to be regarded as an appropriate candidate for *e*-contraction and, more generally, that *every* Horn theory between full meet and some maxichoice *e*-contraction ought to be seen as an appropriate candidate for *e*-contraction.

Definition 22. *For Horn theories H and H', $H' \in H\downarrow_e \alpha$ iff there is some $H'' \in H\perp_e\alpha$ s.t. $(\bigcap H\perp_e\alpha) \subseteq H' \subseteq H''$. We refer to the elements of $H\downarrow_e \alpha$ as the* infra *e*-remainder sets *of H wrt α.*

Definition 23. *Let H be a Horn theory. An infra e-selection function is a function τ such that for every $\alpha \in L_H$, $\tau(H\downarrow_e \alpha) = H$ whenever $H\downarrow_e \alpha = \emptyset$, and $\tau(H\downarrow_e \alpha) \in H\downarrow_e \alpha$ otherwise. We use an infra e-selection function τ to define an infra e-contraction as $H -_\tau \alpha = \tau(H\downarrow_e \alpha)$.*

Booth et al. show that infra *e*-contraction is captured precisely by the AGM postulates for theory contraction, except that **Recovery** is replaced by the **Core retainment** postulate we encountered earlier in the context of defining kernel contraction in Section 3.4.

Theorem 24 ([14]). *Every infra e-contraction satisfies* **Closure**, **Inclusion**, **Success**, **Extensionality**, *and* **Core retainment**. *Conversely, every e-contraction which satisfies* **Closure**, **Inclusion**, **Success**, **Extensionality**, *and* **Core retainment** *is an infra e-contraction.*

It is possible to define a version of kernel contraction for Horn logic, simply by closing under Horn consequence the results obtained from kernel contraction for bases.

Definition 25. *Given a Horn theory H and an incision function σ for H, the* kernel *e*-contraction for H *is defined as* $H \approx^{e}_{\sigma} \alpha = Cn_H(H -_\sigma \alpha)$, *where* $-_\sigma$ *is the base kernel contraction for H obtained from σ.*

Booth et al. prove that kernel *e*-contraction corresponds *exactly* to infra *e*-contraction. From these results it seems that the contraction of Horn theories exhibits a kind of "hybrid" behaviour, somewhere between classical base contraction and classical theory contraction. As evidence for this, recall firstly that in the classical case, partial meet contraction and kernel contraction coincide for theories, but that kernel contraction is more general than partial meet contraction when dealing with the contraction of bases. Furthermore, Horn *e*-contraction for theories does not satisfy the **Recovery** postulate, unlike classical contraction for theories, but similar to classical base contraction. And finally, the set of postulates provided by Booth et al. to characterise infra *e*-contraction (and kernel *e*-contraction) bears a close resemblance to the postulates for characterising Horn contraction for bases in the classical case.

To summarise, these recent investigations into Horn contraction have highlighted the fact that a move away from propositional logic as the underlying logic for belief change can yield interesting and unexpected results. Interestingly enough, although the motivation for initiating research on Horn contraction was partially motivated by an interest in Horn logic in its own right, another reason for doing so is that propositional Horn logic forms the backbone of a group of *description logics*, the class of logics to which we turn to next.

9.2 Belief change for description logics

Description Logics (or DLs for short) are a well-known family of logics used for knowledge representation [6]. They have become the formalism of choice for representing formal ontologies [44]. DLs are decidable fragments of first-order logic, mainly characterised by constructors that allow complex concepts (unary predicates) and roles (binary predicates)

to be built from atomic ones. We provide a brief description of two well-known DLs referred to as $\mathcal{ALC}$ and $\mathcal{EL}$, and show how they relate to belief change.

In the description logic $\mathcal{ALC}$ [67], concept descriptions are built from concept names using the constructors disjunction ($C \sqcup D$), conjunction ($C \sqcap D$), negation ($\neg C$), existential restriction ($\exists R.C$) and value restriction ($\forall R.C$), where C, D stand for concepts and R for a role name. To define the semantics of concept descriptions, concepts are interpreted as subsets of a domain of interest, and roles as binary relations over this domain. An interpretation I consists of a non-empty set Δ^I (the domain of I) and a function $\cdot^I$ (the *interpretation function* of I) which maps every concept name A to a subset A^I of Δ^I, and every role name R to a subset R^I of $\Delta^I \times \Delta^I$. The interpretation function is extended to arbitrary concept descriptions as follows. Let C, D be concept descriptions and R a role name, and assume that C^I and D^I are already defined. Then $(\neg C)^I = \Delta^I \setminus C^I$, $(C \sqcup D)^I = C^I \cup D^I$, $(C \sqcap D)^I = C^I \cap D^I$, $(\exists R.C)^I = \{x \mid \exists y \; s.t. \; (x, y) \in R^I$ and $y \in C^I\}$, and $(\forall R.C)^I = \{x \mid \forall y, (x, y) \in R^I$ implies $y \in C^I\}$. The distinguished concept name $\top$ is always interpreted as $\top^I = \Delta^I$. Similarly, the dinstinguished concept name $\bot$ is always interpreted as $\bot^I = \emptyset$. A DL *Tbox* contains statements of the form $C \sqsubseteq D$ (*inclusions*) where C and D are (possibly complex) concept descriptions. Tboxes are used to represent the terminology part of ontologies in different application areas. The semantics of Tbox statements is as follows: an interpretation I *satisfies* $C \sqsubseteq D$ iff $C^I \subseteq D^I$. I is a *model* of a Tbox iff it satisfies every statement in it. A Tbox statement φ is a *logical consequence* of a Tbox T, written as $T \models \varphi$, iff every model of T is a model of φ.

A concept name A is *concept-satisfiable* wrt to a Tbox T iff there is a model, say I, of T in which $A^I \neq \emptyset$. This turns out to be an important property for ontology construction—if some concept names are *concept-unsatisfiable* wrt a Tbox T it is usually an indication of modelling errors made during the construction of T. For example, Schlobach et al. [66] show the following part of a Tbox for the DICE medical terminology:

brain $\sqsubseteq$ CentralNervousSystem

brain $\sqsubseteq$ BodyPart

CentralNervousSystem $\sqsubseteq$ NervousSystem

NervousSystem $\sqsubseteq \neg$BodyPart

According to this, a brain is a body part as well as a central nervous system, while the latter is a type of nervous system, which, in turn, is not a body part. Formally, the concept *brain* is concept-unsatisfiable wrt the

Tbox. Checking for concept-satisfiability is closely related to checking for logical consequence. Indeed, for many DLs, including $\mathcal{ALC}$, checking for concept-satisfiablity can be reduced to checking for logical consequence. DL reasoners such as RACER [32] and FaCT++ [70] are able to detect concept-unsatisfiability quite efficiently.

The link with belief change comes in with attempts to deal with concept-unsatisfiability in appropriate ways. *Ontology debugging* [46, 66] is concerned with determining the cause of concept-unsatisfiability in a Tbox T, while *ontology repair* [65, 52] aims to modify T in such a way that all concept names beome concept-satisfiable. It turns out that the techniques used for ontology debugging are closely related to the special case of kernel contraction for belief bases known as safe contraction, which was mentioned in Section 3. Recall that the α-kernels of a base B are the minimal subsets of B implying α. Similarly, techniques for ontology debugging identify the minimal subsets of a Tbox T with respect to which at least one concept name is concept-unsatisfiable.

In ontology debugging the Tbox T isn't modified automatically. Instead, the ontology engineer, when presented with the "kernels" of Tbox statements, is expected to use this information to modify T manually in order achieve concept-satisfiability. In contrast, the aim of ontology repair is to modify T automatically to ensure concept-satisfiability. This is achieved by removing exactly one element from each of the "kernels" of Tbox statements, an approach that can be seen as safe contraction applied to concept-satisfiability. Ontology repair, in this sense, has more in common with belief *base* contraction than with *theory* contraction, since it is the Tbox statements occurring explicitly in the Tbox that are used to obtain the TBox "kernels", and not statements in the theory obtained from the the Tbox.

A different application of belief contraction, this time one that is more closely related to *theory* contraction, occurs in ontologies represented in one of the $\mathcal{EL}$ family of DLs [5]. In $\mathcal{EL}$ itself, the basic member of this DL family, concept descriptions are built up from concept names using just conjunction ($C \sqcap D$) and existential restriction ($\exists R.C$). As in $\mathcal{ALC}$, Tbox statements have the form $C \sqsubseteq D$, where C and D are (possibly) complex concepts. The lack of expressivity in $\mathcal{EL}$ is made up for by the efficiency of reasoning algorithms for it. In particular, the task of *computing the subsumption hierarchy* for an $\mathcal{EL}$ Tbox T (determining whether $T \models A \sqsubseteq B$ for all concept names A and B) has polynomial complexity (in the size of the Tbox). Moreover, it turns out that a member of the $\mathcal{EL}$ family is sufficiently expressive to represent a number of biomedical ontologies, including the widely used medical ontology SNOMED [68].

As with $\mathcal{ALC}$, the application of belief change to $\mathcal{EL}$ is also related to the construction of ontologies. In this case, however, it does not address concept-unsatisfiability. Indeed, since $\mathcal{EL}$ does not have negation, concept-unsatisfiability can only occur if the bottom concept $\bot$ is used explicitly. Instead, it relates to a different method for testing the quality of a constructed ontology: asking a domain expert to inspect and verify the computed subsumption hierarchy. Correcting such errors involves the expert pointing out that certain subsumptions are missing from the subsumption hierarchy, while others currently occurring in the subsumption hierarchy ought not to be there. A concrete example of this involves the medical ontology SNOMED [68] which erroneously classified the concept `Amputation-of-Finger` as being subsumed by the concept `Amputation-of-Arm`. Finding a solution to problems such as these is can be seen as an instance of *theory contraction*, in this case by the statement `Amputation-of-Finger` $\sqsubseteq$ `Amputation-of-Arm`. The scenario also illustrates why we are concerned with contraction of theories and not bases. In general, ontologies are not constructed by writing down DL axioms, but rather using ontology editing tools such as SWOOP[10] or Protégé[11], from which the axioms are generated automatically. Because of this, it is the theory obtained from a Tbox that is important, not the axioms from which the theory is generated.

It is only recently that researchers have started to pay attention to theory contraction for $\mathcal{EL}$ [12]. Indeed, much of the work relevant to this topic does not address the $\mathcal{EL}$ family of DLs directly at all. In particular, the work on propositional Horn contraction is of importance in this context. Horn clauses correspond closely to subsumption statements in DLs, since both Horn logic and the $\mathcal{EL}$ family lack full negation and disjunction. In this respect, there is still much work to be done before a claim can be made that belief contraction for $\mathcal{EL}$ has been addressed properly.

Finally, in this section we have focused on recent work related to belief *contraction* for descriptions logics, but it must be pointed out that there has also been some recent work on belief *revision* and related questions [51, 59, 60, 74].

10 Conclusion

In conclusion, we hope that this brief overview of belief change has convinced the reader that research in this area has come a long way over the

[10] http://code.google.com/p/swoop
[11] http://protege.stanford.edu

past 30 years, with the fundamentals of the topic now firmly in place. The main challenge ahead is to build on the established fundamentals and extend the work that has been done to new application areas. As we have seen in Section 9, this is already taking place. And although much remains to be done in this regard with, for example, different underlying logics raising interesting and unexpected questions, it seems clear that the existing body of work provides an appropriate springboard for finding solutions to those new issues that are cropping up.

Acknowledgements

Thanks are due to Guillaume Aucher and an anonymous reviewer for some helpful comments.

References

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] C. Alchourrón and D. Makinson. Hierarchies of regulations and their logic. In Hilpinen, editor, *New Studies in Deontic Logic*, pages 125–148. D. Reidel Publishing Company, 1981.

[3] C. Alchourrón and D. Makinson. On the logic of theory change: Safe contraction. *Studia logica*, 44(4):406–422, 1985.

[4] G. Aucher. Private announcement and belief expansion: an internal perspective. *Journal of Logic and Computation*, to appear.

[5] F. Baader, S. Brandt, and .C Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proceedings of IJCAI*, pages 364–369, 2005.

[6] F. Baader and W. Nutt. Basic description logics. In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[7] A. Baltag and S. Smets. The logic of conditional doxastic actions: A theory of dynamic multi-agent belief revision. In *Proceedings of the Workshop on Rationality and Knowledge, ESSLLI'06*, pages 13–30, 2006.

[8] A. Baltag, H. van Ditmarsch, and L. Moss. Epistemic logic and information update. In *Handbook of Philosophy of Information*, volume 8 of *Handbook of Philosophy of Science*, pages 361–455. 2008.

[9] O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49(1):49–80, 2004.

[10] R. Booth, S. Chopra, A. Ghose, and T. Meyer. Belief liberation (and retraction). *Studia Logica*, 79(1):47–72, 2005.

[11] R. Booth and T. Meyer. Admissible and restrained revision. *Journal of Artificial Intelligence Research (JAIR)*, 26:127–151, 2006.

[12] R. Booth, T. Meyer, and I.J. Varzinczak. First steps in $\mathcal{EL}$ contraction. In *Proceedings of the IJCAI 2009 Workshop on Automated Reasoning about Context and Ontology Evolution (ARCOE 2009)*, 2009.

[13] R. Booth, T. Meyer, and I.J. Varzinczak. Next steps in propositional horn contraction. In *Proceedings of IJCAI*, pages 702–707, 2009.

[14] R. Booth, T. Meyer, I.J. Varzinczak, and R. Wassermann. A contraction core for horn belief change: Preliminary report. In *Proceedings of the 12th International Workshop on Nonmonotonic Reasoning (NMR 2010)*, 2010.

[15] C. Boutilier. Revision sequences and nested conditionals. In *Proceedings of IJCAI*, pages 519–519, 1993.

[16] C. Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

[17] J. Cantwell. Some logics of iterated belief change. *Studia Logica*, 63(1):49–84, 1999.

[18] S. Chopra, A. Ghose, and T. Meyer. Non-prioritized ranked belief change. *Journal of Philosophical Logic*, 32(4):417–443, 2003.

[19] S. Chopra, A. Ghose, T. Meyer, and K.S. Wong. Iterated belief revision and the recovery axiom. *Journal of Philosophical Logic*, 37(5), 2008.

[20] A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89:1–29, 1997.

[21] J.P. Delgrande. Horn clause belief change: Contraction functions. In *Proceedings of KR*, pages 411–421, 2008.

[22] M.A. Falappa, E.L. Fermé, and G. Kern-Isberner. On the logic of theory change: Relations between incision and selection functions. In *Proceedings of ECAI*, pages 402–406, 2006.

[23] E. Fermé and S. O. Hansson. Shielded contraction. In *Frontiers in Belief Revision*, pages 85–107. Kluwer, 2001.

[24] G. Flouris, D. Plexousakis, and G. Antoniou. On Generalizing the AGM postulates. In *STAIRS 2006: Proceedings of the third starting AI researchers' symposium*, pages 132–143, 2006.

[25] A. Fuhrmann and S.O. Hansson. A survey of multiple contractions. *Journal of Logic, Language and Information*, 3(1):39–75, 1994.

[26] P. Gärdenfors and D.. Makinson. Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of TARK*, pages 83–95. 1988.

[27] P. Gärdenfors and D.. Makinson. Nonmonotonic inference based on expectations. *Artificial Intelligence*, 65:197–245, 1994.

[28] Peter Gärdenfors. *Knowledge in Flux : Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge, Massachusetts, 1988.

[29] J. Gerbrandy and W. Groeneveld. Reasoning about information change. *Journal of Logic, Language and Information*, 6(2):147–169, 1997.

[30] P. Gochet and P. Gribomont. Epistemic logic. In *Handbook of the History of Logic*, volume 7, pages 99–195. 2006.

[31] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[32] V. Haarslev and R. Möller. Racer system description. In *Proceedings of IJCAR 2001*, volume LNAI 2100, 2001.

[33] S. O. Hansson. Changes on disjunctively closed bases. *Journal of Logic, Language and Information*, 2:255–284, 1993.

[34] S. O. Hansson. *A Textbook of Belief Dynamics*. Kluwer Academic Publishers, 1999.

[35] S. O. Hansson, E. Fermé, J. Cantwell, and M. Falappa. Credibility-limited revision. *Journal of Symbolic Logic*, 66(4):1581–1596, 2001.

[36] S.O. Hansson. A dyadic representation of belief. In *Belief Revision*, pages 89–121. Cambridge Univ Press, 1992.

[37] S.O. Hansson. Reversing the levi identity. *Journal of Philosophical Logic*, 22(6):637–669, 1993.

[38] S.O. Hansson. Kernel contraction. *Journal of Symbolic Logic*, 59(3):845–859, 1994.

[39] S.O. Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50(2):413–427, 1999.

[40] S.O. Hansson and R. Wassermann. Local change. *Studia Logica*, 70(1):49–76, 2002.

[41] W.L. Harper. Rational conceptual change. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1976, pages 462–494. Philosophy of Science Association, 1976.

[42] M. Hild and W. Spohn. The measurement of ranks and the laws of iterated contraction. *Artificial Intelligence*, 172(10):1195–1218, 2008.

[43] J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

[44] I. Horrocks, P. Patel-Schneider, and F. van Harmelen. From $\mathcal{SHIQ}$ and RDF to OWL: The Making of a Web Ontology Language. *Journal of Web Semantics*, 1(1):7–26, 2003.

[45] Y. Jin and M. Thielscher. Iterated belief revision, revised. *Artificial Intelligence*, 171(1):1–18, 2007.

[46] A. Kalyanpur, B. Parsia, E. Sirin, and J. Hendler. Debugging unsatisfiable classes in OWL ontologies. *Journal of Web Semantics - Special Issue of the Semantic Web Track of WWW2005*, 3(4):268–293, 2005. (To Appear).

[47] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[48] I. Levi. Subjunctives, dispositions and chances. *Synthese*, 34(4):423–455, 1977.

[49] I. Levi. *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford University Press, USA, 2004.

[50] T. Meyer, J. Heidema, W. Labuschagne, and L. Leenen. Systematic withdrawal. *Journal of Philosophical Logic*, 31(5):415–443, 2002.

[51] T. Meyer, K Lee, and R. Booth. Knowledge integration for description logics. In *Proceedings of AAAI05*, pages 645–650, 2005.

[52] T. Meyer, K. Lee, R. Booth, and J. Pan. Finding maximally satisfiable terminologies for the description logic alc. In *Proceedings of AAAI*, pages 269–274. AAAI Press, 2006.

[53] R. Moore. Possible-world semantics for autoepistemic logic. In *Proceedings of the AAAI Workshop on Non-Monotonic Reasoning, New Paltz, NY*, pages 344–354, 1984.

[54] A. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41:353–390, 1994.

[55] A. Nayak, P. Nelson, and H. Polansky. Belief change as change in epistemic entrenchment. *Synthese*, 109:143–174, 1996.

[56] A. Nayak, M. Pagnucco, and P. Peppas. Dynamic belief revision operators. *Artificial Intelligence*, 146:193–228, 2003.

[57] A.C. Nayak, R. Goebel, and M.A. Orgun. Iterated belief contraction from first principles. In *Proceedings of IJCAI*, pages 2568–2573. Morgan Kaufmann Publishers Inc., 2007.

[58] J. Plaza. Logics of public communications. In *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216, 1989.

[59] G. Qi, W. Liu, and D. Bell. Knowledge base revision in description logics. In *Proceedings of JELIA*, pages 386–398, 2006.

[60] G. Qi, W. Liu, and D. Bell. A revision-based approach for handling inconsistency in description logics. *Artificial Intelligence Review*, 2006.

[61] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[62] H. Rott. Two dogmas of belief revision. *The Journal of Philosophy*, 97(9):503–522, 2000.

[63] H. Rott. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. *Towards Mathematical Philosophy*, pages 269–296, 2006.

[64] H. Rott and M. Pagnucco. Severe withdrawal (and recovery). *Journal of Philosophical Logic*, 28:501–547, 1999.

[65] S. Schlobach. Diagnosing terminologies. In *Proceedings of AAAI05*, pages 670–675, 2005.

[66] S. Schlobach and R. Cornet. Non-standard reasoning services for the debugging of description logic terminologies. In *Proceedings of IJCAI*, pages 355–360. Morgan Kaufmann, 2003.

[67] M. Schmidt-Schauß and G. Smolka. Attributive concept descriptions with complements. *Artificial Intelligence*, 48:1–26, 1991.

[68] S. Schulz, B. Suntisrivaraporn, F. Baader, and M. Boeker. SNOMED reaching its adolescence: Ontologists' and logicians' health check. *International Journal of Medical Informatics*, 78(1):S86–S94, 2009.

[69] W. Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. *Causation in decision, belief change, and statistics*, 2:105–134, 1988.

[70] D. Tsarkov and I. Horrocks. Description logic reasoner: System description. In *Proceedings of IJCAR*, pages 292–297, 2006.

[71] J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[72] H. van Ditmarsch, W. van der Hoek, and B. Kooi. Public announcements and belief expansion. In *Proceedings of AiML-2004 (Advances in Modal Logic)*, pages 62–73, 2004.

[73] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2007.

[74] Z. Wang, K. Wang, and R. Topor. Revision of DL-Lite knowledge bases. In *Proceedings of the Description Logics Workshop*, 2009.

Some Remarks on Knowledge, Games and Society

ROHIT PARIKH*

The peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form, but solely as the dispersed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess. The economic problem of society is thus not merely a problem of how to allocate "given" resources – if "given" is taken to mean given to a single mind which deliberately solves the problem set by these "data." It is rather a problem of how to secure the best use of resources known to any of the members of society, for ends whose relative importance only these individuals know.

F. Hayek
Individualism and Economic Order

1 Introduction

We give a sketch of recent developments in Epistemic Logic and Game Theory and explain how they affect our understanding of the workings of society.

2 Knowledge

The Kripke structures for the logic of knowledge have one or more accessibility relations R_i which are *reflexive, symmetric, and transitive*, or in other words, *equivalence* relations.

*City University of New York, Brooklyn College and CUNY Graduate Center, Computer Science, Mathematics and Philosophy, research supported in part by a grant from PSC-CUNY.

Example: Suppose that there is a person who is either male or female, tall or short. Then there are four possible combinations, MS, MT, FS, FT. We write, $W = \{MS, MT, FS, FT\}$ where W is our world of all possibilities. Here FT, for instance, stands for the fact that the person in question is a tall female. Suppose I know the *gender* of the person but not the *height*. Then for me, MS, MT are equivalent, and so are FS, FT. So my space of possibilities splits into classes,

$\{\{MS, MT\}, \{FS, FT\}\}$.

The two persons MS, MT are equivalent as far as my knowledge is concerned, as are the two people FS, FT. So, referring to myself as 1, and the other person as 2, I have the equivalence relation R_1 where for instance, FT is equivalent to both itself and FS. FS is equivalent to both itself and FT. Suppose that someone else knows the height but not the gender. Then for that person the classes would be

$\{\{FT, MT\}, \{FS, MS\}\}$.

Suppose now that the person in question is actually a tall female. Then 1 (that is I) knows that the person is female and 2 knows that the person is tall.

Suppose an announcement is now made that the person is not a tall male. 1 already knew this and has learned nothing about the person in question. 1 is still uncertain between FS and FT. However, 2 now knows that the person is a tall female. The new classes are, for 1,

$\{\{MS\}, \{FS, FT\}\}$,

and for 2,

$\{\{FT\}, \{FS, MS\}\}$.

At this point, 2 knows that 1 does not know that it is a tall female.[1] But 1 does not know that 2 knows that it is a tall female! Why? Because 1 knows that anything that 2 knows has to be true. If 1 knew that 2 knew that it was a tall female, then 1 could also conclude that it is a tall female, and of course 1 does not know that. So (to put it formally) the formula $K_2(\neg K_1(FT))$ is true but $K_1(K_2(FT))$ is false.

2 also knows that 1 does not know.

[1] We assume that the various background facts are common knowledge so that 2 knows that 1 only knows about gender.

3 Some formal developments

The language $\mathcal{L}$ is obtained from atoms $At = \{p, q, ...\}$ by closing under the truth functions and operators K_i. Specifically,

1. An atom p is a formula
2. If A, B are formulas then so are $\neg A$, $A \vee B$
3. If A is a formula then so is $K_i(A)$ for any $i \leq n$

Here n is the number of knowers we are considering. In actual examples we may use letters A, B, or names Ann, Bob, rather than i, j. The letters p, q etc denote basic facts.
A *model* for epistemic logic consists of a world or set of states W, relations $R_i \subseteq W \times W$ for each $i \leq n$ and finally, a map V from $W \times At$ to $\{0, 1\}$ where 0 stands for *false* and 1 stands for *true*. For each i, R_i is the accessibility relation for agent i. The R_i are usually assumed to be equivalence relations. Here $R_i(s, t)$ means that agent i cannot distinguish between the s and t.

Semantics We define the notion $M, s \vDash A$, where M is a model, $s \in W$ is a state, and A is a formula.

1. If p is an atom then $M, s \vDash p$ iff $V(s, p) = 1$
2. $M, s \vDash \neg B$ iff $M, s \nvDash B$
3. $M, s \vDash B \vee C$ iff $M, s \vDash B$ or $M, s \vDash C$
4. $M, s \vDash K_i(A)$ iff $(\forall t)(sR_it \rightarrow M, t \vDash A)$

Note that an equivalence relation gives rise to equivalence classes or to a *partition* of the space W.
Let us use the modalities K_1 for me (1) (the $\Box$ corresponding to R_1), and K_2 for the other person (2).
Then since both R_1, R_2 are reflexive we get the axioms

$K_1(A) \rightarrow A$ and $K_2(A) \rightarrow A$.
By transitivity, we get

$K_1(A) \rightarrow K_1(K_1(A))$ and $K_2(A) \rightarrow K_2(K_2(A))$.
And finally, by symmetry we get

$\neg K_1(A) \rightarrow K_1(\neg K_1(A))$ as well as

$\neg K_2(A) \rightarrow K_2(\neg K_2(A))$.
Of course both operators K_1, K_2 are normal and we get, for instance,

$K_1(A \rightarrow B) \rightarrow (K_1(A) \rightarrow K_1(B))$
If I know A and I know $A \rightarrow B$, then I (can) know B.[2]

Updates
Suppose we are in some model M at some state s and an announcement of some formula A is made. (We assume that the announcement comes from the outside, and not from one of the agents). The announcement is assumed to be public. That is to to say that all agents hear it together. It is also assumed that A is true at M, s.
Then the announcement converts M into a smaller model $M|A$ where the set W is reduced to the set W' defined by $W' = \{t|M, t \vDash A\}$. The states where A was false are simply dropped since they could not be the states from which an announcement of A was made. The relations R_i and the truth value map V remain the same but are now restricted to W'. Note that since A was supposed to be true at s, $s \in W'$.
In the example we had at the beginning, W was {MS, MT, FS, FT} and W' is {MS, FS, FT}. The state $MT$is dropped because the announcement *not MT!* was false at MT.
If A is an atomic formula or even a truth functional combination of atomic formulas then A becomes common knowledge after the announcement.[3] In other words, all the agents know A, all agents know that all agents know A, etc.
It is more subtle if A may contain knowledge operators. Suppose for instance that there is a bug on Bob's shoulder. Let that fact be represented by p. Ann says to Bob, *You don't know this but there is a bug on your shoulder*. So she is announcing $A = (p \wedge \neg K_b(p))$.
After the announcement, p becomes common knowledge, but A does not, since half of A is no longer true!! Bob *does* know p now. In particular, $K_b(p)$ has become true and $\neg K_b(p)$ has become false.
Some of the foundational work in this area was done by Jan Plaza, a Ph.D. student at CUNY. He now teaches at SUNY, Plattsburgh. The following URL is a link to his slides on dynamic epistemic logic.
http://faculty.plattsburgh.edu/jan.plaza/research/logic/public-slides.pdf

[2]But do I *actually* know B? This is the problem of logical omniscience. See [20] for a full discussion of this issue.

[3]This means that everyone knows A, everyone knows that everyone knows A, etc...

4 Games

We shall usually talk about two player games. The players are typically called **Row** and **Column**, but more catchy names may arise in specific contexts. It is a convention that Row is female and Column is male. We may also refer to Row and Column as players 1 and 2 respectively.

In so called normal form games, each player has a finite set of **strategies**, call them S_1 and S_2, and each can choose a particular strategy from their own set. Once the players have chosen their strategies, there are payoffs which depend on *both* the strategies. p_r, p_c are the payoff functions. So suppose that player Row chooses strategy a and Column chooses strategy b, then the payoffs would be $p_r(a, b)$ and $p_c(a, b)$.

Suppose Row has chosen a and Column has chosen b, then (a, b) constitutes a **Nash equilibrium** if, *given that Column is playing* b, Row has nothing better than a, and given that Row is playing a, Column has nothing better than b. In other words $p_r(a, b) \geq p_r(a', b)$ for all a' and $p_c(a, b) \geq p_c(a, b')$ for all b'.

Given two strategies a, a' for Row, we say that a is **dominated** by a' if regardless of what Column plays, a' always gives a better outcome for Row. Thus $p_r(a, b) \leq p_r(a', b)$ for all b and $p_r(a, b) < p_r(a', b)$ for at least one b. Sinilarly for dominance of a Column strategy b by b'. It is normally accepted that a player would never play a dominated strategy, and the opponent may then make his plans based on this fact.

We now give examples of various games in the literature.

4.1 Battle of the Sexes

In this game, the wife (Row) wants to go to the Opera and the husband (Column) wants to watch football. But each would rather go together than watch their favourite thing by themselves. So here are the payoffs. Row's payoffs in each box are listed first.

	Opera	Footb
Opera	2, 1	0, 0
Footb	0, 0	1, 2

If they go to different events, they are not happy, so the payoffs are zero for both. If they go to the *same event*, then both have positive payoffs, but the

wife's is higher if they go to the Opera and the husband's is higher if they go to football. There are two Nash equilibria, the NW one which is (2,1), and the SE one which is (1,2).

The fact that (1,2) is a Nash equilibrium can be seen geometrically. Row can change the row, but if she does her payoff will move from 1 to 0, and she will be worse off. Similarly, Column can change the column, but if he does, his payoff will change from 2 to 0, and he will be worse off.

4.2 Chicken

In this rather dangerous game, two cars race towards each other. If one goes straight and the other swerves, then the one who swerves has shown fear, and is called *chicken*. He is embarrassed while the other feels triumphant. If neither swerves then there is an accident which they both regret – if they survive.

	Swerve	Straight
Swerve	$1,1$	$-1,7$
Straight	$7,-1$	$-10,-10$

There are two Nash equilibria, the NE one which is (-1,7), with Row being the 'chicken' and the SW one which is (7,-1) with Column in that role.

4.3 Matching Pennies

	Heads	Tails
Heads	$1,-1$	$-1,1$
Tails	$-1,1$	$1,-1$

In this game, *Row* is the matcher and *Column* is the mismatcher. Both parties exhibit a penny and if both pennies match (are both showing heads or both showing tails) then Row wins. If one is showing heads and the other tails (mismatch), then Column wins. There are *no* (pure) Nash equilibria in this game (there *is* a mixed strategy equilibrium, but we shall not consider those here[4]).

[4]Still, as anyone knows who has ever played this game, there is a way in which one can on average win half of a series of matching pennies games, by just tossing your own penny on each move. This is called "playing a mixed strategy". The particular mixed strategy where both players choose heads half of the time turns out to be a Nash equilibrium. Such Nash equilibria are called mixed Nash equilibria. The strategies we have discussed above are called "pure strategies", and the equilibria for those are called pure Nash equilibria.

4.4 **Prisoner's dilemma (PD)**

In this game, two men are arrested and invited to testify against each other. If neither testifies, then there is a small penalty (for each) since there is no real evidence. But if one *defects* (testifies) and the other does not, then the defecter goes free and the other gets a large sentence. If both defect they both get medium sentences. Jointly they are better off (The payoffs are 2 each) if neither defects, but for both of them, defecting is the dominant strategy and they end up with (1,1) which is worse.

	Coop	Def
Coop	2, 2	0, 3
Def	3, 0	1, 1

There is a unique, rather bad Nash equilibrium at SE with (1,1), while the (2,2) solution on NW, though better for *both*, is not a Nash equilibrium. This fact has often been taken to imply how, without the existence of an external authority, individuals will harm or destroy each other in order to derive benefit.

We now discuss the first one of our folk examples. We start with an actual example from the Economics literature and then relate it to a story from Indian history.

5 Tragedy of the Commons

From[5] **"The Tragedy of the Commons"** by Garrett Hardin, 1968 [15].

The tragedy of the commons develops in this way. Picture a pasture open to all. It is to be expected that each herdsman will try to keep as many cattle as possible on the commons. Such an arrangement may work reasonably satisfactorily for centuries because tribal wars, poaching, and disease keep the numbers of both man and beast well below the carrying capacity of the land. Finally, however, comes the day of reckoning, that is, the day when the long-desired goal of social stability becomes a reality. At this point, the inherent logic of the commons remorselessly generates tragedy.
As a rational being, each herdsman seeks to maximize his gain. Explicitly or implicitly, more or less consciously, he asks, "What is the utility to me of adding one more animal to my herd?" This utility has one negative and one positive component.

[5] Some material in this section appeared in a previous article of mine [19]

1. The positive component is a function of the increment of one animal. Since the herdsman receives all the proceeds from the sale of the additional animal, the positive utility is nearly +1.

2. The negative component is a function of the additional overgrazing created by one more animal. Since, however, the effects of overgrazing are shared by all the herdsmen, the negative utility for any particular decision-making herdsman is only a fraction of -1.
Adding together the component partial utilities, the rational herdsman concludes that the only sensible course for him to pursue is to add another animal to his herd. And another.... But this is the conclusion reached by each and every rational herdsman sharing a commons. Therein is the tragedy. Each man is locked into a system that compels him to increase his herd without limit – in a world that is limited. Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons. Freedom in a commons brings ruin to all.

But Hardin was anticipated in India by four hundred years!

The following is from the famous **Akbar Birbal** collection of stories. Akbar was the third Mughal emperor and the grandfather of Shah Jehan who built the Taj Mahal as a monument (and mausoleum) for his wife. Birbal was one of his ministers and well known (at least in stories) for his wit and intelligence. Both lived in the second half of the sixteenth century.

5.1 **Birbal story:**

One day Akbar Badshah said something to Birbal and asked for an answer. Birbal gave the very same reply that was in the king's own mind. Hearing this, the king said, "This is just what I was thinking also." Birbal said, "Lord and Guide, this is a case of *a hundred wise men, one opinion* " (in Hindi, *sau siyane ek mat*). The king said, "This proverb is indeed well-known". Then Birbal petitioned, "Refuge of the World, if you are so inclined, please test this matter". The king replied, "Very good."
The moment he heard this, Birbal sent for a hundred wise men from the city. And the men came into the king's presence that night.
Showing them an empty well, Birbal said, "His Majesty orders that at once every man will bring one bucket full of milk and pour it in this well."
The moment they heard the royal order, every one reflected that "where there were ninety-nine buckets of milk, how could one bucket of water be

detected?" Each one brought only water and poured it in. Birbal showed it to the king.
The king said to them all, "What were you thinking, to disobey my order? Tell the truth, or I'll treat you harshly!" Every one of them said with folded hands, "Refuge of the World, whether you kill us or spare us, the thought came into this slave's mind that where there were ninety-nine buckets of milk, how could one bucket of water be detected?"
Hearing this from the lips of all of them, the king said to Birbal, "What I'd heard with my ears, I've now seen before my eyes: "a hundred wise men, one opinion!"

Birbal lived from 1528 to 1586, and died in the battle of Malandari Pass, in Northwest India.[6]

Analysis:
What is common between the example which Hardin gives and the Akbar-Birbal story? In each case, the individual benefits at the cost of the group. In the Hardin case, the herdsman benefits by having one more animal.[7] In the Birbal case, the "wise man" benefits by saving one pot of milk. In each case the group is harmed. In the case of the herdsmen, the common is overgrazed and the grass dies. In the Akbar-Birbal case, there is a danger that if the cheating is discovered, all hundred men face the threat of prison or even execution. Akbar was a benign king,[8] but not entirely immune to anger.
Also, in each case, cheating is a dominant strategy. If most of the others are cheating, it does no extra harm if you cheat too. And if most of the others are not cheating, then again it does no extra harm if you are one of the rare cheaters. But if everyone practices their dominant strategy and cheats, then there can be disaster for the whole group.
A very nice discussion of *information* as a sort of commons is to be found in [9]. Elinor Ostrom was the first woman to win the Nobel prize in Economics, 2009.
However, the contention that the PD is a typical game in society, has been disputed by many scholars. Brian Skyrms claims that the *Stag hunt* is a better example of a game which arises typically in society.

[6]http://en.wikipedia.org/wiki/Akbar-the-Great , http://en.wikipedia.org/wiki/Birbal

[7]We should acknowledge some differences between the two scenarios. The Akbar-Birbal story is a comedy rather than a tragedy and its economic consequences are surely minor. We submit, however, that the logical structure is very similar and Birbal should surely receive some credit.

[8]Akbar, though a Muslim, worked hard to create amity between Hindus and Muslims, even marrying a Hindu wife, and having endless discussions on religion with Hindus, Christians and Jains.

5.2 The Stag hunt

	Stag	Hare
Stag	2, 2	0, 1
Hare	1, 0	1, 1

The game Stag hunt has two Nash equilibria, at (stag, stag) and at (hare, hare). But the (stag, stag) equilibrium, which involves co-operation is better for both parties.

In this game, each person can decide to hunt hare, by himself. The hare is a small animal, and a person can hunt a hare by himself, but the reward is also small. Or two people could co-operate to hunt a stag, which requires joint effort, but the reward is substantially better.

In [22], Brian Skyrms argues that the Stag hunt is a better model for human co-operation than the Prisoners Dilemma. In the PD, both parties are better off betraying the other. In the Stag hunt, co-operation is better, although it works only if the other party also co-operates.

Amadae and Lempert also argue in [1] that the prisoner's dilemma is not a good model for social interactions. They recommend instead the game hi-lo whose payoff matrix looks as follows.

	High	Low
High	2, 2	0, 0
Low	0, 0	1, 1

In this game, there are two Nash equilibria. The (high,high) equilibrium yields a payoff of 2 to each. The (low,low) equilibrium yields a payoff of 1 to each. Clearly the first, (high,high) equlibrium is better. But in order for one player, row, to choose "high" she needs confidence that so will the other player. Similarly, for column to choose high, *he* needs confidence that row is choosing high. Thus each needing the other to justify his/her action, we have an infinite regress. Amadae and Lempert argue that some sort of team reasoning or social reasoning is needed in order for the two players to co-ordinate on (high, high).[9]

This point is not new and belongs to a new way of looking at game theory in terms of what is called *team reasoning*, a notion which is due to Michael Bacharach and others [4, 13]. [6] provide experimental evidence in favor of team reasoning where individuals put the interest of a group before their own personal interest.

[9] See however, [10, 11] who use the notion of focal point to similar effect.

One evolutionary argument in favor of team reasoning or group reasoning is that groups whose members co-operate with each other and do not put individual interest first are likely to prosper compared to other groups where selfishness reigns. Thus genes which are conducive to co-operation will be passed on to future generation.
See Tomasello's article on co-operation among humans as compared to chimpanzees in the New York Times [14]. The doctoral dissertation of Tithi Bhatnagar [5] also makes the point that relationships are primary for human beings.
In fact note that while dogs *are* social creatures which move around in packs, cats are more solitary and tend to hunt alone. Nonetheless, cats have closer and more affectionate relationships with human beings than they have with other cats! It is true that humans do go to war against each other, but wars imply armies, and humans could not form armies without a natural tendency to co-operate.
It may be a rash prediction but it may well be that in the future, when Science and Technology have become common knowledge of all humanity, Eastern societies which emphasize co-operation and duty will do better than Western societies which emphasize individual rights, and (consequently) selfish behavior. But this is speculation, and only time will tell.

6 The Role of Knowledge in Society

We now return to the issue of knowledge which was introduced by Hajek at the beginning of this paper. Hajek, when he speaks of resources, is probably thinking of goods, but we can use the term more generally. For instance, resources could mean *medical expertise*.
In [17] the authors consider the case of a physician Uma whose neighbour Sam is sick, but she does not know this. [17] argue that she does not then have an obligation to treat him. However, she acquires such an obligation when she is informed.
Consider now, by comparison, the problem of *parking*, which is a also a knowledge problem. When people are looking for parking in a busy area, they tend to cruise around until they find a space. What they are trying to acquire is *knowledge*, knowledge of where an empty space is.
This fact has unfortunate consequences as Shoup [12] points out.

> *When my students and I studied cruising for parking in a 15-block business district in Los Angeles, we found the average cruising time was 3.3 minutes, and the average cruising distance half a mile (about 2.5 times around the block). This*

> *may not sound like much, but with 470 parking meters in the district, and a turnover rate for curb parking of 17 cars per space per day, 8,000 cars park at the curb each weekday. Even a small amount of cruising time for each car adds up to a lot of traffic.*
>
> *Over the course of a year, the search for curb parking in this 15-block district created about 950,000 excess vehicle miles of travel – equivalent to 38 trips around the earth, or four trips to the moon. And here's another inconvenient truth about underpriced curb parking: cruising those 950,000 miles wastes 47,000 gallons of gas and produces 730 tons of the greenhouse gas carbon dioxide. If all this happens in one small business district, imagine the cumulative effect of all cruising in the United States.*

An algorithmic solution to the problem of parking might well be possible using something like a GPS system. If information about empty parking spaces was available to a central computer which could also accept requests from cars for parking spaces, and allocate spaces to arriving cars, then a solution could in fact be implemented. The information transfer and the allocation system would in effect convert the physically distributed parking spaces into the algorithmic equivalent of a queue. There would be little wasteful consumption of gasoline, and the drivers would save a great deal of time and frustration.

As our final example we consider elections. An election is a way of finding out who is the most popular candidate running for office. As Arrow [3], has pointed out, defining "most popular candidate" is not devoid of problems. But let us assume that in some particular election there is such a person.
Then the purpose of the election is to find out who he (say) is, that is to say, to acquire knowledge. But here, more than knowledge is involved, for the public also needs to be convinced that the candidate is the *right* one. A good election procedure also makes sure that no one knows anyone else's vote, for otherwise bribery and bullying could become endemic.
Suppose for instance that when you vote you receive a receipt showing whom you voted for. This would give you assurance that your vote was counted because you could check your vote in a database. But at the same time, someone else could say to you, "Show me the receipt that you voted for *my* favorite candidate, or I will beat you up." Thus the asset, that you can check your vote, can also become a liability.
Moreover, campaigning is a process by which the candidates inform the public of their positions so that the public acquires knowledge. This issue

is addressed in [7].
Of course, is it actually knowledge which the public acquires, or only, perhaps, a false belief in case the election was stolen? It is also possible that an election was not in fact stolen but that a large proportion of the public (typically the losing party) believes that it was stolen.
These issues are starting to be addressed, but this paper is probably not the place for that.

Further reading: The source [8] contains valuable papers on *Social Software*, a project started by the author at a lecture given at the *FSTTCS* conference in Hyderabad in 1996, and starting with [18], followed up by various authors. The book [2] gives amusing insights into how society operates with real people with their many foibles. [17] gives a knowledge theoretic analysis of moral obligations, and offers an analysis of the Kitty Genovese case. See [21] for the importance of social norms, trust, etc.

References

[1] S.M. Amadae, and Daniel Lempert, Deriving an “ought" from an “Isn’t," pitfalls in modeling social interactions, research report, Ohio State University, May 2010.

[2] Dan Ariely, *Predictably Irrational*, Harper Collins 2008.

[3] Kenneth Arrow, *Social Choice and Individual Values*, Wiley 1951, 1963.

[4] Michael Bacharach, *Beyond Individual Choice: Teams and Frames in Game Theory*, edited by Natalied Gold and Robert Sugden, Princeton U. Press 2006.

[5] Tithi Bhatnagar, *Subjective Well Being in the Indian Context: Concept, Measure and Index*, doctoral dissertation, Indian Institute of Technology, Bombay, India (2010).

[6] C.M. Colman, B.D. Pulford and J. Rose, Team reasoning and collective rationality: piercing the veil of obviousness, *Acta Psychol*, 128, June 2008, 409-12.

[7] Walter Dean and Rohit Parikh, The logic of campaigning, in M. Banerjee and A. Seth (Eds) *ICLA 2011*, LNAI 6521, pp. 38-49, 2011.

[8] Jan van Eijck, and Rineke Verbrugge, *Discourses on Social Software*, Amsterdam University Press, (Paperback - Apr 1, 2010)

[9] Charlotte Hess and Elinor Ostrom, Ideas, artifacts and facilities: information as a common-pool resource, *Law and Contemporary Problems*, 2003, 111-145.

[10] David Lewis, *Convention, a Philosophical Study*, Harvard U. Press, 1969.

[11] Thomas Schelling, *A Strategy of Conflict*, Harvard U. Press, 1960.

[12] Shoup, D.: Gone Parkin', Op-Ed page, *The New York Times*, March 29, 2007.

[13] Robert Sugden, The Logic of team reasoning, *Philosophical Explorations*, 6 (2003), 165-181.

[14] Michael Tomasello, How are humans unique?, *The New York Times*, May 25, 2008.

[15] Garrett Hardin, The Tragedy of the Commons, *Science*, **162**, No. 3859 (December 13, 1968), pp. 1243-1248.

[16] Martin Osborne and Ariel Rubinstein, *A Course in Game Theory*, MIT Press (1994)

[17] Eric Pacuit, Rohit Parikh and Eva Cogan The logic of knowledge based obligation, presented at Society of Exact Philosophy meeting in Maryland, and at *DALT 2004.* In *Knowledge, Rationality and Action*, a subjournal of *Synthese*, (2006). **149** 311-341.

[18] Rohit Parikh, Social Software, *Synthese*, **132**, Sep 2002, 187-211.

[19] Rohit Parikh, Knowledge, games and tales from the East, in *Logic and its Applications*, Ramanujam and Sarukkai, editors, Springer 2009, 65-76.

[20] Rohit Parikh, Sentences, belief and logical omniscience, or what does deduction tell us?, *Review of Symbolic Logic*, **1** (2008) 459-476.

[21] Karl Sigmund, *The Calculus of Selfishness*, Princeton, 2010.

[22] Brian Skyrms, *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, 2004

List of Contributors

Guest Editors

Amitabha Gupta,
Former Professor,
Department of Humanities & Social Sciences,
Indian Institute of Technology Bombay,
agcg503@gmail.com

Johan van Benthem,
Henry Waldgrave Stuart Professor of Philosophy,
Stanford University, California, USA,
University Professor of Pure and Applied Logic,
University of Amesterdam, Amesterdam, The Netherlands and
Weilun Visiting Professor of Humanities, Tsinghua University, Beijing.
http://philosophy.stanford.edu/profile/Johan+Benthem;van;van+Benthem/
http://staff.science.uva.nl/~johan

Part 1
History of Logic

1. **Wilfrid Hodges**,
British Academy, London.
http://wilfridhodges.co.uk/

Stephen Read
Professor of History and Philosophy of Logic,
Department of Logic and Metaphysics,
University of St. Andrews
http://www.st-andrews.ac.uk/~slr/read.html

2. **Fabien Schang**,
Postdoctoral researcher at the Technical University of Dresden and
Associate member at the Henri Poincaré Archives.
http://poincare.univ-nancy2.fr/Presentation/?contentId=1557

3. **Prabal Sen**,
Professor of Philosophy,
University of Calcutta, Kolkata 700 027
pksen_cu@rediffmail.com

Amita Chatterjee
Vice Chancellor,
Presidency University, Kolkata
amita_ju@yahoo.com

4. **Fenrong Liu**,
Associate Professor of Logic, Department of Philosophy,
Tsinghua University, Beijing, China.
http://fenrong.net/

Wujing Yang
Department of Philosophy,
Renming University of China, Beijing, China.
wujinyang3032@sohu.com

Part 2
Mathematical Logic and Foundations

5. **Anand Pillay**,
   Professor of Mathematical Logic,
   Department of Pure Mathematics,
   University of Leeds.
   http://www.maths.leeds.ac.uk/~pillay/

6. **Jouko Väänänen**,
   Professor, ILLC, University of Amsterdam, and
   Department of Mathematics and Statistics,
   University of Helsinki, Finland and University of Amsterdam.
   http://www.math.helsinki.fi/logic/people/jouko.vaananen/

7. **Jeremy Avigad**,
   Professor, Department of Philosophy and
   Department of Mathematical Sciences,
   Carnegie Mellon University.
   http://www.andrew.cmu.edu/user/avigad/

8. **S. Barry Cooper**,
   Professor, Department of Pure Mathematics,
   University of Leeds.
   http://www.amsta.leeds.ac.uk/~pmt6sbc/

9. **Hiroakira Ono**,
   Distinguished Professor, Research Center for Integrated Science and
   Japan Advanced Institute of Science and Technology (JAIST)
   http://www.jaist.ac.jp/profiles/info_e.php?profile_id=00035

Part 3
Logics of Processes and Computation

10. **Frank Wolter**,
    Professor of Logic and Computation,
    Department of Computer Science,
    University of Liverpool.
    http://www.csc.liv.ac.uk/~frank/

Michael Wooldridge,
Professor of Computer Science,
Department of Computer Science,
University of Liverpool.
http://www.csc.liv.ac.uk/~mjw

11. **Samson Abramsky, FRS**,
Christopher Strachey Professor,
Oxford University Computing Laboratory,
Fellow, Wolfson College, Oxford University.
http://www.comlab.ox.ac.uk/people/samson.abramsky/

12. **Ramaswamy Ramanujam**,
Professor of Logic and Distributed Systems,
Institute of Mathematical Sciences, Chennai, and
Lorentz Fellow, NIAS.
http://www.knaw.nl/cfdata/agenda/
agenda_detail.cfm?agenda__id=1433

Part 4
Logics of Information and Agency

13. **Eric Pacuit**,
Resident Fellow and Assistant Professor,
Department of Philosophy,
University of Tilburg.
http://www.tilburguniversity.nl/faculties/
humanities/tilps/people/ResidentFellows/

14. **Richard Booth**,
Faculty of Informatics,
Mahasarakham University,
Mahasarakham, Thailand
http://italpha.msu.ac.th/richard/

Tommie Meyer,
Extraordinary Professor,
School of Computing,
University of South Africa.
http://krr.meraka.org.za/people/tmeyer

15. **Rohit Parikh**,
Distinguished Professor,
Department of Computer Science and Information Science,
Brooklyn College and
CUNY Programs in Computer Science, Philosophy and
Mathematics, Graduate Center,
City University New York.
http://www.sci.brooklyn.cuny

Part 5
Logic and Interfaces with Philosophy

16. **Isidora Stojanovic**,
Jean Nicod Institute and
Centre for National Scientific Research, Paris.
http://www.ub.edu/petaf/?q=content/isidora-stojanovic

17. **Jeffrey Helzner**,
Assistant Professor,
Department of Philosophy,
Columbia University.
http://www.columbia.edu/cu/philosophy/
fac-bios/helzner/faculty.html

Vincent Hendricks,
Professor of Formal Philosophy,
Department of Philosophy,
University of Copenhagen and
Department of Philosophy, Columbia University.
http://akira.ruc.dk/~vincent/

18. **Bas C. van Fraassen**,
Former Professor of Philosophy of Science,
Princeton University, Professor, Department of Philosophy,
San Francisco State University.
http://www.princeton.edu/~fraassen/index.html

19. **Sven Ove Hansson**,
Professor of Philosophy and Head of the Division of Philosophy,
Royal Institute of Technology, Stockholm.
http://www.infra.kth.se/~soh/

20. **Horacio Arlo Costa**,
Associate Professor,
Department of Philosophy,
Carnegie Mellon University.
http://www.hss.cmu.edu/philosophy/faculty-arlocosta.php

21. **Prof. Dr. Hannes Leitgeb**,
Chair and Head of the Munich Center for Mathematical Philosophy,
Ludwig-Maximilians-Universität, München.

22. **Edward Zalta**,
Senior Research Scholar, Center for the Study of Language and Information (CSLI),
Stanford University.
http://mally.stanford.edu/zalta.html

Part 6
Logic and Other Disciplines

23. **Sonja Smets**,
Faculty of Philosophy and
Faculty of Mathematics and Natural Sciences,
University of Groningen.
http://www.vub.ac.be/CLEA/sonja/

24. **Kenny Easwaran**,
Assistant Professor,
School of Philosophy,
University of Southern California.
http://college.usc.edu/phil/people/faculty_display.cfm?Person_ID=1022684

25. **Dov Gabbay**,
Augustus De Morgan Professor of Logic,
Group of Logic, Language and Computation,
King's College London.
http://www.dcs.kcl.ac.uk/staff/dg/

26. **Alistair M.C. Isaac**,
Department of Philosophy,
Stanford University.
http://philosophy.stanford.edu/profile/Alistair\%20Isaac/cv/

Jakub Szymanik,
Postdoctoral Researcher,
Department of Philosophy,
Stockholm University.
http://www.jakubszymanik.com/

27. **Olivier Roy**,
Postdoctoral Researcher, Faculty of Philosophy,
Rijksuniversiteit, Groningen.
http://www.philos.rug.nl/~olivier/

28. **Petr Hajek**,
Professor and Director of the Institute of Computer Science,
Academy of Sciences, Czech Republic.
http://www.uivt.cas.cz/~hajek/