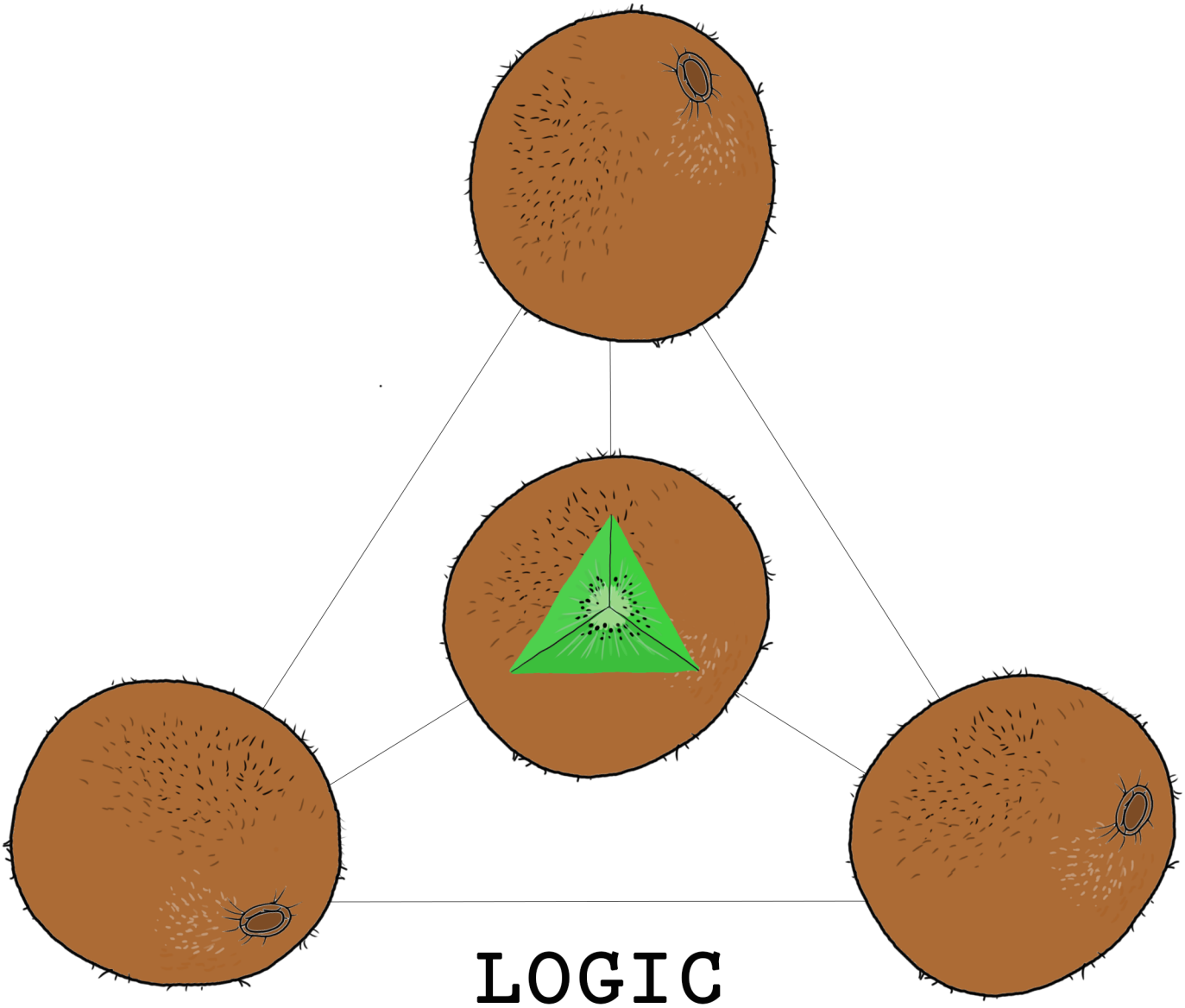
R I F A J

Rivista Italiana di Filosofia Analitica Junior

-peer reviewed journal-



LOGIC

Special Issue

Sponsored by the Italian Society of Analytic Philosophy since 2011

Edited by:
Pietro Casati
Marco Grossi

www.rifanalitica.it
ISSN: 2037-4445

Editorial team:
Stefano Canali
Mattia Sorgon

VOLUME 9
NUMBER 2

2018

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Rivista Italiana di Filosofia Analitica Junior Vol. 9, n. 2

Table of Contents

Copyright. © 2018 Published in Italy. Some rights reserved.

Articles (Special Issue)

Articles

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Editorial Letter

Pietro Casati, Marco Grossi

RIFAJ is almost nine years old and we are happy to say that we have reached the 17th volume. In all these years, logic has always been present, indirectly or directly, throughout all of the issues we have published so far: either hidden in the formalisms of some paper on *philosophy of language*, or out in the open, in an article on *artificial intelligence* or on *modality*. Logic, with its careful argumentation and its praise of rigor, is at the very centre of the analytic method that we endorse in philosophy and that we are trying to advertise through this journal.

However, we had not yet granted logic an issue of its own. This is where the idea for this volume comes in. We have gathered new and old articles, along with interviews and reports about topics on logic, broadly conceived. In this issue, we cover a vast array of arguments, from *truth* and *paradoxes* to *modality* and *computability*. The issue is divided into sections: there is a first section for interviews, reports and book reviews, then one for articles on logic and, lastly, one for articles on other related topics.

Given the positive feedback we have received in response to our call for papers, some new articles are yet to be peer-reviewed. We have therefore decided to extend the topic of this issue into the next one, which will once again generally cover logical issues. So, please do keep sending us more articles on logic, for the next issue!

Copyright. 2018 Pietro Casati, Marco Grossi. Published in Italy. Some rights reserverd.
Authors. Pietro Casati, pietroangelo.casati@gmail.com. Marco Grossi, 10marcogrossi@gmail.com.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Interview with Francesco Berto

Pietro Casati, Fabio Ceravolo

Introduction. Francesco 'Franz' Berto is Professor of Philosophy at the University of Amsterdam. Currently a passionate metaphysician and an explorer of deviant logics, he started studying Hegelian dialectic at the Ca' Foscari University of Venice, tutored and inspired by Emanuele Severino. At the time, he published *La dialettica della struttura originaria* (2003) and *Che cos'è la dialettica hegeliana?* (2005). As a postdoc in Padova and, successively, Paris, he obtained the 2007 Castiglioncello Prize for young philosophers with his *Teorie dell'assur do* [English version: *How to Sell a Contradiction*, 2007]. In 2012 he has been appointed senior lecturer at the University of Aberdeen, taking part at Crispin Wright's research project on the metaphysical basis of logic and working with Graham Priest on paraconsistent semantics, impossible worlds, and dialetheism. The results are contained in *Existence as a Real Property* (2013) [Italian version: *L'esistenza non è logica* (2010)]. Among students, though, he is chiefly well-known for his textbooks on quantified first-order logic [*Logica da zero a Gödel*, 2007] and Gödel theorems [*There's Something About Gödel: The Complete Guide to the Incompleteness Theorem*, 2009], published in Italy by Laterza [*Tutti pazzi per Gödel: La guida completa al teorema di incompletezza*, 2008]. In the introduction, he declares that he has written the books mostly in order to pay his bills. Personally, we are proud of our economic contribution. We have questioned him on his view on non-existent objects and on his "impressionistic impressions" on the reception of Hegel in analytic philosophy.

Copyright. CC BY NC SA 2018 . Published in Italy. Some rights reserverd. First published on *Rivista Italiana di Filosofia Analitica Junior* 5:2 (2014), pp. 1-8.
Authors. Pietro Casati, pietroangelo.casati@gmail.com. Fabio Ceravolo, fabioceravolo.012@gmail.com.

Having been concerned with inexistent objects, non-standard logic, (even) Hegelian dialectic (!), you proved yourself an uncommon, yet extremely interesting philosopher. It shouldn't be easy to work in isolation with respect to the more mainstream and fashionable debates… Thanks for calling me interesting!

Walking away from established debates can be a lot of fun: it allows you to explore original paths where less is taken for granted.

So-called analytic philosophy (*turbo-capitalistic philosophy*, as my esteemed colleague Diego Fuffaro would call it https://www.facebook.com/DiegoFuffa) often takes a lot for granted. Analytic philosophers work comfortably, we may say in Kuhnian fashion, within customary philosophical paradigms and without asking too many radical questions – until the paradigms enter a phase of recession.

Something of the sort is happening nowadays in ontology – precisely, in *meta*-ontology: the methodology of ontology. Here the Quinean paradigm of the Forties was taken for granted for a long time: ontological commitment is captured by the quantifier ("To be is to be the value of a variable" was one of Quine's famous rhetorical mottos). And the task of ontology is to write down the complete catalogue of the furniture of the world – of everything there is. Ontology gets the list right insofar as it includes nothing that isn't there, and leaves out nothing that is there. And that's it.

21st Century ontology is dominated by reactions against such paradigm: gro unding theorists like Kit Fine, Jonathan Schaffer and Fabrice Correia, neo-Meinon gians like Graham Priest, fictionalists like Stephen Yablo and Hartry Field, quantifier variantists like Eli Hirsch, and ontological pluralists like Jason Turner and Kris McDaniel, all claim that there's something seriously wrong with the Quinean framework. They propose to reform it in various ways, or even to reject it altogether. None of this would have happened if these folks hadn't felt the need to walk away from the established path.

In *Existence as a Real Property*, you write: "It is thanks to Doyle's creativity as a fantasy writer, that Holmes is available for reference and quantification at @" (p. 224). "Creativity" lays the foundations of the Comprehension Principle (UCP), according to which there is an object matching every possible combination of properties. But is the principle really so limitless? For instance, could Doyle write a story on the adventures of the round square in the actual world, or would he fall afoul to a modal error? Doyle could certainly do that, with no modal mistake. Here's why.

First, your "There is an object matching every possible combination of properties" is one formulation of the so-called Naïve or Unrestricted Comprehension Principle:

(UCP) Any condition A[x] is satisfied by something.

This principle quickly goes down in flames, for it delivers triviality. Let A[x] be: '$x = x$ & B'. By UCP, something – say, o – satisfies A[x]. Thus: $o = o$ & B. By conjunction elimination: B. But B was arbitrary. So one can use UCP to derive any conclusion one wills!

On the other hand, *nobody* has ever endorsed the UCP – not even Meinong. Instead, Modal Meinongianism (MM), the view endorsed by Graham Priest and myself, subscribe to, is a Qualified Comprehension Principle:

(QCP) Any condition A[x] is satisfied by some object at some world.

And "world" here means any situation or state of affairs, including ones that could *not* obtain, that is impossible situations.

Now, something's being round and square is an impossible situation for sure: the world could not be like that. But suppose Doyle writes about an *actual* round square, that is, something that is characterized by Doyle as round and square at the actual world @. How does MM handle this?

Easily enough: once ways the world could not be, that is, so-called non-normal or impossible worlds (see http://plato.stanford.edu/entries/impossible-worlds) are admitted in the semantics, one cannot expect modal operators to work uniformly across all worlds. And "actually" is one such operator. The truth and falsity conditions for "actually" given on pp. 175-6 of *Existence as a Real Property* (ERP) are as follows. Where w is a *possible world*:

'Actually A' is true at w iff A is true at @.

'Actually A' is false at w iff A is false at @.

If A is false at @, then 'Actually A' is not a necessary truth. But when w is a non-normal world, A can hold there even if A does not hold at @. So given A[x] = "x actually is a round square", we still have, as per the QCP, that something has, at some world, the property of being a round square at @. But this doesn't give us anything which is *actually* round and square, for the relevant world is a non-normal one, and the truth conditions above don't apply there.

The insight is obvious: you can imagine your impossible dreams to be actually realized, but that doesn't make them real. What your imagination produces is only something that is represented as being such-and-so-at-@, but nothing that *is*, at @, such-and-so.

The modal semantics developed by you and Graham Priest (2005) allow for rigid designation (denotation operators have no world-indexes). At the same time, QCP possibly assigns every property – 'of course' even inconsistent ones – to some object. There seems to be at least one property, though, such that, if both rigid designation and QCP hold, cannot be instantiated in any world. We

are talking about self-diversity: $\lambda x(x \neq x)$. For is it right to say that there can be no world in which Holmes fails to be identical to itself, on pain of abandoning rigid designation? And thus, doesn't rigid designation validate the necessity of self-identity, at least in the sense that there is no world in which it doesn't hold – even though there are (impossible) worlds in which it holds accompanied by its negation? Yes, MM is perfectly compatible with the necessity of identity: just have "=" get the same extension/anti-extension at all possible worlds. This gives you that if $a = b$, then '$a = b$' holds at all possible worlds.

Third, as for rigid designation and the "textbook Kripkeanism" story you have rehearsed: there being ways the world could not be, such that something is not self-identical, does not interact with rigid designation at all. Here's my latest logical fiction story:

> "Nonsy was non-self-identical and this made him very unhappy: he had done many different jobs, tried many experiences, met so many people, still he couldn't find his identity: 'Who am I?' – he kept wondering. Luckily one day Nonsy became friends with Selfy, an unselfish self-identical girl …"

…And so on. According to MM, in the above story "Nonsy" rigidly refers to Nonsy. Of course, Nonsy doesn't really exist – he's just a fictional character invented by me. At the actual world, though, Nonsy is perfectly self-identical: Nonsy just is Nonsy, and nobody else. And so he is at any other possible world: there is no way the world could be, such that something fails to be self-identical.

That "Nonsy" designates rigidly simply means that it refers to that guy even in counterfactual situations where he is supposed to have features he doesn't actually have. We can, for instance, wonder what Nonsy would do if he fell in love with Selfy. Nonsy isn't actually in love with Selfy (he doesn't really exist, recall? Nonexistents cannot fall in love). Still we refer to *Nonsy* when we use "Nonsy" to describe the counterfactual situation we are wondering about.
Now, given non-normal worlds, some counterfactual situations will also be counterpossible: situations in which Nonsy has properties he cannot have. $\lambda x(x \neq x)$ is just one such property. The usual textbook-Kripkean story holds: first we fix the reference of "Nonsy" at the actual world. Then we hold the reference fixed across nonactual circumstances, including circumstances that could not obtain, aka impossible worlds. Kripke said that we ought to keep our language constant across alternative worlds. He was, of course, right. There is no incompatibility between things being non-self-identical in some impossible circumstance or other, and rigid designation.

It is another issue, *how* we can actually fix the reference of "Nonsy". This is, in my view, the most serious problem of MM. It is called the Selection Problem. I tried to address it in the last Chapter of ERP and elsewhere, but it's a tricky issue

and all bets concerning it are still off.

Ok, we understand that you handle designation as a mere semantic clause allowing you to refer to Nonsy in every world-domain in which she (he? neither? both?) figures. But our textbook-like Kripkean stubbornness (and a good deal of Putnam *for dummies*, as you certainly guessed) suggests us that it is only conditionally upon admitting the necessity of self-identity that you can admit rigid designation, and that the two things are somehow connected – the former being a necessary condition for the latter. After all, when you baptize Nonsy (supposing you can) in the actual world, you baptize a necessarily self-identical thing. Given this relationship, it doesn't seem to be possible to refer to him (her?) in contexts where that very Nonsy is not the same thing as itself – would you really achieve reference? That Nonsy should be self-identical in every world, and that it is odd to utter truths on a non-self identical Nonsy are at least powerful non-semantic intuitions... I think there's a confusion between (a) referring at @ to something that, at world $w \neq @$, is such-and-so, and (b) referring at w to something that, at world w, is such-and-so.

It might be that, in a (closest) world w where Nonsy is not self-identical ("in contexts where that very Nonsy is not the same thing as itself", as you say), we cannot refer to him, say, because we keep referring to something else (!) or to nothing at all.

That goes under case (b). So it's not our problem here at @. We are in case (a): We can refer to Nonsy at @ (I just did it), as he is perfectly self-identical around here. We can then describe a counterfactual (indeed, counterpossible) scenario, w, in which Nonsy is non-self-identical (I did it above). That we *would* have problems in referring to a non-self-identical object in a counterpossible scenario where there are such things around doesn't affect our actually referring to something, which is then represented as non-self-identical in a logical fantasy.

Modal Meinongian metaphysics seems to lack an important feature: essential properties. Particularly, your semantics combines rigid designation and UCP, from which we can derive that every object can change every property, maybe even lose its self-identity, by yet 'remaining' the same object we are referring to in the actual world. Does any deflationism about essential properties hide behind this omission? And if so, how can it go together with rigid designation? Luckily, MM is perfectly neutral with respect to essentialism.

MM allows worlds where, for instance, Socrates is an iPhone 6. Essentialists may not like this. But they would be wrong. For MM semantics includes non-normal worlds, which are ways things *cannot* be. And the theory does not mandate taking worlds where Socrates is an iPhone 6 as possible.

Suppose you, qua essentialist, want Socrates to be essentially human. Then just impose to (the formal language counterpart of) "is human" the constraint that whatever makes it actually true also makes it true at any possible world (where the thing exists). Then Socrates, being human, will also be human at all possible worlds (ditto): situations in which Socrates is a smartphone will be ruled out from the realm of possibilities.

If, on the other hand, you don't like Socrates to be essentially human, because you are an anti-essentialist, just avoid imposing such a constraint. MM can make you happy either way.

Can we *conceive* Socrates as a glossy black new iPhone 6, even if the essentialist supposition that this is not a possible scenario is right? I claim that we can, but this issue, having to do with the connections between conceivability and possibility, is a tangled one, and we may avoid getting into this during our chat.

Right. Please, allow us to change our subject. In the previous discussion, we talked about contradiction, and you dealt with this problem by using paraconsistent logic. In the past you've also worked on dialectic. But what are the main differences exactly? And more generally, what is the relationship between dialectic and formal logic? The interest in dialectic still has an influence, perhaps subliminally, on your current research, or is it a closed chapter? "Dialectic" means lots of different things. If what you have in mind is Hegel's dialectic (or Hegel's dialectical method, or whatnot) – the relation between that stuff and formal logic is extremely complicated.

In the Sixties and Seventies, when the interest in Marx's and Hegel's thought was very lively, various people tried to "formalize" Hegel's dialectic (there's a great anthology on this, *La formalizzazione della dialettica*, edited by one of my philosophical heroes: Diego Marconi). Some of these formalizations used paraconsistent logic. The idea was that, since Hegel believed there to be true contradictions and took them to be essential to his "dialectical method", we had better adopt some paraconsistent logic to make sense of his views. For any non-paraconsistent logic, in the face of true contradictions, will allow you to infer that everything is true – and Hegel cannot have been *that* foolish.

However, that Hegel's dialectic requires there to be true contradictions is controversial: Bob Brandom, Diego Marconi, Emanuele Severino, Pirmin Stekeler-Weithofer, and others, deny this. If they are right, there's little need for paraconsistency to make sense of Hegel's dialectic.

As for my research on Hegel, I sometimes wish to go back to that. But I have so little time. Well, time will tell.

Introducing *Empiricism and the Philosophy of Mind* by Sellars, Richard Rorty argues that along with Quine, the later Wittgenstein, and then Brandom and McDowell there has been a transition of analytic philosophy from an initial empiricist phase to a Kantian phase, to finally land to Hegel. In this turn, you went in the opposite direction, departing from Hegel and arriving to analytic metaphysics. Given the historical perspective proposed by Rorty, in what would be the originality of analytic philosophy in repeating the same steps?
Ah, the sociology of philosophy is a difficult subject! Well here are my impressionistic impressions, shaped by my personal experience and unsupported by statistical data.

After having been dipped for some years in the analytic philosophy of the Anglo-Saxon countries, I'd say that the transition Rorty envisaged hasn't happened. The later Wittgenstein is less and less popular in the analytic camp, his main supporters being nowadays mostly interpreters of Wittgenstein who talk to each other, rather than people who engage in systematic philosophy.

As for Brandom and McDowell, while their work has had some impact on so-called continental philosophers and on some analytic folks, their positive influence on the analytic camp at large has been controversial. Some of the best analytic philosophers – people like Tim Williamson or my former boss in Scotland, Crispin Wright – have engaged with their work, but in a sharply critical way and in order to essentially dismiss it.

I understand why Rorty would have *wished* analytic philosophy to follow a path from Kant, through Hegel, and, possibly, into post-Hegelian and possibly relativistic thought. But nothing of the sort seems to me to be happening. This is what is actually happening, according to Tim Williamson's 2007 *The Philosophy of Philosophy* – he speaks of:

> [...] the liveliest, exactest, and most creative achievements of the final third of the [20th] Century: the revival of metaphysical theorizing, realist in spirit, often speculative, sometimes commonsensical, associated with Saul Kripke, David Lewis, Kit Fine, Peter van Inwagen, David Armstrong, and many others [...].
>
> On the traditional grand narrative schemes in the history of philosophy, this activity must be a throwback to pre-Kantian metaphysics: it ought not to be happening – but it is. (p. 19)

As for *my* going in the non-Rortian direction, that's purely accidental: I was initially raised a continental, and heavily trained in the history of philosophy. I discovered analytic philosophy later. And I have had a lot of fun doing analytic philosophy, continental philosophy, anything in the middle, and also any philosophy I've done that I wouldn't know how to label.

While the English-speaking world was once again interested in the Hegelian thought, in your *Che cos'è la dialettica hegeliana?* [*What is Hegelian Dialectic?*] you point out that the Italian community continues to believe it best to avoid even reading it. How do you explain this lack of interest? I guess you mean the community of Italian philosophers who consider themselves to be *analytic* philosophers. I think it's because the infamous analytic/continental divide is felt more strongly in the countries with a robust continental tradition, and Italy is one such country. I have experienced a somewhat similar situation when I was working at the Ecole Normale Supérieure in Paris. You may imagine the early analytic Italian folks from the Seventies and Eighties, trying to establish themselves as a research community, and the people surrounding them: neo-idealists, Hegelians, wannabe-Heideggerians, postmodernists, and so on. Not that analytic philosophers in the US or UK spend their whole day reading Hegel. But they never had a strong continental counterpart, in their own philosophy departments or in their national cultural milieu at large, against which a cultural reaction was called for.

When you were a student you have explored the work of Emanuele Severino, to whom you often recognize your intellctual debt. Do you consider his dialectic as an evolution of the Hegelian dialectic? And would you wish for a reception of his thought on the part of the analytic tradition similar to that of Hegel? I'd say that are various common points between the way I can make sense of Hegel's dialectic, and a view Severino labels as "dialectic" in some core chapters of (what I take to be) his most beautiful book, *La struttura originaria*. Not sure if it's an "evolution", for this may mean too many things. Certainly, Severino's dialectic helped me to understand Hegel's.

I don't think there will be any "analytic reception" of Severino's thought. This is due to various issues, one being that Severino's works are accessible almost only to an Italian readership and Italian is, regrettably, not a very important language from the viewpoint of top-level international research. Another issue is that Severino himself is usually recalcitrant when one tries to propose similarities and affinities between his thought and someone else's; this has to do partly with his own philosophy, partly with the man. It's a pity, but I don't think I can do a lot about this.

By way of consolation, consider that many young analytic philosophers were raised in Venice, firstly exposed to Severino, still admirers of his work, now pursuing brilliant careers of international profile: people like Elia Zardini, Roberto Loss, Matteo Plebani, and others.

Last question: can we learn something in advance on *Existence as a Real Property II* and the book you would like to write on Wittgenstein, if it does/will exist

in the actual world? Ha! Whether ERP II is to remain an unactualized *possibile* or not depends on my finding the time to work on that stuff again, and this is something I cannot predict at the moment. You know what Iris Murdoch once said? No philosophy book is ever finished: it is only abandoned.

I'm not sure I ever expressed the wish to write a whole book on Wittgenstein (maybe I once did, but I forgot). If I do, I hope I will manage to follow Wittgenstein's own recommendation, according to Malcom's memoirs: I hope it will be a philosophical book entirely composed of jokes.

Rivista Italiana di Filosofia Analitica Junior 9:2 (2018)

References

Berto, F. (2005). *Che cos'è la dialettica hegeliana?* Il Poligrafo.

— (2010). *L'esistenza non è logica: dal quadrato rotondo ai mondi impossibili*. Laterza. expanded eng. ed., *Existence as a Real Property*, 2013, Synthèse Library, Springer.

Priest, G. (2005). *Towards Non-Being: The Logic and Metaphysics of Intentionality*. Oxford University Press.

Sellars, W. (1956/1997). *Empiricism and the Philosophy of Mind: with an Introduction by Richard Rorty and a Study Guide by Robert Brandom*. Ed. by Robert Brandom. Harvard University Press.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Interview with Kevin Scharp

Marco Grossi

Introduction. Kevin Scharp is Reader in Philosophy and Director of Arché Philosophical Research Centre at the University of St Andrews.

Kevin has developed a conceptual engineering approach to the liar paradox, according to which the concept of truth is intrinsically defective and needs to be replaced with other concepts, when doing semantics. This method can be applied across philosophy, giving rise to a new methodology and a new way of thinking about philosophical problems. Kevin is also interested in the semantics for normative concepts. He and Bryan Weaver are going to publish a book titled *Semantics for Reasons* at Oxford University Press in 2019.

Copyright. 2018 Marco Grossi. Published in Italy. Some rights reserverd.
Author. Marco Grossi, 10marcogrossi@gmail.com.

Thanks for having us.
Thanks for inviting me.

I have a first question about what is philosophy for you. At the very beginning of your book *Replacing Truth* you say that you see philosophy, basically, as the study of inconsistent concepts. What do you mean by that?
Ok, so, I think that some of our concepts are defective in the sense that they can be used properly, but even when used properly they can lead someone to believe a contradiction or pursue contradictory plans, and be irrational. So, it is not that someone is misusing or misapplying the concept, for the person is using them properly, but is the concept itself that is defective. And so, I think that this way of thinking about concepts is very plausible in the case of standard philosophical concepts: the concepts that philosophers have been investigating for the last 2000 years. I think that is the case with respect to Truth, but also Goodness, Beauty, Value, Freedom, Knowledge etc. So, I think it makes sense to say that all of these standard philosophical concepts are defective, or inconsistent, and that this goes someway to explain why philosophical disputes are so intractable, and why they make such slow progress, and why philosophy is not like the sciences, at least the hard sciences.

Ok, and do you think that science is basically old philosophy? That when a concept becomes tractable it goes into the sciences?
It does not always happens, but yes, frequently it does. Though, sometimes it just gets kicked out completely. For example, alchemy was part of philosophy and all these various alchemical concepts were part of philosophy, and then most of them got kicked out. It is not that they became a science, but eventually a different science started, chemistry. And when the concepts of oxygen etc. got "cleaned up" from their alchemical background, and when those were sufficiently clean and in good shape, then they could be spun off as a science and they were no longer part of philosophy. I think you can see the sequence of sciences, you know, being immitted by philosophy over the last 400 years.

A question on Brandom. You did your PhD in Pittsburgh, Brandom was there. You said that he, more than everyone, has had an impact on your work. So can you tell us more about how he helped shape your current philosophical views?
I first encountered Brandom when I was a graduate student in Northwestern. So, I first started my PhD at Northwestern. I took a couple of seminars on Brandom's work. And I became really obsessed with his book *Making it explicit* and I felt like this was a really different and intesting way of thinking about standard philosophical issues and problems. He was drawing this very big picture: big

sweeping claims about the history of philosophy and how parts of philosophy interact with one another and about how whole vocabolaries relate to one another. And that really struck me as well: I like big pictures, and I think he was underepresented in philosophy and I though he was doing really well. So, then the people I wanted to work with at Northwestern left and so I got lucky and got accepted at Pittsburgh as a transfer student. And the paper that I wrote for my application was a criticism of Brandom. And he ended up liking it and I was in, and I was able to work with him, and he was the supervisor for my PhD. That was sort of a dream come true. I had never met him before, so being able to have him as a mentor was spectacular.

Is Brandom's pragmatism related to your conceptual engineering ideas?
I think so. He actually uses the term "conceptual engineering" in a paper from 2001, and The paper is a transcript of a talk he gave in 1999, and around the same time Blackburn uses the term in his book *Think*. But both of them when they use the term "conceptual engineering" they were just throwing remarks, they are both describing what do and what I care about, but they are not doing it in a kind of detailed way. I mean, the word occurs only once in each piece. But yes, the way Brandom thinks of concepts: the idea that even though surely we have a grip on concepts, in the way we "possess them", concepts also "have a grip on us", they "do things" to us, too. They almost have a sort of life of their own, to some extent. They are not these pure 100% accurate gems handed down from heaven or something like that. They are messy, and made up by humans over 1000 of years sometimes. It would be ridicoulus to expect them being anything other than a mess. And that is a big aspect of Brandom's work that had an influence on me.

Another aspect is that he enphasises that a theory of meaning should be "expressively complete", which means that it should be able to give the meaning of its own sentences, the sentences that make up the theory. So, for example, a verificationist theory of meaning, the one the positivists had, says that the meaning of some term is its method of verification. Yet this claim itself cannot be verified. This was one of the problems critics where pointing out: it looks like the positivists theory of meaning is self-undermining. And so Brandom made an explicit constrain on his theory of meaning so that it should not have this feature, so that it should not be self-undermining like that.

So, when I started thinking about truth, I saw the same kind of problem: lots of theories of truth do not apply to the language in which they are formulated. And so for me I think seeing Brandom having that kind of constrain on his views on meaning made a bit easier for me to insist on something like that in the realm of truth, while everyone else on truth was saying: "That is wrong, that's a mistake". There's a couple of other people, Vann McGee I think is a great example

of one other person out there insisting on this sort of constrain. When I started taking this constrain seriously, a lot of my ideas on truth kind of fell into place.

Thanks for bringing this out, since I wanted to talk about your book on Truth. So, the stuff you just said is related to revenge paradoxes. People usually come up with a solution to the Liar paradox, but somehow what they say about the paradox can get turned around and made into a new paradox that the theory can't handle. And so, you insist that this is a serious issue, it's not just a "puzzling thing". Since no conceptual analysis can solve this, so we need to replace the concept itself, at least in logical frameworks: in everyday life we can just get way with using the old inconsistent concept.
Exactly. There are a couple of more steps there. The concept is defective, yet it is useful. The defect is a problem for some of its uses. In particular, doing a semantics for a particularly rich language, like natural language. And so, in a situation like that, when you have an inconsistent concept that is preventing you from doing what you want to do with it, then it makes sense to replace it with one or more other concepts that do that job without having that defect. So, these are the boxes that you have to check in order to get to the replacement.

Gotcha. So, we can go a little deeper into this. Let me summarize briefly the issue and your solution. In standard semantics we want the so called T-schema: P if and only if P is true, for every sentence P. Yet, having this schema unrestrictedly gives rise to the liar paradox: there is a sentence that says of itself that it is not true. So, if it is true, it is not true, and if it is not true it is true. To avoid this, you say, we should replace Truth with two other concepts: Ascending and Descending Truth. With either of these concepts, you don't have the full T-schema. One satisfies only the left to right part of the T-schema, the other only the right to left. How can we be sure that the job that Truth does can be done with these concepts?
That is a difficult thing to show. I focus only on this one job of doing semantics, and on attributing truth-conditions to the sentences in question. If you want a theory that attributes truth-conditions then it's going to run into a liar paradox problem. The kind of theory that I propose does not offer truth-conditions to sentences, but instead it offers ascending truth-conditions and decending truth-conditions, using ascending and descending truth, respectively. And so then, the question is: "Great, why would I want these ascending and descending truth-conditions? How is that even doing semantics?" Well, in almost all cases, the ascending and descending truth-conditions are the same: they are just the truth-conditions. There are only differences when it comes to things like liar sentences, and only then they differ, and that is the key to avoiding the paradox. So in what sense is the theory that I offer really doing what we want to do, but bet-

ter? The sense is that it reduces to the truth-conditional theory in all the normal circumstances, it changes only in the circumstances that the truth-conditional theory simply can't handle, at all.

What about the idea that some people have, that paraconsistent logic is the answer. They say: "Yeah, that is a paradox, and that's it", you have to learn to live with the fact that something is both true and not true. How do you argue with that?

Yes, the price there is that you end up with a very weak logic as a result, and a very counterintuitive one, as well. One where usually modus ponens fails, if, for example, we are in Priest's Logic of paradox. I mean, you can go paraconsistent without that, right? You can have BX, which is a relevant logic, which is what Beall endorses. So you don't have to have that counterintuitive consequence of no modus ponens, but you still have others. Whenever you go paraconsistent, you are gonna have extremely counterintuitive consequences.

And this is the main problem you have with paraconsistent approaches.

I think there is a number of problems here, and I want to emphasize the fact that the dialethic paraconsistent view on paradoxes has a lot going for it. It's not just a crazy view. I lived through a time in my graduate school when people thought it was just crazy. And that was the standard objection, and people laughed and poke fun about it. Thankfully those days are over. At least for philosophy of language and logic people don't act like that anymore and that's good. It is sort of difficult to come up with decent objections when people just have this standard just-gut reaction: "This is gonna be wrong". My own way of thinking about what is wrong with the dialethic paraconsistent view is the following. There's lots of concepts and vocaboulary that we have in natural language that paraconsistent logician has trouble dealing with and the standard example is "just true". So, if I wanna call something "just true", what I mean is that it's not false, it is just true. How do I do that in a paraconsistent view? It is a standard problem. Priest has a solution to this, which I think it does not work. And J. C. Beall recently put out his "shrieking approach" to just true and that generated a decent bit of attention. I just put out a criticism of his shrieking approach. First, here is what the shrieking approach consists in. If you take the shrieking line, everytime I am calling something "just true" what I am calling "just true" is actually a bit different than what I think it is, and I am just calling that thing just true. Suppose you want to say that a theory T is just true. Now J. C. Beall comes along and says: "Here s what you really said. You really said that this other thing, not T but T-shrieked is true, where in the shrieked theory you added a consistency assumption.

You add a bunch of rules to the thing, and you change the theory.
Exactly. Now suppose we are having a conversation about something and I say: "Hey, guess what! T is both true and false". And you say: "No, it s not! it's just true!" Now, according to the shrieking approach you are not talking about T, you are talking about something else.

It's a change of subject.
Yes, exactly. we are not even talking about the same thing anymore. That seems bad. It gets worse than that: the approach is part of the bigger project by paraconsistent logicians to understand what happens in situations in which we are assuming to be consistent. And you cannot just add a consistency assumption and "get back" classical reasoning in paraconsistent systems.

This is actually a big asymmetry between paracomplete theories like Field's and paraconsistent theories like Priest's and Beall's. A paracomplete logician can get classical reasoning back by assuming excluded middle. It works pretty well compare to the paraconsistent theorist. For, if you just assume consistency, that is perfectly consistent with paraconsistency for the paraconsistent theorist. So, it does not solve anything at all. Here you have a big asymmetry, there.

One of the other problems I point out is that, in the shrieking approach, we must assume that there are shrieking operators all over language, hidden inside every sentence.

Yes, you would have to change the logical structure of sentences, and that would imply some big semantic blindness of the speakers.
Yes, exactly, it would imply a massive semantic blindness.

What do you think in general about paraconsistent dialetheist approaches?
Well, in general paraconsistent dialetheist and conceptual engineers agree on a lot of stuff. We both think that something like the analytic principles for truth are inconsistent. Yet the dialetheist says: they are analytic so they must be true but they imply their falsity so they are also false, so they are both true and false. What I say, on the other hand, is that they are meaning-constitutive but that does not mean they are true: they are just false. In this way I can go in a different direction, by not abandoning classical logic. I am rather after a consistent theory of inconsistent concepts.

The method you suggest has tons of ramifications. If you are right about truth, it might be that a lot of the philosophical discussion right now might be methodologically faulty, and we might rather try to revise our terminology in our philosophical discussion.
Yes. Firstly, the question is: how do you think about the pile of paradoxes that

show up around philosophy about different concepts? I am thinking about the paradox of Free will and Determinism, or the paradoxes of Knowledge, or of Goodness, or of Naturalness in metaphysics. Does that show that these concepts are defective? If so, you have to think about the use of these concepts. What do you want to do with them? And how might these defects get in the way, and if they do get in the way how do you replace these concepts with another team of concepts that do not have those defects to begin with. I think that in each of these cases the answers are different, because the uses of these concepts are different. So, there is no across-the-board recipe that works in every case. You really have to think through the details of each particular case and then make some suggestions on how to replace these concepts with others that are tailored to that particular case.

Now, let's say that you have replaced truth with two concepts, and knowledge and Goodness with some bunch of concepts. Now instead of three concepts we have, say, fifteen concepts. How three or four concepts interact with one another is quite a complicated issue but not crazy complicated, right? How fifteen concepts interact with one another: that's way more complicated. So, the number of decisions you have to make in thinking through how these replacement concepts interact with one another goes up as you replace one concept with two or three. So, that is one kind of a headache for my kind of a method. Yet, in some sense it is a good thing because it makes the project more rich and interesting. It is an additional step in the project that other philosophical methodologies do not have, at all. So, what I am thinking about now is how to extend this methodology across philosophy and thinking of philosophy as the study of what turned out to be defective concepts. If that's the case quite often, then how should our methodology be like? I think conceptual engineering in such cases is the right answer. But it needs a lot of detail: it turns out quite difficult to say much in general about how to do conceptual engineering, without looking at particular cases.

Is conceptual engineering what you mainly focus on in your work on deontic modals, as well? I know that you are going to publish a book titled *Semantics for Reasons* with Bryan Weaver in a few months.

So, my angle on Ought is somewhat different. It comes from thinking about reasons. There are a lot of people that want to say that Ought and Reasons are connected in important ways. So, if you have good reasons to do something, then that's what you ought to do, and what you ought to is what you have most reasons to. That's the basic idea. Now, the work on Ought in the literature is really interesting because it's been kind of "invaded" by philosopher of language and linguists and people doing natural language semantics. There is a tremendous amount of smart people thinking hard about how to understand the se-

mantics for 'Ought' statements. And what are the philosophical consequences of that? I think they are huge and they spread all across meta-ethics, moral psychology, normative ethics etc. But there has been not, at least since a couple of years ago, a semantics about reasons, at all. So, this is what the book I co-authored with Bryan Weaver is about: a semantics for reasons. Now, I don't think that one's approach for the semantics for reasons dictates one's approach for the semantics for Ought. I am kind of partial to the Kratzer style semantics for Ought[1] even though it does not always play super-nicely for the semantics for reasons-claims me and Weaver ultimately endorse. Two main things to think about when it comes to the semantics for reasons and that kind of lessons you might want to draw from other areas are these. Firstly, in the literature on reasons in meta-ethics there is a vast number of distinctions: internal-external, normative-motivating, and million others. Now, how are those distinctions related to one another? No one knows, right? And how are they related to the meaning of the word "Reasons"? Are these distinctions disambiguations? So, are they different meanings of the word "Reason"? Or are they rather different assumptions on what reasons are and they presuppose the same meaning for "Reasons"? How do they work?

So, one of the big things with our semantics theory is that we lay out six or seven major distinctions people appeal to in the literature and we explain each of those and how are they related to one another in term of their semantics. And so, you can see exactly how they are related to one another and you can see which are semantic distinctions and which are not semantic distinctions. We give a nice sort of method for distinguishing those in a simple way, with a simple linguistic test.

So, basically, the book would help the discussion between different sort of people, am I right? Like the psychologists and the linguists and the philosophers.

Yeah, that's the idea. The main focus of the whole project though is getting to a philosophical payoff. The idea is: now that we see what the semantics for reason is, we can think about whether the ontology of reasons are any good. We can think about whether the discussion between reasons and rationality are any good. And we can also judge the debate about moral reasons and other kind of reasons, as well. One of the main things that we do is use the semantics to develop a "reasons first" approach in general, where all normative phenomena can be explained in a certain sense in terms of reasons. But it's not the case that reasons can be explained in terms of other normative phenomena. So, reasons

[1] A Kratzer style semantics is a Neighbourhood semantics for modal logic: it can handle modal logics where K ($\Box(P \rightarrow Q) \rightarrow (\Box P \rightarrow \Box Q)$) or Necessitation (if P is a theorem then $\Box P$ is a theorem) fail.

are fist in the normative realm. This is one of the main philosophical stance you can defend using this semantics.

That all sounds really interesting. Thanks for the good chat and for your time.
Cool great! Thanks so much!

References

Beall, J. C. (2009). *Spandrels of Truth*. Oxford: Oxford University Press.

— (2013). "Shrieking against gluts: the solution to the 'just true' problem". In: *Analysis* 73.3, pp. 438–445.

Blackburn, Simon (1999). *Think: A Compelling Introduction to Philosophy*. Oxford: Oxford University Press.

Brandom, Robert B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Field, Hartry H. (2008). *Saving Truth From Paradox*. Oxford: Oxford University Press.

Kratzer, Angelika (1977). "What 'must' and 'can' must and can mean". In: *Linguistics and Philosophy* 1.3, pp. 337–355.

Priest, Graham (2005). *Doubt Truth to Be a Liar*. Oxford: Oxford University Press.

Scharp, Kevin (2007). "Replacing truth". In: *Inquiry: An Interdisciplinary Journal of Philosophy* 50.6, pp. 606–621.

— (2018). "Shrieking in the face of vengeance". In: *Analysis* 78.3, pp. 454–463.

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

Scuola estiva di Logica

Palazzo Feltrinelli – Gargnano sul Garda
26-31 agosto 2013

Matilde Aliffi

La Scuola Estiva di logica, organizzata dall'Associazione Italiana di Logica e sue Applicazioni (AILA) e dalla Società Italiana di Logica e Filosofia della Scienza (SILFS), ha avuto luogo dal 25 al 31 agosto a Gargnano sul Garda.

La Scuola, giunta alla quindicesima edizione, ha offerto due corsi istituzionali, il primo di carattere filosofico in Storia della Logica, tenuto dal Prof. Massimo Mugnai, docente di Storia della Logica alla Scuola Normale Superiore di Pisa, e il secondo in Logica Computazionale, tenuto dal Prof. Davide Sangiorgi, docente di Informatica dell'Università degli Studi di Bologna e membro dell'INRIA (Institut National de Recherche en Informatique et en Automatique). Accanto ai corsi istituzionali si sono svolte due lezioni magistrali, tenute dalla Dott.ssa Sonia L'Innocente, ricercatrice all'Università di Camerino e dal Prof. Vincenzo Marra, ricercatore all'Università degli Studi di Milano.

L'obiettivo della Scuola è quello di permettere a studenti e dottorandi in Filosofia, Informatica, Matematica, Fisica e Ingegneria di ampliare e integrare le proprie conoscenze in logica, favorendo inoltre un incontro tra una diversità di approcci e uno scambio di idee tra i partecipanti. La possibilità di incontrarsi viene percepita dagli studenti come particolarmente preziosa, data la mancanza in Italia di un percorso di studi specificamente indirizzato allo studio della logica. Durante le edizioni passate della Scuola infatti da questa esigenza è nata l'idea di creare ulteriori occasioni di condivisione delle proprie ricerche, e dal 2007 è stato istituito il Seminario di Logica Permanente (SELP)[1], che ha avuto modo di presentare le sue iniziative anche in questa edizione. Il tempo libero tra le lezioni e la cornice paesaggistica suggestiva hanno contribuito a favorire relazioni tra le persone e lo scambio di contatti e idee per condividere e realizzare ulteriori progetti futuri.

[1] http://selp.apnetwork.it/sito/.

Copyright. © 2018 Matilde Aliffi. Pubblicato in Italia. Alcuni diritti riservati. Prima pubblicazione su *Rivista Italiana di Filosofia Analitica Junior* 4:2 (2013), *Epistemologia*, pp. 52-61.
Autore. Matilde Aliffi, matilde.aliffi@gmail.com.

In questo report si tratterà sinteticamente una parte dei contributi delle lezioni mattutine, esprimendo alcuni dei concetti più importanti che sono emersi durante la Scuola.

Indice

1 *Lezioni in Storia della Logica*

Prof. Massimo Mugnai

Mugnai ha dedicato le sue lezioni alla nozione di "seguire da", ad alcuni elementi di logica antica e medievale, al pensiero di Leibniz e al processo di matematizzazione della logica. Durante il corso è emerso come fare storia di una disciplina "scientifica" richieda scelte di metodo, che riguardano il modo in cui si comprende il rapporto tra presente e passato, la possibilità di sostenere l' unità della disciplina, e un impegno sulla natura della logica. Servendosi anche dell'aiuto dei testi originali, Mugnai ha preferito leggere il passato evitando il ricorso alla logica dei "precorrimenti", secondo cui ciò che si afferma più ampiamente diventa punto di riferimento per analizzare il passato, mentre le possibili soluzioni alternative vengono lette come "rami secchi", privi di un reale interesse storico. Per Mugnai invece gli antichi non sono semplicemente degli anticipatori di ciò che più tardi nella storia si affermerà, ma vanno letti in tutta la loro ricchezza, all'interno di una adeguata contestualizzazione; il riferimento a soluzioni cronologicamente successive è stato usato infatti più come un confronto utile per differenziare e chiarire i concetti che come chiave interpretativa.

Mugnai ha inoltre insistito sulla peculiarità della logica, come disciplina "corta", nella quale la distanza tra lo stato attuale del suo sviluppo e il momento in cui è nata risulta meno marcata di quella di altre discipline scientifiche. Anche se la differenza tra la logica aristotelica e la logica matematizzata è piuttosto marcata, è difficile rifiutare di riconoscere che antichi e contemporanei condividessero problemi e concetti riguardanti la stessa materia. Mugnai ha quindi privilegiato una concezione continuista, ritenendo che sarebbe fuorviante cercare di stabilire una cesura tra la logica "classica", prefregeana, e la logica matematizzata. All'interno di questa cornice Mugnai ha svolto le sue lezioni, interessanti non solo per lo studente di filosofia, ma anche per lo studioso di scienze dure che ha potuto così arricchire di profondità storica i propri concetti.

Per quanto riguarda la ricostruzione storica della nozione di "seguire da" ci si è soffermati su tre diverse concezioni, riconducibili a Filone di Megara e Crisippo di Soli, logici e filosofi della scuola megarico-stoica e ad Abelardo[2], logico e filosofo medievale. Filone di Megara sosteneva che il condizionale fosse vero quando non si dà il caso che cominci col vero e finisca col falso. Le condizioni di verità del condizionale filoniano, quindi, si possono rappresentare attraverso la tavola di verità nella tabella 1, nella quale 0=falso e 1=vero.

α	β	$\alpha \rightarrow \beta$
1	1	1
1	0	0
0	1	1
0	0	1

Tabella 1: Tavola di verità del condizionale filoniano.

Secondo Filone, quindi, per determinare le condizioni di verità di un condizionale è sufficiente tener conto solamente dei valori di verità di antecedente e conseguente, evitando che l'antecedente sia vero senza che lo sia il conseguente. Non si richiede dunque alcun tipo di connessione tra antecedente e conseguente, infatti per Filone la conseguenza logica è valida anche per un enunciato del tipo «se la terra vola, la terra esiste» nel quale l'antecedente è falso e il conseguente vero. Questa concezione tuttavia potrebbe lasciare perplessi, infatti per essa qualsiasi conseguenza logica con un antecedente falso risulta vera, così come qualsiasi conseguenza nella quale antecedente e conseguente sono veri, indipendentemente dalla connessione tra essi.

Crisippo di Soli invece non effettua una valutazione del condizionale semplicemente componendo i valori di verità dei due membri, ma richiede che l'opposto del conseguente sia incompatibile con l'antecedente per concludere la verità della connessione. La conseguenza «se è giorno, c'è luce» quindi per Crisippo è vera perché l'opposto del conseguente è incompatibile con l'antecedente, infatti «non c'è luce» è incompatibile con «è giorno», mentre l'asserzione «se è giorno, Dione passeggia», vera nell'interpretazione filoniana, per Crisippo è falsa, infatti l'opposto del conseguente non è incompatibile con l'antecedente, dal momento che «Dione non passeggia» non è incompatibile con «è giorno».

Abelardo tuttavia critica il condizionale crisippeo perché secondo esso si è costretti anche ad accettare come veri tutti i condizionali che si basano su un antecedente impossibile. Mentre secondo Crisippo la conseguenza «se Socrate è una pietra, Socrate è un asino» è sempre vera, dal momento che è impossibile che «Socrate è una pietra» sia vero e «Socrate è un asino» falso, Abelardo nega la verità di questo condizionale. Egli infatti richiede una inseparabilità concettua-

[2]Per una lettura approfondita leggere (Mugnai 2013, cap. VI).

le, ossia che il senso del conseguente sia contenuto in quello dell'antecedente, una concezione che, letta con gli occhi del logico contemporaneo può definirsi quasi "rilevante". In questa prospettiva inoltre Abelardo fu il primo a distinguere l'argomento

(1) $\alpha \vdash \beta$

da

(2) $\alpha \rightarrow \beta$

Infatti per Abelardo, mentre l'argomento «se Socrate è un uomo allora Socrate non è una pietra» è corretto, non lo è necessariamente il condizionale corrispondente. Quindi «Socrate è un uomo *implica* Socrate è una pietra» è vero mentre «Socrate è un uomo, *dunque* Socrate è una pietra» è falso, poiché nel primo caso non è necessario il contenimento, mentre nel secondo sì.

Queste tre diverse concezioni del condizionale si ritrovano anche in autori contemporanei. Mentre Peirce ritiene che nell'ambito della logica formale il condizionale filoniano sia il più adatto, Hugh McColl nel 1880 presenta un calcolo logico su *Mind* analogo a quello di Crisippo (McColl 1880). Lewis invece adotta una interpretazione analoga a quella di Abelardo, proponendo nel 1912 su *Mind* un calcolo logico basato sulla implicazione stretta, secondo cui il condizionale è vero quando è impossibile che l'antecedente sia vero ed il conseguente falso (Lewis 1912).

Un altro problema trattato durante le lezioni è stato quello del rapporto tra la logica e la matematica. Esso è stato affrontato individuando in una prospettiva storica le origini del problema e le differenti soluzioni adottate, che hanno portato la logica alla sua matematizzazione. Anche se con il tramonto della Scolastica e l'affermarsi dell'Umanesimo si diffonde in Europa una generale diffidenza verso la logica, è dalla seconda metà del sedicesimo secolo che i rapporti tra logica e matematica iniziano ad essere discussi. Infatti nell'antichità logica e matematica venivano considerate due discipline distinte e nel medioevo furono in pochi ad occuparsi del problema dei rapporti tra le due discipline, nonostante la grande fioritura che ebbe lo studio della logica. Con la riscoperta dei testi di Euclide invece la matematica diventò esempio di rigore dimostrativo, e logica e geometria iniziarono ad avvicinarsi in un processo che vede un "movimento" della logica verso la matematica e un "movimento" della matematica verso la logica.

Il movimento della logica verso la matematica iniziò a realizzarsi nel sedicesimo secolo dalle idee di Conrad Dasypodius e Christian Herlinus. In quel periodo infatti ci si chiedeva se la logica tradizionale, di impianto aristotelico-scolastico, fosse adeguata a svolgere le dimostrazioni matematiche. Mentre secondo Alessandro Piccolomini la dimostrazione per eccellenza della tradizione

aristotelica non poteva essere applicata alla matematica, per Dasypodius e Herlinus era possibile rendere esplicita la struttura logica di ciascuna dimostrazione degli *Elementi* di Euclide attraverso la logica aristotelica con la aggiunta di altre regole e principi della tradizione stoica, come il *modus ponens* e la *legge di contrapposizione*.

Il movimento della matematica verso la logica invece ha origine con Thomas Hobbes; secondo il filosofo inglese, infatti, ragionare significa addizionare e sottrarre. Verso la fine del sedicesimo secolo, con François Viète iniziò a svilupparsi l'idea che fosse possibile utilizzare lettere dell'alfabeto per eseguire calcoli, al fine di ottenere una elevata generalità. Leibniz, con la scoperta del calcolo infinitesimale aveva mostrato che i calcoli non usavano solo numeri, ma lettere, estendendo l'ambito del calcolabile a qualsiasi tipo di simboli. Questa scoperta tuttavia generò un'ampia disputa per stabilire chi tra Newton e Leibniz ne meritasse la priorità, anche se in realtà, come oggi si può affermare, la scoperta del calcolo infinitesimale fu effettuata indipendentemente da entrambi. A conseguenza della disputa l'approccio newtoniano, fondato su una concezione "geometrico-dinamica" delle grandezze si diffuse soprattutto tra i matematici del Regno Unito, mentre nel continente, e in particolare in Francia e Germania, si preferì la notazione leibniziana, più facile da usare e svincolata dall'interpretazione di tipo fisico-cinematico, propria dell'approccio di Newton. In seguito a questa disputa i matematici inglesi rimasero in una situazione di relativo isolamento, finché, verso la metà dell'Ottocento, Augustus De Morgan e William Rowan Hamilton non rinnovarono con i loro studi l'interesse per la logica in Gran Bretagna. I due logici si impegnarono in una controversia sulla "quantificazione del predicato", che riguardava chi per primo avesse sostenuto, contrariamente al parere di Aristotele, che negli enunciati categorici tradizionali era legittimo esprimere la quantità del predicato, oltre a quella del soggetto. In questo clima George Boole ricevette lo stimolo a occuparsi di logica. Egli in *The Mathematical Analysis of Logic* distinse l'interpretazione dei simboli utilizzati nel calcolo dalle leggi che regolano la combinazione degli stessi simboli, e affermò che l'interpretazione quantitativa di essi non era l'unica possibile. Per Boole infatti i simboli possono essere usati anche per designare operazioni logiche o concetti generali, per esempio classi di oggetti qualsiasi promuovendo una evoluzione della matematica da "scienza della quantità" a "scienza della qualità". La logica viene quindi ricondotta nell'ambito di una trattazione algebrica ed il calcolo logico diventa un particolare settore della matematica applicata.

Mentre Boole completa il processo di avvicinamento della logica verso la matematica, arrivando ad assorbirla *nella* matematica, dal momento che la logica viene considerata un ramo della matematica applicata, Frege portò a termine il movimento della matematica verso la logica, fino a sostenere una preminenza della logica rispetto alla matematica. Secondo il filosofo tedesco infatti la ma-

tematica era concepita come una struttura originata dallo sviluppo di nozioni e principi logici fondamentali, attraverso definizioni e teoremi. Nella misura in cui le cui parti superiori possono essere ricondotte al fondamento, una volta che esso sia risolubile in assiomi e definizioni logiche, secondo Frege si è mostrato che l'intera matematica, ad eccezione della geometria, non sia altro che logica applicata. Per realizzare questo progetto Frege costruì un linguaggio caratteristico artificiale, l'ideografia, funzionale al progetto di logicizzazione della matematica.

È interessante notare che i due progetti portati avanti da Boole e da Frege erano già presenti nel pensiero logico leibniziano. Leggendo i suoi scritti infatti ci si rende conto che Leibniz avesse già sognato di "matematizzare" la logica, e questo secondo due prospettive parallele. Da un lato quella di tipo combinatorio, simile al progetto realizzato successivamente da Boole, per cui, dato un insieme di simboli si procede a "manipolarli" attraverso operazioni, avendo di mira fondamentalmente il risultato finale. Dall'altro nei testi di Leibniz si legge continuamente che in una dimostrazione tutto deve essere specificato nei minimi dettagli in modo rigoroso, senza salti e senza affidarsi ad espressioni delle quali non si controlla il significato. Questa insistenza rivolta a trovare una dimostrazione rigorosa conferisce una preminenza alla logica rispetto alla matematica, aspetto che può essere accostato al progetto fregeano.

La prospettiva di Frege, tuttavia, ha introdotto una distanza tra la logica matematizzata e la logica tradizionale, ed ha aperto la strada ad un approccio quasi normativo della logica, che, sempre più distante dal modo di pensare umano, cerca essere canone di come *dobbiamo pensare*. Alcuni logici contemporanei tuttavia hanno proposto una reazione a questa prospettiva. Johan van Benthem, in particolare, ha cercato di avvicinare logica e pensiero, lavorando all'interno di un programma di ricerca volto ad affermare l'esistenza di una logica "naturale" sottostante al linguaggio comune, che stia alla base della capacità umana di inferire e di pensare. Per realizzare questo progetto egli ha analizzato la logica sillogistica, individuando un legame tra il principio di monotonicità e la teoria della distribuzione della Scolastica medievale, riuscendo così a motivare perché le inferenze che venivano svolte nell'ambito della logica tradizionale fossero corrette. Secondo Van Benthem, infatti, il principio di monotonicità dovrebbe spiegare perché la sillogistica autorizzasse una sostituzione di predicati con predicati con una estensione più grande o più piccola. L'esempio presente in letteratura che mostra l'inadeguatezza della sillogistica medievale e la sua inferiorità nei confronti della logica moderna di Boole e Frege risale a De Morgan ed è il seguente:

(3) Ogni cavallo è un animale. Dunque ogni coda di cavallo è la coda di un animale

De Morgan aveva osservato che la logica sillogistica non riesce a rendere conto di inferenze come (3) poiché per capirne la validità bisognerebbe ricorrere a relazioni binarie, mentre la logica tradizionale era basata su predicati monadici. Invece, attraverso il linguaggio logico del primo ordine, è possibile mostrare la validità dell'argomento nel seguente modo, dove H=cavallo, A=animale e T=coda:

$$\frac{\forall x(Hx \rightarrow Ax)}{\forall x((Tx \wedge \exists y(Hy \wedge Rxy)) \rightarrow (Tx \wedge \exists y(Ay \wedge Rxy)))}$$

Gli antichi tuttavia effettuavano inferenze analoghe a (3) come la seguente:

$$\frac{\text{La grammatica è un'arte}}{\text{Colui che impara la grammatica impara un'arte}}$$

Come già Sanchez Valencia aveva affermato, queste inferenze nella logica tradizionale erano valide poiché il sillogismo veniva applicato in un senso *ampio*, per il quale valgono alcune leggi sillogistiche aggiuntive, ossia che la specie (il termine più piccolo) può prendere il posto del genere (il termine più ampio) quando si parla di tutto il genere oppure che il genere (il termine più ampio) può prendere il posto della specie (termine più piccolo) quando qualcuna delle specie è menzionata. Anche se non si dovesse condividere con van Benthem l'esistenza di una logica "naturale" sottostante al linguaggio comune, secondo Mugnai, questo approccio ha comunque il merito importante di rivalutare alcuni aspetti della logica tradizionale, poiché talvolta la concezione moderna, basata sulla nozione di sistema formale, non rende giustizia di come la sillogistica funzionava realmente.

Riferimenti bibliografici

- George Boole (1847). *The Mathematical Analysis of Logic, Being an Essay Towards a Calculus of Deductive Reasoning*. Macmillan, Barclay, & Macmillan. URL: http://www.gutenberg.org/ebooks/36884. Reprinted in Oxford by Basil Blackwell, 1951
- Clarence Irving Lewis (1912). "Implication and the Algebra of Logic". In: *Mind* 21, pp. 522–531
- Hugh McColl (1880). "Symbolical reasoning". In: *Mind* 5.17, pp. 45–60
- Massimo Mugnai (2013). *Possibile necessario*. Bologna: Il Mulino
- Victor Sànchez Valencia (1997). "Head or Tail? De Morgan on the bounds of traditional logic". In: *History and Philosophy of Logic* 18, pp. 123–138
- Johan van Benthem (2008). "A Brief History of Natural Logic". In: *Technical Report PP-2008-05*, pp. 123–138

2 *Bisimulazione e coinduzione*

Prof. Davide Sangiorgi

Davide Sangiorgi nel suo corso ha presentato una introduzione ai concetti di bisimulazione e coinduzione[3], privilegiando il loro utilizzo come tecniche di prova per stabilire una uguaglianza tra processi. In questo report saranno presentate le nozioni di base del concetto di bisimulazione, sintetizzando il contenuto delle prime lezioni del prof. Davide Sangiorgi.

Per avere una idea intuitiva dell'esigenza di introdurre la tecnica di bisimulazione, si immagini di avere una macchina per il caffè, molto semplice, con una apertura dove mettere i soldi e due tasti, che permettono di scegliere rispettivamente tè o caffè. Dopo aver inserito la moneta si può richiedere la bevanda premendo, a seconda della propria scelta, il tasto del tè o il tasto del caffè. Immaginiamo quindi che sulla macchina sia presente una etichetta che spiega il comportamento della macchina nel modo seguente:

- Inserisci la moneta.
- Dopo avere inserito la moneta puoi premere il tasto del tè oppure il tasto del caffè.
- Dopo che hai premuto il tasto del caffè ottieni il caffè.
- Dopo che hai premuto il tasto del tè ottieni il tè.
- Dopo che la bevanda è stata erogata, la macchina è pronta per un nuovo servizio.

Immaginiamo che ad un certo punto la macchina si rompa e che la ditta che aveva prodotto la macchina guasta sia fallita. A questo punto ci si rivolgerà ad una nuova ditta per ordinare una macchina nuova capace di erogare tè o caffè. La ditta designata allora fornisce una nuova macchina, che tuttavia funziona diversamente dalla precedente. Infatti, dopo aver inserito la moneta essa eroga non deterministicamente tè o caffè, quando il tasto raffigurante il tè o il caffè viene premuto. Essa può quindi erogare correttamente la bevanda selezionata, ma anche servire tè quando si è premuto il tasto del caffè o il caffè quando si è premuto il tasto del tè.

A questo punto immaginiamo di chiamare la ditta che ha fornito la macchina, chiedendo di volerne un altra perché questa non si comporta come la precedente che avevamo richiesto. La ditta tuttavia non accetta di sostituirla, perché a suo avviso sostiene di avere fornito una macchina che soddisfaceva le precedenti richieste, infatti la possibilità di premere un tasto del caffè e uno del tè e di bere la bevanda erogata viene da essa garantita.

[3] Per una esposizione precisa e completa leggere (Sangiorgi 2012).

Grazie a questo esempio ci si rende conto della esigenza di poter esprimere quando due processi hanno un comportamento equivalente. È bene ricordare che nel cercare questa relazione non si è interessati a dettagli riguardanti forma o colore della macchine, bensì al loro comportamento. Una descrizione del comportamento di una macchina di questo tipo si può rendere con i Labelled Transition System (LTS).

Un Labelled Transition System è una tripla $\langle P, Act, T\rangle$ dove:

- P è l'insieme (non vuoto) di stati o di processi;
- Act è l'insieme delle azioni (eventualmente infinito);
- $T \subseteq \langle P, Act, P\rangle$ è la relazione di transizione.

Si scrive quindi $P \xrightarrow{\mu} P'$ se $(P, \mu, P') \in T$ quando il processo P accetta una interazione con l'ambiente e effettua l'azione μ per diventare il processo P'. P' è un derivato di P se ci sono $P_1, \ldots, P_n, \mu_1, \ldots, \mu_n$ tale che $P \xrightarrow{\mu_1} P_1 \ldots \xrightarrow{\mu_n} P_n$ e $P_n = P'$.

Un LTS dice quindi quali sono gli stati o i processi in cui un sistema può essere e, per ogni stato, le interazioni possibili. Il comportamento della prima macchina caffè quindi può essere rappresentato come LTS nel modo seguente:

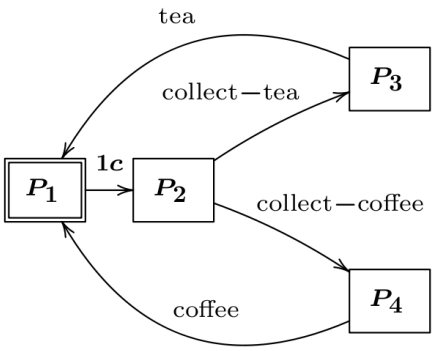


Figura 1: LTS (1).

In questo esempio quindi si ha un insieme di processi non vuoto, ossia $\{P_1, P_2, P_3, P_4\}$, delle azioni, che in questo caso sono:

$$\{1c, collect - tea, collect - coffee, tea, coffee\}$$

e relazioni di transizione, ossia:

$$\{(P_1, 1c, P_2), (P2, collect{-}tea, P_3), (P_2, collect{-}coffee, P_4), (P_4, coffee, P_1), (Q_3, tea, P_1)\}$$

Il comportamento della seconda macchina è invece rappresentato dal seguente LTS:
Intuitivamente (1) e (2) sono macchine che esprimono un comportamento diverso, poiché quello che intendiamo per uguaglianza tra due macchine è la possibilità di eseguire la stessa operazione con la prima e la seconda macchina, e

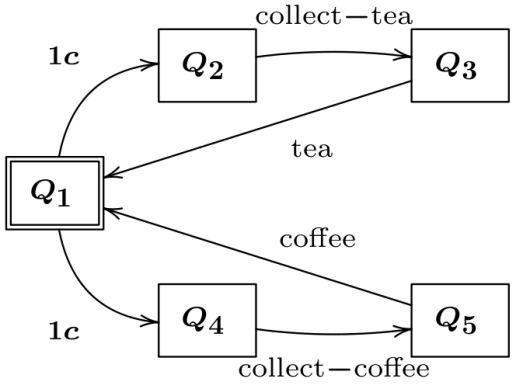


Figura 2: LTS (2).

lo stesso anche per i due stati in cui le macchine evolvono. Si può formalizzare quindi il concetto di bisimulazione e di bisimilarità nel seguente modo:

> Si definisce una bisimulazione, in un singolo LTS, come la relazione $\mathcal{R}$ su processi se ogniqualvolta $P\mathcal{R}Q$:
>
> 1. $\forall \mu$, P' tale che $P \xrightarrow{\mu} P'$, allora $\exists Q'$ tale che $Q \xrightarrow{\mu} Q'$ e $P'\mathcal{R}Q'$;
> 2. $\forall \mu$, Q' tale che $Q \xrightarrow{\mu} Q'$, allora $\exists P'$ tale che $P \xrightarrow{\mu} P'$ e $P'\mathcal{R}Q'$.
>
> P e Q sono bisimili, scritto $P \sim Q$ se $P\mathcal{R}Q$ per qualche bisimulazione $\mathcal{R}$.

La definizione data sopra dà origine ad una tecnica di prova per verificare che due processi sono bisimili. Siano date le seguenti figure:

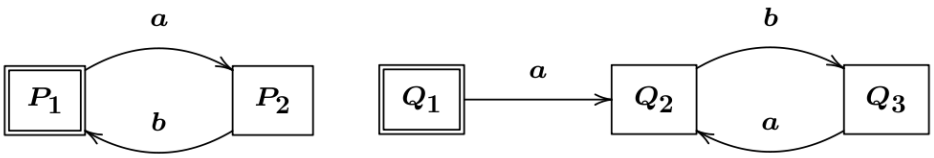


Figura 3: LTS (3), a sinistra, e LTS (4), a destra.

Per provare che sia $P_1 \sim Q_1$ bisogna trovare una relazione $\mathcal{R}$ di bisimulazione che contenga la coppia (P_1, Q_1). Affinché una relazione $\mathcal{R}$ sia una bisimulazione, tutti i derivati di P_1 e Q_1 devono apparire in $\mathcal{R}$, come da definizione. Si supponga di voler definire $\mathcal{R} = \{(P_1, Q_1), (P_2, Q_2)\}$. Si avrà quindi il seguente diagramma di bisimulazione per (P_1, Q_1):

$$\begin{array}{ccc} P_1 & \mathcal{R} & Q_1 \\ {\scriptstyle a}\downarrow & & {\scriptstyle a}\downarrow \\ P_2 & \mathcal{R} & Q_2 \end{array} \qquad \begin{array}{ccc} P_1 & \mathcal{R} & Q_1 \\ {\scriptstyle a}\downarrow & & {\scriptstyle a}\downarrow \\ P_2 & \mathcal{R} & Q_2 \end{array}$$

Per la coppia (P_2, Q_2) tuttavia non è possibile trovare una relazione di bisimulazione, dal momento che un derivato di Q_2, in questo caso Q_3, rimane scoper-

to. Mentre effettuando una transizione da P_2 si ottiene P_1, l'unica transizione possibile da Q_2 è $Q_2 \xrightarrow{b} Q_3$, e la coppia (P_1, Q_3) non appartiene a $\mathcal{R}$.

$$\begin{array}{ccc} P_2 & \mathcal{R} & Q_2 \\ {\scriptstyle b}\downarrow & & {\scriptstyle b}\downarrow \\ P_1 & \not{\mathcal{R}} & Q_3 \end{array} \qquad \begin{array}{ccc} P_2 & \mathcal{R} & Q_2 \\ {\scriptstyle b}\downarrow & & {\scriptstyle b}\downarrow \\ P_1 & \not{\mathcal{R}} & Q_3 \end{array}$$

Aggiungendo la coppia (P_1, Q_3) si ottiene invece una bisimulazione. Infatti se:

$$\mathcal{R} = \{(P_1, Q_1), (P_2, Q_2), (P_1, Q_3)\}$$

la relazione $\mathcal{R}$ nel diagramma precedente è verificata, e per la coppia (P_1, Q_3) si avrà che:

$$\begin{array}{ccc} P_1 & \mathcal{R} & Q_3 \\ {\scriptstyle a}\downarrow & & {\scriptstyle a}\downarrow \\ P_2 & \mathcal{R} & Q_2 \end{array} \qquad \begin{array}{ccc} P_1 & \mathcal{R} & Q_3 \\ {\scriptstyle a}\downarrow & & {\scriptstyle a}\downarrow \\ P_2 & \mathcal{R} & Q_2 \end{array}$$

Dato che (P_2, Q_2) appartiene a $\mathcal{R}$, per definizione segue che $P_1 \sim Q_1$.

Durante il corso questi concetti sono stati sviluppati ulteriormente e si è insistito sull'utilità di queste tecniche, utilizzate non solo in informatica, ma anche in intelligenza artificiale, scienze cognitive, matematica, filosofia e fisica, prevalentemente per spiegare fenomeni che coinvolgono un certo tipo di circolarità. In informatica per esempio la bisimulazione è prevalentemente utilizzata in teoria della concorrenza e nel *model checking*, in filosofia negli ambiti di ricerca che fanno uso della logica modale, in matematica, per esempio, nello studio di insiemi che non soddisfano l'assioma di regolarità, in fisica nello studio di modelli di sistemi quantistici.

La bisimulazione[4] è un ambito di ricerca particolarmente recente, e questo è dovuto in parte al fatto che, benché quando la teoria degli insiemi venne assiomatizzata da Zermelo rimanesse ancora aperta la possibilità di definizioni che coinvolgessero una certa forma di circolarità, dopo la scoperta dell'insorgere di paradossi come quello di Russell o di Burali-Forti si cercò di rigettare qualsiasi forma di circolarità. Si affermò quindi la teoria dei tipi proposta da Russell che permette di costruire solamente costruzioni stratificate, eliminando qualsiasi circolarità. La forte influenza di questo approccio stratificato ha contribuito a ritardare la scoperta della bisimulazione, che avvenne solamente negli anni '70 indipendentemente in informatica, matematica e logica modale. Fu scoperta in informatica in seguito ai lavori di Hennessy e Milner nello studio di processi in teoria della concorrenza, in teoria degli insiemi in alcuni studi intrapresi per formalizzare una nuova fondazione per la matematica che ammettesse l'esistenza di insiemi non ben fondati, e in logica modale per studiarne l'espressività.

[4] Per approfondire, si veda (Sangiorgi 2009).

Questo ambito di ricerca inoltre è ancora molto fertile, alcuni problemi, per esempio, riguardano la bisimulazione di linguaggi di ordine superiore, il suo sviluppo come metodo di prova, linguaggi con costrutti probabilistici o nozioni unificanti. Bisimulazione e coinduzione inoltre sono concetti con una forte natura interdisciplinare, che possono essere applicati in ambiti diversi e che quindi possono anche permettere di comprendere alcune analogie e similitudini tra fenomeni che a prima vista possono sembrare molto diversi tra di loro.

Riferimenti bibliografici

- Davide Sangiorgi (2009). "On the origins of bisimulation and coinduction". In: *ACM Transactions on Programming Languages and Systems (TOPLAS)* 31.4, pp. 1–41
- Davide Sangiorgi (2012). *Introduction to bisimulation and coinduction.* Cambridge: Cambridge University Press

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

Logica

Graham Priest

[Codice Edizioni, Torino 2012]

recensione a cura di Matilde Aliffi

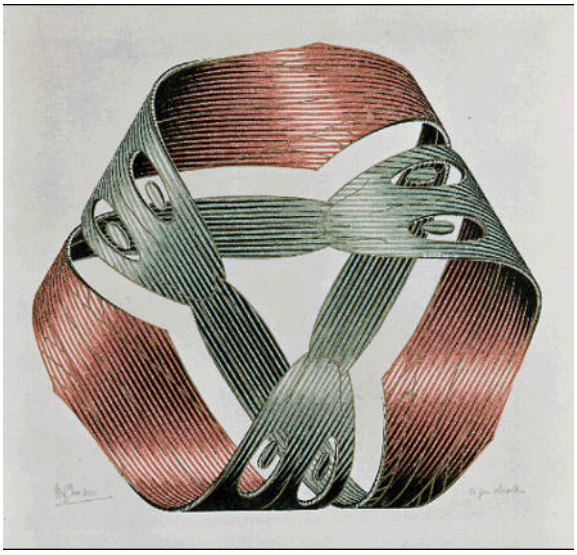


Figura 1: MÖbius Strip I, 1961, Escher

Può una proposizione essere contemporaneamente vera e falsa? Graham Priest, logico inglese professore all'università di Melbourne, afferma che in alcune particolari proposizioni questo è possibile. Come in un nastro di Möbius a causa di una torsione la parte interna del nastro diventa esterna, e quella esterna interna, così in alcuni enunciati il vero e il falso sembrano tramutarsi l'uno nell'altro (cfr. Priest 2012, p. 46); proposizioni come «Questa frase che sto pronunciando è falsa», infatti, sembrano essere sia vere che false, poiché, come si nota, se la proposizione è vera, essa allora dev'essere falsa, ma se è falsa, allora è vera. Accettare tuttavia che una proposizione possa essere sia vera sia falsa significa ammettere delle contraddizioni, e questo genera numerosi problemi.

Copyright. © 2018 Matilde Aliffi. Pubblicato in Italia. Alcuni diritti riservati. Prima pubblicazione su *Rivista Italiana di Filosofia Analitica Junior* 4:1 (2013), pp. 67-72.
Autore. Matilde Aliffi, mxa654@bham.ac.uk.

Che fare, per esempio, del *principio di non contraddizione* e della *legge del terzo escluso*?

Questa «faccenda molto ingarbugliata» (Priest 2012, p. 53) ed altri problemi sono trattati in *Logica*, agevole testo edito da Codice Edizioni, traduzione di *Logic. A Very Short Introduction* (Oxford University Press, 2000). Il libro è breve, chiaro e discorsivo, capace di avvicinare il lettore ai problemi principali della logica moderna evitando tecnicismi, o notazioni complesse. Autoreferenza, vaghezza, logica *fuzzy* e teoria delle decisioni sono alcuni dei temi trattati, che vengono presentati anche attraverso riferimenti classici e letterari, e un accurato uso delle immagini. Il tentativo di Priest è quello di dimostrare che la logica è una materia viva, in evoluzione, e per questo solleva questioni e problemi, senza pretendere di presentare una visione definitiva della materia. Al libro non compete fornire risposte, ma problemi, che invoglino il lettore curioso ad approfondire le questioni che più lo hanno interessato. Per questo nell'appendice del testo Priest fornisce anche dei rimandi utili a letture più approfondite e inserisce anche alcuni esercizi risolti.

Nonostante ciò, questo libro non si può considerare propriamente una introduzione allo studio della logica, poiché la prospettiva seguita non è quella classica. La legge di non contraddizione (LNC), secondo la quale ogni proposizione non può essere contemporaneamente vera e falsa, e la legge del terzo escluso (LTE), che afferma che il valore di verità di una proposizione è sempre opposto a quello della proposizione contraddittoria, infatti, dopo essere state presentate nei primi capitoli vengono messe in dubbio a seguito dell'analisi dei paradossi basati sull'autoreferenza e sulla vaghezza.

Come si è visto, infatti, nel caso dell'enunciato «questa frase che sto pronunciando è falsa» sembra non sia possibile sottrarsi da una continua alternanza tra V e F.

Tuttavia Priest, invece di cercare di eliminare la contraddizione in esso presente, si limita a riconoscere la presenza della contraddizione, e ad accettare il paradosso come un dato di fatto. Dunque «assumiamo che, in qualsiasi situazione, ogni proposizione può essere vera ma non falsa, falsa ma non vera, sia vera sia falsa, né vera né falsa» afferma Priest (2012, p. 49), ammettendo quindi che ci possano essere delle lacune di valori di verità (*truth-value gaps*), nel caso di una proposizione né vera né falsa, e una situazione invece in cui ci siano eccessi di valori di verità (*truth-value gluts*), come nel caso di una proposizione sia vera che falsa. La legge del terzo escluso e la legge di non contraddizione quindi sono violate, e in questo risiede la principale differenza tra la logica classica e quella non classica.

Secondo Graham Priest infatti, nel linguaggio naturale esistono vere contraddizioni, chiamate "dialetheie", dal greco *dialétheia*, ossia "doppia verità". Una *dialétheia* è secondo Priest una affermazione vera della forma «α e non è il ca-

so che α». (cfr. Priest 2006b, p. 4). Priest infatti rappresenta uno dei principali esperti mondiali del dialeteismo, ossia di una concezione che ammette l'esistenza di enunciati che siano contemporaneamente veri e falsi, e così questo testo, anche se privo di un diretto riferimento al dialeteismo, si può considerare perfettamente inscritto in questa prospettiva.

La violazione della LNC tuttavia comporta numerosi problemi; già Aristotele avvertiva infatti nel quarto libro della *Metafisica* che chi nega il principio di non contraddizione non riesce più a dire nulla, perché disdice tutto quello che dice, e non gli è quindi più possibile pensare. «E se non sostiene nulla, ma crede e non crede in egual modo, che differenza ci sarà tra lui e le piante?» (Aristotele, *Met. IV*, p. 52). Priest tuttavia non afferma che tutto sia contemporaneamente vero e falso, ma che esistono solo *alcune* contraddizioni. La difficoltà del dialeteista consiste allora nel mostrare come sia possibile accettarne solo alcune (cfr. D'Agostini 2011, pp. 148–158). Il problema è quello classico dell'«esplosione», ossia del fatto che se si accetta una contraddizione $p \wedge \neg p$, sembra poi possibile derivare qualsiasi cosa, e quindi la logica sembra esplodere, perché tutto diventa vero. Infatti, se $p \wedge \neg p$ è vera, lo è anche p, e quindi qualsiasi disgiunzione $p \vee q$ in cui figuri p è vera. Ma secondo la logica classica, data una disgiunzione vera, e la negazione di un congiunto ($\neg p$), ne segue la verità dell'altro disgiunto. Ma a questo punto qualsiasi falsità potrebbe essere dimostrata, da una contraddizione quindi si potrebbe derivare qualsiasi proposizione, e la logica potrebbe esplodere (cfr. D'Agostini 2011, pp. 153–154).

In *Logica* Priest risponde a questo tipo di obiezione affermando che il sillogismo disgiuntivo funziona solo nel caso in cui se p è vero, $\neg p$ è falso. Ammettendo una violazione del principio di non contraddizione cambia infatti anche il modo di intendere la negazione. Secondo le regole classiche la verità di $\neg r$ esclude la verità di r, mentre invece nella prospettiva dialeteista «la verità di $\neg r$ non esclude quella di r. Ciò avverrebbe solo nel caso in cui fosse impossibile per una proposizione essere sia vera, sia falsa». L'esempio presentato nel testo è il seguente:

$$\frac{\text{La regina è ricca o i maiali possono volare} \qquad \text{La regina non è ricca}}{\text{I maiali possono volare}}$$

In simboli:

$$\frac{r \vee m \qquad \neg r}{m}$$

Questa inferenza nella prospettiva classica è valida poiché non si da il caso in cui entrambe le premesse sono vere e la conclusione invece è falsa, come è reso evidente da questa tavola di verità (cfr. Priest 2012, p. 19):

Se invece accettiamo che r è sia vero che falso allora $\neg r$ non esclude la verità di r e quindi il ragionamento non può considerarsi valido. Infatti nella prospettiva dialeteista la contraddizione non è né priva di contenuto, come sostenuto per

r	***m***	***r*** ∨ ***m***	¬***r***	***m***
V	V	V	F	V
V	F	V	F	F
F	V	V	V	V
F	F	F	V	F

esempio da Aristotele, né ha un contenuto totale nel senso della complementazione, come sostenuto dalla logica intuizionista; la contraddizione possiede invece un contenuto parziale, né nullo né totale (Priest 2006a, p. 31).

In *Logica* sono inoltre trattati altri paradossi come quelli del *sorite*, che si generano quando il predicato utilizzato è vago, ossia quando la sua applicabilità è tollerante a modifiche piccole (cfr. Priest 2012, p. 95). Predicati come «essere un mucchio» o «essere alto» oppure «essere giovane» non hanno infatti confini precisi. Dato un mucchio di sabbia, infatti, se si elimina un granello dal mucchio avremo ancora un mucchio, e così se si elimina anche un altro granello. Tuttavia, eliminando ancora un granello, e poi ancora uno, il mucchio diventerà sempre più piccolo, finché rimarrà un solo granello di sabbia. È ancora un mucchio, quando rimane un solo granello? E se un solo granello non è un mucchio, allora in quale momento quel mucchio iniziale non è più un mucchio?

Il paradosso del sorite ha quindi questa struttura:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{a_0 \qquad a_0 \to a_1}{a_1} \qquad a_1 \to a_2}{a_2} \qquad \dots}{\dots} \qquad \dots}{a_{k-1}} \qquad a_{k-1} \to a_k}{a_k}$$

La risposta data da Priest a questo problema è l'uso della logica *fuzzy*. Questi predicati enunciano proprietà che gradualmente si dissolvono, così come il valore di verità di «Jack è giovane» si trasforma gradualmente dal vero al falso. Questi gradi possono essere misurati con numeri compresi tra 0 e 1, dove 0 è la completa falsità e 1 la completa verità. Con l'avanzare del tempo, il valore di verità di «Jack è giovane», cui prima avevamo attribuito il valore 1, lentamente diminuisce, fino ad arrivare a 0.

a	**¬a**
1	0
0,75	0,25
0,5	0,5
0,25	0,75
0	1

Mentre per la logica classica una congiunzione $p \wedge q$ è vera, ossia ha valore 1, solo se entrambi p e q hanno valore 1, nella logica *fuzzy* se p e q hanno valori

intermedi, la congiunzione ha il valore minimo. Così, anche il condizionale, se l'antecedente è meno vero del conseguente, è completamente vero. Mentre, se l'antecedente è più vero del conseguente, il condizionale è pari a 1 diminuito della differenza tra i due valori di verità. In simboli:

$$\text{Se } |a| \leq |b|\text{: } |a \rightarrow b| = 1$$

$$\text{Se } |b| < |a|\text{: } |a \rightarrow b| = 1 - (|a| - |b|)$$

Una proposizione per Priest è vera in una situazione se e solo se il suo valore di verità è almeno pari ad un certo livello di accettabilità, chiamato, per esempio, ϵ. Esso funge da standard epistemico, fissando quindi il limite a partire dal quale un enunciato è considerato vero. Quale sia il valore esatto di ϵ può essere stabilito dal contesto.

Con la logica *fuzzy* Priest riesce mostrare che il paradosso non regge. Infatti, se si considera la proposizione «Jack è un bambino dopo n secondi», e suppone che il bambino cresca in soli quattro secondi, si avrà una questo registro di valori di verità, dove a_n è la proposizione al secondo n-esimo.

$\mathbf{a_0}$	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$	$\mathbf{a_4}$
1	0,75	0,5	0,25	0

Il *modus ponens*, ossia $p \rightarrow q$, $p \vdash q$, alla base della catena di inferenze che costituisce il paradosso è valido solo se la soglia di accettabilità fissata ϵ è uguale a 1, poiché:

$$|a| = 1, \text{ e } |a| \leq |b|, \text{ da cui segue che } |b| = 1$$

Tuttavia nell'esempio preso qui in considerazione, come fa notare Priest, ciascun condizionale che funge da premessa nel *modus ponens* ha valore inferiore a 1. Infatti $a_0 \rightarrow a_1$ ha valore 0,75, infatti $a_1 < a_0$ e quindi 1–(1–0,75)=0,75. Anche fissando una soglia di accettabilità ϵ pari a 0,75 l'argomento non funziona, infatti sia a_1 che $a_1 \rightarrow a_2$ hanno valore 0,75, e quindi secondo il nostro standard epistemico sono veri, tuttavia a_2 ha come valore 0,5, che è minore di ϵ. La conclusione quindi non è vera, e l'inferenza non è valida.

I problemi che questo approccio genera comunque non sono pochi. Innanzitutto nelle situazioni di confine, dove a e $\neg a$ hanno valori di verità coincidenti, è possibile che si abbia quindi una *dialétheia* (cfr. D'Agostini 2011, p. 170). Inoltre, come afferma Priest a conclusione del capitolo trattato, qualsiasi valore di ϵ si scelga, esso è completamente arbitrario, e il problema della vaghezza anche in questa nuova formulazione non può dirsi risolto, ma solo spostato. Infatti, quando una proposizione come «Jack è un bambino» modifica il suo valore di verità, passando da 1 ad un valore minore di 1? La questione rimane aperta.

Questi sono solo alcuni degli argomenti trattati in *Logica*, libro composto da quattordici capitoli, ognuno dei quali destinato ad un tema specifico. Non un

manuale, quindi, ma un'introduzione alla profondità delle questioni della logica moderna, che mira a porre più interrogativi che risposte. Un libro utile per chi voglia lasciarsi appassionare dalla logica, in particolare dai suoi ultimi sviluppi, in un modo divertente e ricco di stimoli.

Riferimenti bibliografici

Aristotele. *Il principio di non contraddizione. Libro quarto della Metafisica*. A cura di Emanuele Severino. Brescia: La scuola 2001.

D'Agostini, Franca (2011). *Introduzione alla verità*. Torino: Bollati Boringhieri.

Priest, Graham (2006a). *Doubt truth to be a Liar*. Oxford: Oxford Scholarship Online.

— (2006b). *In contradiction*. New York: Oxford University Press.

— (2012). *Logica*. Torino: Codice Edizioni.

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

Possibile/necessario

Massimo Mugnai

[Il Mulino, Bologna 2013]

recensione a cura di Mattia Cozzi

Possibile/Necessario, edito da Il Mulino all'interno della collana "Lessico della filosofia", è prima di tutto una *prospettiva storica* sui due concetti modali richiamati nel titolo. In un numero piuttosto esiguo di pagine (poco più di 200), Mugnai ripercorre la storia della possibilità e della necessità a partire dall'antichità (con riferimento in particolare ad Aristotele e alla scuola stoica) e arrivando fino al '900, con la fondamentale ripresa del dibattito sulla modalità ad opera di autori come Carnap, Hintikka, Lewis, Kripke, Stalnaker, Plantinga, Goodman e Quine, per citare quelli che ricevono maggiore spazio nella parte finale del volume.

L'opera di Mugnai si propone come un testo per *non specialisti*: fin dalle prime pagine, infatti, vengono introdotti alcuni termini "tecnici" spiegati concisamente, ma nonostante ciò riuscendo a dare tutte le informazioni richieste per la comprensione degli argomenti trattati (Mugnai in questo senso riesce ad ottenere un'ottima sintesi tra leggibilità da parte di un lettore non-tecnico e corretta specificazione dei termini). Per citare alcuni esempi di questo modo di procedere, si vedano la p. 8 (caratterizzazione generale dei concetti modali), le pp. 9-11 (distinzione tra "proposizione" ed "enunciato"), le pp. 37-38 ("*axíoma*" e "*lektón*"), la p. 120 (spiegazione del concetto di "dimostrazione") e le pp. 146-153 ("calcolo degli dei predicati", "calcolo degli enunciati", "formula ben formata", "tavola di verità", "relazione binaria", "modello", "valutazione", ecc.).

Mugnai pone fin da subito l'accento sul fatto che il suo interesse in questo libro è legato «prevalentemente all'aspetto *logico-gnoseologico*» dell'analisi dei concetti modali (Mugnai 2013, p. 8, corsivo nell'originale), trattando in modo più marginale l'aspetto *etico-teologico* della possibilità e della necessità.

Il primo dei nove capitoli del libro è dedicato alla trattazione dei concetti modali da parte di Aristotele e della scuola stoica, come anticipato. Il primo,

Copyright. 2018 Mattia Cozzi. Pubblicato in Italia. Alcuni diritti riservati. Prima pubblicazione su *Rivista Italiana di Filosofia Analitica Junior* 5:1 (2014), pp. 88-93.
Autore. Mattia Cozzi, cozzimattia@gmail.com.

una volta messa in luce da parte di Mugnai la non distinzione tra i concetti di "logicamente possibile" e "possibile nel mondo attuale", definisce il necessario come "ciò che è impossibile che sia altrimenti", secondo l'equivalenza tra $\Box\alpha$ e $\neg\Diamond\neg\alpha$, riuscendo così ad ottenere il *quadrato modale*, analogo modale del classico quadrato delle proposizioni categoriche:

A	E
$\Box\alpha, \neg\Diamond\neg\alpha$	$\Box\neg\alpha, \neg\Diamond\alpha$
$\neg\Box\neg\alpha, \Diamond\alpha$	$\neg\Box\alpha, \Diamond\neg\alpha$
I	O

Aristotele stabilisce inoltre un'inferenza tra l'esistenza in ogni tempo e la necessità e «siccome in altre circostanze Aristotele inferisce dalla necessità di certi enti la loro eternità» (Mugnai 2013, p. 25) possiamo far valere l'equivalenza tra "necessario" e "vero o esistente in ogni tempo". Introduciamo così un nuovo modo di valutare i concetti modali, cioè facendo riferimento al tempo. Quest'ultimo approccio è detto *interpretazione statistico-frequentista della modalità*: si intenderà quindi "possibile" come "vero in qualche tempo" e "necessario" come "vero in ogni tempo" (a partire da osservazioni come quelle appena fatte, viene successivamente proposta la concezione *diacronica* della modalità, trattando anche del concetto di "contingenza").

Dopo aver trattato il problema dei futuri contingenti in Aristotele in riferimento alla modalità (citando il famoso passo del *De interpretatione* in cui lo Stagirita si chiede come si possano valutare enunciati come "Domani ci sarà una battaglia navale"), si passa, sempre nel primo capitolo, alla dottrina stoica della modalità, o meglio, a quanto è possibile sapere di essa a partire dalle scarse fonti disponibili. A partire dal commento di Severino Boezio al *De interpretatione*, vengono esposte le dottrine di Filone di Megara e di Diodoro Crono. Il primo avrebbe sostenuto che possibile è ciò che è suscettibile di esser vero, *indipendentemente* da circostanze esterne: ci sarebbero pertanto proposizioni possibili eppure false[1]; il secondo si rifarebbe invece ad una concezione diacronica della modalità, secondo la quale sarebbe possibile ciò è vero o che sarà vero, escludendo pertanto l'esistenza di possibilità che mai si realizzeranno. Si noti in particolare che Filone tratta la modalità facendo riferimento al contenuto intrinseco degli *axíomata* (nella nota 64, a p. 50, si parla di "struttura lektologica dell'*axíoma*").

Il secondo capitolo è dedicato alla trattazione medievale della modalità, avente come intermediario con l'antichità ancora Boezio e di discendenza principalmente aristotelica (anche se altre concezioni, come quella stoica, giocano co-

[1] L'esempio proposto è quello di "quel pezzo di legno brucia", che risulta possibile, anche se falsa, nel momento in cui quel pezzo di legno si trova immerso nell'acqua.

munque un ruolo di tutto rilievo). Uno dei limiti e al contempo delle particolarità della logica medievale è quello di utilizzare in massima parte una sezione molto specializzata del linguaggio naturale, le proposizioni categoriche, del tipo "S è P", cui vengono ricondotti tutti gli altri tipi di proposizioni. Questa concezione della logica sarà molto influente, tanto che si può ritrovare ancora nel XIX secolo (si pensi ad esempio a *The Mathematical Analysis of Logic* di Boole, del 1847), anche se già nel Medioevo si possono trovare indizi in merito alla differenza tra forma logica e forma grammaticale. Tema fondamentale di questo secondo capitolo è la distinzione tra modalità *de rebus* e *de sensu* in Abelardo, vicina a quella tra modalità *de re* e *de dicto* nello Pseudo-Tommaso (Mugnai è peraltro molto chiaro qui nell'esplicitare le differenze tra modalità *de re*/*de dicto* in epoca medievale e in epoca "post-fregeana"). Sempre nel capitolo dedicato al Medioevo, viene trattata la modalità secondo Guglielmo di Ockham, il quale, con una sensibilità che l'autore di questa recensione ha trovato sorprendentemente moderna, «riconduce la necessità di una proposizione al suo *esser vera e all'impossibilità di essere falsa, a meno che non muti il significato dei termini*» (Mugnai 2013, p. 78, corsivo nell'originale).

Dopo il breve terzo capitolo, un rapido *excursus* sulla modalità nel Quattro-Cinquecento, il capitolo 4 compie un piccolo passo indietro, introducendo la modalità in riferimento al Cristianesimo e al suo Dio creatore-architetto, questione che a sua volta pone il problema dell'onniscienza e dell'onnipotenza divina. Questo capitolo permette a Mugnai di introdurre almeno due concetti fondamentali: quello di *possibile logico*[2] e quello di *mondo possibile*. Il mondo attuale è quindi quello che Dio ha deciso di creare, mentre gli altri restano delle possibilità inespresse, che restano *in mente Dei* (si noti che si è così passati da una concezione diacronica della possibilità ad una sincronica).

Il capitolo seguente è interamente dedicato al concetto di "mondo possibile" e in particolare al lavoro di Leibniz, per molti versi vicino a Duns Scoto e al suo "possibile logico". «Per certi versi, si potrebbe dire addirittura che Leibniz sistematizzi e amplifichi la concezione scotista, inserendola in una metafisica incentrata sul concetto di *mondo possibile*» (Mugnai 2013, p. 111, corsivo nell'originale). Fondamentali per comprendere la posizione di Leibniz sono i "concetti completi", aggregati non contraddittori (e quindi logicamente possibili) di concetti semplici. Se i concetti completi sono descrizioni esaustive di individui logicamente possibili, i mondi possibili sono insiemi non contraddittori di concetti completi. Due concetti completi sono inoltre detti "compossibili" se la loro congiunzione non genera contraddizioni. La questione immediatamente successiva è quella dell'analiticità: nella ricostruzione del pensiero di Leibniz

[2]Duns Scoto nel XIII secolo collega il possibile logico alla *non ripugnanza* tra i concetti, che tuttavia non deve essere confusa con l'*immaginabilità*: Scoto ritiene infatti che esistano cose immaginabili che sono tuttavia impossibili.

proposta da Mugnai, l'analiticità di una proposizione non ne implica la necessità, per via del *concetto analitico di verità* (ovvero l'idea per cui una proposizione risulta vera se la nozione del predicato è contenuta nella nozione del soggetto, come "Un parallelogrammo ha angoli opposti congruenti"). Due sono le proposte di Leibniz in merito alla questione: la prima distingue due generi di necessità, mentre la seconda, che nell'opinione di chi scrive è decisamente la più interessante, fa riferimento alla *dimostrazione*[3].

La trattazione prosegue andando ad analizzare il rapporto che esiste tra modalità e condizionale; così si esprime l'autore:

> Fin dagli inizi della riflessione sulla logica, già al tempo degli stoici, ci si è accorti che esistono vari tipi di condizionale e ben presto emerge il legame che almeno un tipo di essi mantiene con le nozioni di possibile e necessario. (Mugnai 2013, p. 127)

Sia Filone di Megara, sia Diodoro Crono, sia Crisippo di Soli avevano una propria visione del condizionale: il primo propone quello che oggi chiamiamo "condizionale materiale", il secondo lega la verità di una proposizione ipotetica al tempo, mentre il terzo introduce il concetto di *incompatibilità* (Mugnai interpreta: "incompatibilità logica"). Un'altra dottrina (della quale è difficile dire la provenienza, se dalla scuola stoica o se da Sesto Empirico, la fonte principale cui si fa riferimento) collega la verità del condizionale al "contenimento in potenza" del conseguente nell'antecedente[4]. Di tali idee si trova poi una eco nel Medioevo, ancora per opera di Boezio, che nel *De hypoteticis syllogismis* distingue tra condizionale accidentale e condizionale "naturale", ripresa poi da Abelardo nella distinzione tra inseparabilità di natura e inseparabilità concettuale. La prima dipende dal modo in cui le cose del mondo stanno, mentre la seconda dipende dai collegamenti di ordine concettuale tra i termini utilizzati. Chiarissime le parole di Mugnai in merito:

> Mentre per determinare le condizioni di verità di un *condizionale filoniano* è sufficiente tener conto dei soli valori di verità di antecedente e conseguente, per stabilire se un condizionale è vero o falso in base all'inseparabilità naturale occorre qualcosa di più: bisogna accertare se è *impossibile* o no che l'antecedente sia vero senza che lo sia il conseguente. Ancora qualcosa in più, rispetto al condizionale crisippeo, lo richiede la condizione di inseparabilità concettuale: perché sia vero un condizionale basato su questo tipo di inseparabi-

[3] Risulta essere in questo senso necessaria una proposizione per la quale si possa dimostrare entro un numero finito di passi l'inerenza del predicato al soggetto. L'autore di questa recensione avrebbe molto apprezzato, per quanto riguarda la seconda proposta di Leibniz, una bibliografia leggermente più ampia cui attingere.

[4] Risulterebbe pertanto falso un condizionale del tipo $\alpha \rightarrow \alpha$.

> lità, occorre che il *il senso del conseguente «sia contenuto» in quello dell'antecedente*. (Mugnai 2013, pp. 124-135, corsivi nell'originale)

Arriviamo ora al XIX secolo, con il progetto di matematizzazione della logica portato avanti, tra gli altri, da George Boole con *The Mathematical Analysis of Logic* e da Gottlob Frege con la sua *Ideografia*. Anche all'interno di questa nuova corrente, abbiamo diversi autori che si interrogano sul condizionale, come Charles Sanders Peirce, Hugh MacColl (che, nel suo *Symbolic Logic* propone un condizionale analogo a quello di Crisippo)[5] e soprattutto Clarence Irving Lewis, con il quale comincia "ufficialmente" la storia della logica modale moderna. Lewis introduce la ben nota *implicazione stretta*, un condizionale che risulta vero quando è impossibile che l'antecedente sia vero e che il conseguente sia falso. Lewis è allora in grado di costruire un sistema logico basato sull'implicazione stretta. Lewis nel 1932, insieme a Cooper H. Langford, costruirà infine alcuni calcoli modali, in particolare i sistemi $S1$[6] e $S2$[7] ancora oggi usati.

Della semantica per la logica modale tratta poi il settimo capitolo del testo, a partire dalle intuizioni di Leibniz, cui ridà vita Rudolph Carnap con la nozione di *descrizione di stato* (ovvero una classe che contiene, per ogni enunciato atomico α, α stesso o la sua negazione $\neg\alpha$):

> Così, le descrizioni di stato rappresentano i mondi possibili di Leibniz oppure i possibili stati di cose di Wittgenstein. (Carnap, *Meaning and Necessity*, cit. in Mugnai 2013, p. 145)

L'apparato semantico della logica modale si arricchisce inoltre della *relazione di accessibilità* tra mondi possibili, grazie al lavoro di logici come Tarski, Prior, Hintikka e soprattutto di Saul Kripke, oltre alla più volte citata nel testo analogia tra operatori modali e quantificatori[8]. Senza entrare nel dettaglio, basti qui ricordare, come giustamente fa Mugnai che con l'introduzione della relazione di accessibilità si stanno relativizzando i concetti modali: "necessario" diventa infatti "vero in tutti i mondi (possibili) accessibili". La semantica per la logica modale, che quantifica sui mondi possibili, pone inoltre vari problemi filosofici analizzati negli ultimi due capitoli, con la trattazione della teoria del riferimento diretto di Kripke, del realismo modale di David Lewis, del nominalismo di W.V.O. Quine e Nelson Goodman, del realismo moderato di Robert Stalnaker e

[5] Chi scrive ha qui sentito la mancanza di un maggiore spazio dedicato a questa sezione di algebra della logica, anche se questa piccola opinione rientra certamente nei gusti e negli interessi personali.

[6] http://www.cc.utah.edu/ nahaj/logic/structures/systems/s1.html.

[7] http://www.cc.utah.edu/ nahaj/logic/structures/systems/s2.html.

[8] Durante la trattazione della semantica per la logica modale Mugnai si premura di dare tutti gli strumenti affinché anche il lettore meno avvezzo alla logica possa comprendere il testo, come già abbiamo anticipato all'inizio di questa recensione.

di quello di Alvin Plantinga e infine del deflazionismo, dell'agnosticismo modale e del modalismo[9].

Nella conclusione del testo, Mugnai tratta infine delle ormai famose critiche alla modalità prodotte da Quine, facendone una breve ma efficace panoramica, la quale mette in luce (a) come Quine non sia il solo ad avversare i problemi metafisici posti dalla modalità[10] e (b) quali critiche abbiano colto nel segno e quali invece non lo abbiano fatto. Nel finale trova anche posto un riferimento all'epistemologia, ma preferiamo evitare gli *spoilers*.

Il testo di Mugnai raggiunge nell'opinione di chi scrive una notevole chiarezza, che si accompagna (ed è un pregio non da poco) ad una precisa e meticolosa scelta degli argomenti da affrontare (e conseguentemente di quelli da *non* affrontare). *Possibile/Necessario* risulta pertanto un vero piacere per il lettore, anche quello meno avvezzo ai tecnicismi logici. L'unico piccolo appunto che può essere fatto è l'assenza di un piccolo indice analitico, di cui in certi momenti si può sentire la mancanza. Questo libro ha inoltre il grande merito di valorizzare l'approccio storico alla logica e ai suoi concetti, approccio che oggi fin troppo spesso viene lasciato in secondo piano in favore di una "semplice" esposizione dei risultati.

[9] Sono teorie sulle quali, per ragioni di spazio e di chiarezza, preferiamo non soffermarci in questa sede.

[10] Una citazione di Hintikka riportata a p. 198, che lasciamo da leggere a chi deciderà di leggere questo libro, è detta da Mugnai, e chi scrive non può che essere d'accordo, addirittura "eloquente"!

Riferimenti bibliografici

Beth, Evert Willem (1953). "On Padoa's method in the theory of definition". In: *Indagationes Mathematicae* 15.1, pp. 330–339.

Boole, George (1847). *The Mathematical Analysis of Logic*.

Chang, Chen Chung e H. Jerome Keisler (1973). *Model Theory*. Studies in Logic and the Foundations of Mathematics 73. Amsterdam: North-Holland.

Frege, Gottlob (1879). *Ideografia (*t.o. *Begriffsschrift)*.

Hilbert, David (1899). *Fondamenti della geometria* (t.o. *Grundlagen der Geometrie*). A cura di Dario Narducci. Trad. dal tedesco da Pietro Canetta. Con introd. di Renato Betti. Milano: FrancoAngeli 2009.

Hodges, Wilfrid (2001). *Tarski's Truth Definitions*. A cura di Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: http://plato.stanford.edu/entries/tarski-truth/.

— (2007). *Tarski on Padoa's method*. Handout. URL: http://wilfridhodges.co.uk/history06.pdf.

— (2008). "Tarski's Theory of Definition". In: *New Essays on Tarski and Philosophy*. A cura di Douglas Patterson. Oxford: Oxford University Press, pp. 94–132.

MacColl, Hugh (1903). *Symbolic Logic*.

Mugnai, Massimo (2013). *Possibile/Necessario*. Lessico della filosofia 9. Bologna: Il Mulino.

Swijtink, Zeno (1998). "Beth's Theorem and Craig's Theorem". In: *Routledge Encyclopedia of Philosophy*. A cura di Edward Craig. Vol. 1: *A posteriori* to *Bradwardine*. Londra: Routledge, pp. 760–764.

Tarski, Alfred (1956). *Logic, Semantics, Metamathematics. Papers from 1923 to 1938*. Cur. e trad. da John Woodger. Oxford: Clarendon Press.

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

Che cos'è una Contraddizione

Francesco Berto e Lorenzo Bottai

[Carocci editore, Roma 2015]

recensione a cura di Marco Grossi

"*Che cos'è una Contraddizione*" è un libro raro. Raro perché fa una cosa difficilissima: riesce ad essere divulgativo trattando di questioni complicate. Sempre, quando ci si trova a scrivere un libro di divulgazione, ci si trova davanti ad un apparente *trade-off*: più scrivo semplice e mi faccio capire, più il contenuto sarà banale, più il contenuto è interessante e specifico, più sarò ostico e incomprensibile. Il trucco sta nel saper bilanciare le due cose: esser semplici, senza esser banali. In questo libro, i due autori Francesco Berto e Lorenzo Bottai ci son riusciti pienamente. La cosa è ancora più strana, se si pensa che stanno trattando di logica: materia notoriamente ostica, già di per sé.

Questo libro parla di contraddizioni e paradossi. Per prima cosa: cosa è esattamente una contraddizione? Il primo capitolo ne esplora il concetto: la formulazione standard di una contraddizione è $P\&\neg P$, dove P sta per un qualsiasi enunciato completo e interpretato, tipo "$2 + 2 = 4$". Ma oltre a questa definizione, ce ne sono altre. Si può parlare di contraddizione semantica: in tal caso, una contraddizione è che P sia vero e falso. O che sia vero che P e vero che $\neg P$. Oppure si può parlare di contraddizione metafisica: in tal caso, si parla di un oggetto che soddisfa proprietà contradditorie.

Perché le contraddizioni fanno così paura ai logici? Perché i paradossi sono assolutamente da evitare? La risposta la si trova nel capitolo 2, sulla "detonazione": le contraddizioni sono, normalmente, qualcosa di "esplosivo": se in un sistema una contraddizione è vera, allora all'interno di quel sistema tutto è dimostrabile. Si dice dunque che il sistema diventa "triviale", perché in esso ogni enunciato è vero. In termini più tecnici, il seguente è un teorema della logica classica: '$P\&\neg P \rightarrow Q$', dove Q è un enunciato arbitrario. Oppure, che è lo stesso, in logica classica questa è una inferenza valida: '$P\&\neg P \models Q$'. Da una contraddizione segue qualsiasi cosa. Come mai? La dimostrazione è di una semplicità

Copyright. 2018 Marco Grossi. Pubblicato in Italia. Alcuni diritti riservati. Prima pubblicazione su *Rivista Italiana di Filosofia Analitica Junior* 8:2 (2017), pp. 65-68.
Autore. Marco Grossi, 10marcogrossi@gmail.com.

disarmante: assumete $P\&\neg P$. Per eliminazione della congiunzione, ottenete rispettivamente P e $\neg P$. Per introduzione della disgiunzione, ottenete $P \vee Q$. Per il sillogismo disgiuntivo, da $P \vee Q$ e $\neg P$ inferite Q. QED

Gli autori fanno notare che, se si accetta questa inferenza, ogni sistema contradditorio è anche triviale. Lo slogan è: se accetti una contraddizione, le devi accettare tutte (ricordate che Q è un qualsiasi enunciato; quindi, può essere anche una qualsiasi contraddizione). Ecco spiegato perché i paradossi non piacciono ai filosofi: fanno esplodere il sistema. Per questo motivo, l'inferenza da una contraddizione ad un enunciato arbitrario viene anche chiamato "principio di esplosione".

Le critiche al principio di esplosione non sono poche: non a caso, $P\&\neg P \rightarrow Q$ è considerato uno dei cosiddetti "paradossi dell'implicazione materiale", data la sua stranezza semantica. Normalmente, infatti, non diremmo che se ora sono in piedi e sono seduto allora gli asini volano. Eppure, la dimostrazione sembra molto solida. Ricapitolando, essa usa tre principi: eliminazione della congiunzione, introduzione della disgiunzione e sillogismo disgiuntivo. Per evitare l'esplosione, dunque, occorrerà negare uno di questi tre principi di inferenza. È impresa disperata attaccare i primi due, e quindi di solito si attacca il sillogismo disgiuntivo. Questo principio di inferenza è però molto utilizzato nella vita di tutti i giorni, e forse addirittura dagli animali: immaginate un cane che sta fiutando la preda, arriva ad una diramazione. Vede che nel secondo sentiero la strada è bloccata, e imbocca il primo. Cosa ha appena fatto? In modo probabilmente inconscio, ha ragionato così: la preda è o nel primo o nel secondo sentiero. Ma non può esser nel secondo, perché la strada è bloccata, dunque è nel primo. Al di là della intuitività del principio in questione, c'è anche un grosso problema tecnico, che rende molto dispendioso sbarazzarsi del sillogismo disgiuntivo: il principio, infatti, è anche chiamato *modus tollendo ponens*, perché "togliendo" qualcosa ($\neg P$) pone qualcos'altro (Q). Formalmente, si scrive così: $P \vee Q$; $\neg P$, quindi Q.

Ora, in logica classica sono valide un certo tipo di trasformazioni, per cui si può trasformare una formula in un'altra. Basta rispettarne le tavole di verità. È semplicissimo: prendete il condizionale materiale "$\rightarrow$", esso ha questa tavola di verità:

A	$\rightarrow$	B
V	V	V
V	F	F
F	V	V
F	V	F

Come si può notare, il condizionale è falso se solo se l'antecedente è vero e il conseguente è falso. Intuitivamente, questo rispecchia l'idea per cui la verità

dell'antecedente di un condizionale "assicura" la verità del conseguente. Ora, però, notate che questa tavola è equivalente a quella sopra:

$(\neg$	P	$\vee$	$Q)$
F	V	V	V
F	V	F	F
V	F	V	V
V	F	V	F

Questo vuol dire che la nuova tavola dice cose uguali in modo diverso: questa dice che o l'antecedente è falso, o il conseguente è vero. Da questa tavola, però, si può dimostrare che il sillogismo disgiuntivo è equivalente ad un ben più importane principio: il *modus ponens*. Il *modus ponens* è considerato il principio base, ed è quello che dice che se è vero che se A allora B, e si dà il caso che A sia vero, allora anche B è vero. La dimostrazione dell'equivalenza è la seguente: si è mostrato prima che $P \rightarrow Q$ è equivalente a $\neg P \vee Q$. Il sillogismo disgiuntivo, ricordate, è il seguente: $P \vee Q$; $\neg P$, quindi Q. Ma $P \vee Q$ è equivalente a $\neg P \rightarrow Q$. Dunque, l'inferenza si può riscrivere così: $\neg P \rightarrow Q$; $\neg P$, quindi Q. Questo è un esempio di *Modus Ponens*.

Nonostante le insidie dell'abbandono del sillogismo disgiuntivo, si sono sviluppate recentemente delle logiche chiamate "paraconsistenti", in cui si tenta di rendere "inesplosive" le contraddizioni. In poche parole, in queste logiche passare da una contraddizione ad un enunciato arbitrario non è una inferenza valida. Nel libro se ne discutono svariati esempi: in particolare si tratta di logiche non-aggiuntive, logica del paradosso di Priest, logiche positive-plus e logiche della rilevanza.

Partiamo dalle prime, trattate nel capitolo terzo: le logiche non aggiuntive sono logiche in cui dalla verità di A e dalla verità di B non si può passare alla verità di $A\&B$. Da qui il termine "non aggiuntivo". È un po' come se le verità del mondo se ne stiano atomizzate, e non possano fondersi tra di loro. Logiche del genere sono utili per mimare le discussioni tra le persone, ed effettivamente il primo esempio di tali logiche, elaborato da Jaskowski, (1979), serviva proprio a questo. Possiamo infatti modellare una discussione come un incontro tra due mondi diversi, incompatibili tra di loro. Lo scopo del gioco è riuscire a costruire un mondo condiviso, che sia coerente. I "mondi" rappresentano le idee e opinioni di ciascun parlante. Quando si è in disaccordo su qualcosa, significa che ci sono dei "fatti" in mondi diversi incompatibili tra di loro. Per poter riuscire a discutere nonostante queste incompatibilità, i fatti di ciascun mondo non si possono "fondere" tra di loro, con la congiunzione classica: di qui la regola di non aggiunzione. Ad esempio, se nel mio mondo è vero che A e nel tuo è vero che $\neg A$, non possiamo fondere i nostri mondi, altrimenti avremmo che $A\&\neg A$, e

questo violerebbe le "regole del gioco", che proibiscono l'incoerenza. Si può già capire come mai le logiche non aggiuntive siano utili ad evitare il principio di esplosione: ora il fatto che si possa inferire P e $\neg P$ dal nostro discorso condiviso non è sufficiente per poter inferire che $P\&\neg P$, ma solo che P si dà in qualche mondo, e $\neg P$ si dà in qualche mondo. Nessuna contraddizione in questo. Certamente, da P posso inferire che $P \vee Q$, ma non posso poi usare $\neg P$ per derivare Q, almeno che $\neg P$ si dia nello stesso mondo di P. Notate che, in un certo senso, non si è evitato il principio di esplosione: se in un mondo di un parlante fosse vero che $P\&\neg P$, allora quel mondo sarebbe effettivamente triviale. Ma tal mondo è, per così dire, scartato fin dal principio dalle regole del gioco, che ci impongono di evitare le incoerenze, e che quindi squalifica automaticamente un parlante nel casi in cui il suo mondo sia incoerente.

La logica dialeteista forse più famosa, tuttavia, è sicuramente quella di Graham Priest, chiamata "logica del paradosso" (LP). Essa è trattata nel capitolo quinto. In LP, ci sono 3 valori di verità: vero; falso; vero e falso. Di solito vengono indicati rispettivamente da 1; 0; 1, 0. Una contraddizione può avere valore "vero e falso". Per poter far "spazio" al nuovo valore di verità, occorre cambiare le tavole di verità. Questa è la proposta di Priest:

A	$\neg A$
1	0
0	1
1, 0	1, 0

A	B	$A \wedge B$	$A \vee B$
1	1	1	1
1	0	0	1
0	1	0	1
0	0	0	0
1	1, 0	1, 0	1
0	1, 0	0	1, 0
1, 0	1	1, 0	1
1, 0	0	0	1, 0
1, 0	1, 0	1, 0	1, 0

La tavola della negazione è semplice: è quella classica con l'aggiunta della clausola per cui, quando un enunciato è paradossale, lo è anche la sua negazione. L'altra tavola è quella più intrigante. L'idea con cui è costruita è spiegata in modo efficace dagli autori in questo modo: immaginate di "ordinare" i valori di verità. 1 è il massimo, 0 è il minimo, e 1, 0 è "a metà". La disgiunzione tra due o più enunciati ha sempre come valore il valore minimo tra i suoi componenti, laddove la congiunzione ha quello massimo. Quindi, ad esempio, se A è 1 e B è 1, 0, $A \vee B$ ha valore 1, $A\&B$ ha valore 1, 0.

Come fa LP ad evitare il principio di esplosione? Molto semplice: prendete il sillogismo disgiuntivo, per esempio; esso fa leva sul fatto che è sempre falso

che $P\&\neg P$. Quindi, se ho $P \vee Q$ e $\neg P$, posso sicuramente star certo che non è vero che P, e dato che $P \vee Q$ ha bisogno di almeno un disgiunto per esser vero, e quel disgiunto non può essere P, posso inferire che il il disgiunto vero sia Q. Ma seguendo ora le tavole di verità di Priest, in LP le cose cambiano: può darsi che $\neg P$ sia sì vero, ma anche falso. In tal caso, anche P sarà sia vero che falso, e quindi anche vero, e ciò basterebbe a render almeno vero $P \vee Q$. In particolare, posto che P sia paradossale, mal che vada la disgiunzione sarebbe paradossale, e sicuramente non solamente falsa.

Il principio di esplosione è disinnescato: cosa comporta? Intanto non è più necessario evitare che certi paradossi siano veri. Ad esempio, prendete il paradosso del mentitore:

(1) (1) è falso.

Se (1) è vero, allora è vero che (1) è falso, quindi è falso. Se (1) è falso allora è falso che (1) è falso, quindi (1) è vero. Quindi (1) è falso se e solo se (1) è vero. Questo paradosso, fin dall'antichità, ha travagliato le menti dei filosofi. La "soluzione" di Priest è la seguente: e se non ci fosse nulla di sbagliato in (1)? Al posto di evitare in qualche modo che (1) sia contraddittorio, o che si possa costruirlo all'interno del nostro sistema, perché invece non accettare semplicemente che (1) ci sia, e sia paradossale? In LP si può, senza far esplodere il sistema. In sintesi, il punto di Priest è che, nel momento in cui si "disinnesca" il meccanismo di trivializzazione delle contraddizioni, si possono semplicemente accettare enunciati come (1). In un certo senso, la soluzione è che non c'è soluzione al mentitore![1]

Quali sono le problematiche di LP? Intanto LP non riesce a disinnescare tutti i paradossi. Ci sono i cosiddetti "*revenge liars*", dei paradossi che sono immuni a svariate soluzioni al principio di esplosione. Sono un po' come dei batteri farmaco-resistenti, che neanche il più forte antibiotico sembra esser in grado di debellare. Uno dei più insidiosi che gli autori citano è il cosiddetto paradosso di Curry:

(2) Se (2) è vero allora Q.

Se (2) è vero, allora il condizionale è vero, quindi l'antecedente è vero, e quindi ha da esserlo anche il conseguente. Se (2) fosse falso, allora il condizionale avrebbe da esser falso, ma questo è impossibile, perché l'antecedente sarebbe falso, e quindi automaticamente l'intero condizionale sarebbe vero. Quindi, in ogni caso, il condizionale è vero e quindi Q è vero. LP non riesce a disinnescare (2), perché in LP il condizionale si comporta in modo classico, e quindi (2) trivializza il sistema. Un secondo, grosso problema di LP è che in essa il *modus*

[1] Priest ha raffinato negli anni la sua teoria. Il primo esempio di LP è in Priest, (1979). Una lunga giusticazione filosofica de suo progetto si trova in Priest, (2005, 2006). Per una difesa dell'approccio paraconsistente al paradosso del mentitore, un classico è Beall, (2003).

ponens non è una inferenza valida, e il *modus ponens* è, come abbiamo detto, forse il miglior principio di inferenza che abbiamo.

Il libro di Berto e Bottai si conclude in modo aporetico, per così dire: il campo delle logiche paraconsistenti è in continua espansione, e si spera che nascano nuovi modelli che sopperiscano alle carenze degli attuali. Le logiche paraconsistenti hanno trovato utilizzo nei campi più svariati, ed hanno avuto un vero e proprio "boom" in questi ultimi anni. Gli autori non ne parlano diffusamente, ma, anche al di là della filosofia, le loro affascinanti applicazioni sono innumerevoli. Per esempio: possono esser usate per modellare il nostro sistema di credenze, dato che probabilmente abbiamo (senza accorgercene) delle credenze incompatibili tra di loro (Tanaka, (2005); Girard e Tanaka, (2016)); sono utili in intelligenza artificiale (Akama, 2016), ad esempio per il "*quantum computing*" (Agudelo e Carnielli, 2009); si tenta di applicarle in fisica quantistica, per modellare il concetto di super-imposizione di stato (Da Costa e De Ronde, (2013); De Ronde, (2015)); offrono nuovi modi di risolvere i paradossi del mentitore e altri, come quelli del sorite, riguardo al concetto di vaghezza (Priest, (2005); Beall, (2003)). Questo libro permette al lettore anche meno ferrato e senza strumenti formali di farsi una opinione chiara e informata su questo affascinante campo della logica.

Riferimenti bibliografici

Agudelo, Juan C. e Walter Carnielli (2009). "Paraconsistent machines and their relation to quantum computing". In: *Journal of Logic and Computation* 20.2, pp. 573–595.

Akama, Seiki (ed.) (2016). *Towards Paraconsistent Engineering*. A cura di Seiki Akama. Intelligent Systems Reference Library 110. Dordrecht: Springer.

Beall, J.C. (ed.) (2003). *Liars and Heaps: New Essays on Paradox*. A cura di J.C. Beall. Oxford: Oxford University Press.

Berto, Francesco e Lorenzo Bottai (2015). *Che cos'è una Contraddizione*. Roma: Carocci editore.

Da Costa, N. e Christian De Ronde (2013). "The Paraconsistent Logic of Quantum Superpositions". In: *Foundations of Physics* 43.7, pp. 845–858.

De Ronde, Christian (2015). "Epistemological and Ontological Paraconsistency in Quantum Mechanics: For and Against Bohrian Philosophy". In: *The Road to Universal Logic*. Springer International Publishing, pp. 589–604.

Girard, Patrick e Koji Tanaka (2016). "Paraconsistent Dynamics". In: *Synthese* 193.1, pp. 1–14.

Jaskowski, Stanislaw (1979). "Calcolo delle proposizioni per sistemi deduttivi contraddittori". In: *La formalizzazione della dialettica. Hegel, Marx e la logica contemporanea*. A cura di Diego Marconi. Torino: Rosenberg & Sellier, pp. 281–303.

Priest, Graham (1979). "The logic of paradox". In: *Journal of Philosophical Logic* 8.1, pp. 219–241.

— (2005). *Doubt Truth to Be a Liar*. Oxford: Oxford University Press.

— (2006). *In Contradiction: A Study of the Transconsistent*. Oxford: Oxford University Press.

Tanaka, Koji (2005). "The AGM Theory and Inconsistent Belief Change". In: *Logique et Analyse* 48.189-192, pp. 113–150.

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

La computabilità: algoritmi, logica, calcolatori

Marcello Frixione e Dario Palladino

[Carocci Editore, Roma 2011]

recensione a cura di Michele Herbstritt

Chi conosce già le *Bussole* di Carocci, sa cosa è lecito aspettarsi da *La computabilità: Algoritmi, Logica, Calcolatori*, di Marcello Frixione e Dario Palladino. Chi non le conosce, d'altra parte, avrà sicuramente di che stupirsi. Dando per scontato che ci si possa stupire tanto negativamente quanto positivamente, si può certamente dire che i modi di stupirsi per questo libro sono (almeno) due, diversi nelle cause e opposti negli effetti. Se, in quanto segue, lo "stupore" sembra un concetto troppo forte, si sostituisca a esso la "sorpresa" o l'"inaspettato": dovrebbero andar bene ugualmente.

Il modo di dire *in cauda venenum* esercita indubbiamente un certo fascino, ma in questa sede si preferisce lasciarlo da parte, cominciando a trattare (brevemente) proprio a partire dalla "brutta sorpresa", se così si può chiamare, che aspetta il lettore digiuno di Bussole e, magari, già parzialmente satollo di computabilità. I temi affrontati nelle cento pagine e poco più sono tra quelli che ci si aspetterebbe: introduzione agli algoritmi, esposizione del modello di Turing, definizione delle funzioni ricorsive, tesi di Church. E poco altro. Niente gradi di resolubilità e niente teoremi di Kleene. Chi di computabilità sa (o cerca) qualcosa di più, non lo può certo trovare qui.[1]

Fortunatamente, le "brutte sorprese" hanno anche una funzione, si passi l'esagerazione, educativa. Affrontare la lettura di una Bussola con aspettative sbagliate non può che portare a delusioni. Ma quali sono, dunque, le giuste aspettative? A questo proposito viene in aiuto la quarta di copertina: «Chiare, essenziali, accurate: le guide Carocci per orientarsi nei principali temi della cultura contemporanea», questa la definizione ufficiale di cosa siano le *Bussole*, e poi,

[1] Chi cercasse una discussione più completa e approfondita dei temi trattati può consultare (Frixione e Palladino 2004), manuale che gli autori stessi indicano come fonte principale. In lingua inglese, invece, un manuale di teoria della computabilità sicuramente ottimo ma anche più tecnico e impegnativo è (Boolos, Burgess e Jeffrey 2007).

Copyright. © 2018 Michele Herbstritt. Pubblicato in Italia. Alcuni diritti riservati. Prima pubblicazione su *Rivista Italiana di Filosofia Analitica Junior* 2:2 (2011), pp. 131-134.
Autore. Michele Herbstritt, michele.herbstritt@gmail.com.

poco più in basso: «Il testo si propone di esporre i concetti fondamentali della computabilità senza presupporre alcuna conoscenza tecnica preliminare [...]». Ebbene, tenendo a mente queste premesse, le aspettative sono giustamente ridimensionate.

Si tratta quindi di ricominciare da capo, come si è detto, con le giuste aspettative. Ed è a questo punto che subentra il secondo motivo di stupore, la "bella sorpresa", se così la si vuole chiamare. Perché se ci si accontenta della quantità e del grado di approfondimento degli argomenti trattati, non si può non rimanere colpiti dalla chiarezza con cui questi sono esposti. Tutta la trattazione è una sorta di piacevole passeggiata in cui il lettore è accompagnato per mano (talvolta addirittura portato in lettigia) verso la comprensione di temi, va detto, non sempre semplicissimi.

Il primo capitolo ha un carattere del tutto informale e introduttivo: il lettore che avesse anche solo una vaga idea del significato della parola *algoritmo* ha a sua disposizione una dozzina abbondante di pagine per farsene un'idea più precisa e rigorosa. Gli esempi, di carattere matematico e non, sono chiari e funzionali; la sezione relativa ai diagrammi di flusso molto utile in particolare per assimilare il concetto di ciclo (finito e non); l'accenno ai metodi di codifica dei dati, facendo riferimento al mondo (più quotidiano) dei calcolatori elettronici, risulta efficace. Forse, gli autori avrebbero potuto mettere l'accento un po' di più sul fatto che i concetti di cui si tratta rimangono esplicitati solo a livello intuitivo e, *quindi*, che sarà necessario renderli rigorosi in seguito.

Come è naturale, il concetto (informale) di algoritmo viene collegato a quello di funzione (calcolabile), nel secondo capitolo. Qui una qualche abitudine al ragionamento e alla notazione matematici risulta senz'altro di una certa utilità, ma la trattazione è comunque portata avanti per gradi, a partire dalla definizione generale di funzione, raffinata poi in quella di funzione calcolabile (totale o parziale), per giungere con un'accelerata finale a mostrare addirittura che l'insieme delle funzioni calcolabili non esaurirà mai l'insieme delle funzioni aritmetiche (quest'ultima dimostrazione, fondata sul ragionamento diagonale *a là* Cantor, è effettivamente più impegnativa, perché presuppone un certo grado di familiarità con alcune nozioni più avanzate di teoria degli insieme, come quella di cardinalità).

Una volta introdotti i concetti fondamentali della disciplina, non resta che "svelare l'assassino", rivelando come sia possibile rendere rigorosi i concetti di algoritmo e di funzione calcolabile. Il terzo capitolo ha precisamente questo scopo, essendo dedicato alle macchine di Turing. Dopo una concisa introduzione storica, che dà un'idea del contesto scientifico in cui inserire il famoso articolo di Turing (1936), gli autori forniscono un'esposizione piuttosto tradizionale di che cosa siano le macchine di Turing e in cosa consista il modello computazionale che si basa su di esse (le funzioni T-computabili). Anche in questo caso, gli

esempi non mancano e la trattazione risulta molto chiara. Il cosiddetto *problema della fermata*, in cui certamente risiede una parte del merito scientifico di Turing, viene rimandata al quinto capitolo. Prima di allora, è necessario presentare almeno un "complice", un altro esempio di sostituito formale: le funzioni ricorsive.

Il quarto capitolo si occupa di funzioni ricorsive, e risulta essere il più impegnativo del libro. La classe delle funzioni ricorsive è definita a partire da quella delle funzioni ricorsive primitive, le quali sono sì funzioni calcolabili, ma non certo sufficienti a esaurire la classe di queste ultime (e due controesempi sono forniti a riguardo: la cosiddetta *funzione diagonale* e la funzione di Ackermann. Per ovviare ai problemi delle funzioni ricorsive primitive, si introduce il famoso operatore di minimalizzazione, ottenendo la classe delle funzioni ricorsive generali, vere candidate allo scopo di sostituire il concetto intuitivo di funzione calcolabile.

Come si è già detto, il quinto capitolo si occupa del problema della fermata. Ma non solo, ovviamente. Anzi, prima di trattare di problemi non risolubili per mezzo di un algoritmo, risulta necessario citare la tesi di Church: i modelli rigorosi elaborati per sostituire il concetto intuitivo di algoritmo, così ci si può esprimere, colgono nel segno; in altri termini: le funzioni calcolabili sono tutte e sole quelle T-computabili/ricorsive generali. A favore della tesi di Church gli autori forniscono tre argomenti: quello di *evidenza euristica* (così la chiamano), relativo sostanzialmente al fatto che tutte le funzioni calcolabili a noi note sono effettivamente T-computabili/ricorsive generali; quello di carattere più logico, relativo al fatto che *tutte* le sistemazioni formali elaborate con lo scopo di sostituire il concetto intuitivo di algoritmo risultano equivalenti fra loro (si citano, fra gli altri, la lambda-ricorsività, Herbrand-Gödel ricorsività e gli algoritmi di Markov); quello che fa perno sulla natura indipendente e sostanzialmente non matematica del modello di Turing. Una volta stabilita l'alta plausibilità della tesi di Church, non resta agli autori che introdurre la macchina di Turing universale e il noto procedimento diagonale che porta alla definizione di un problema insolubile per mezzo di macchine di Turing (il problema della fermata), e dunque, modulo tesi di Church, insolubile in maniera algoritmica.

Giunto in fondo al quinto capitolo, il lettore (precedentemente) digiuno avrà sicuramente un bel po' di materiale su cui riflettere e, in fin dei conti, potrebbe decidere di fermarsi, almeno temporaneamente, avendo appreso i basilari di teoria della computabilità. Per chi decidesse poi di continuare la lettura, il libro garantisce altri due capitoli, di carattere più generale, dedicati a mostrare alcuni dei collegamenti vigenti fra la teoria in questione e altri interessanti temi.

Il sesto capitolo, in una dozzina di pagine, vuole essere un introduzione al problema dei Fondamenti della Matematica, a partire dalle geometrie non euclidee fino al teorema di incompletezza di Gödel (di cui è addirittura presente

una dimostrazione). L'*excursus* storico è interessante, ma risulta un po' frettoloso e, forse, sarebbe stato più utile all'inizio del libro. Il teorema di incompletezza di Gödel ha senza dubbio un suo valore intrinseco che ne rende interessante l'esposizione a prescindere, ma la decisione di fornirne una dimostrazione (per quanto informale) in un libro introduttivo come questo forse non è stata delle migliori.

L'ultimo capitolo del libro, decisamente più azzeccato, è dedicato innanzitutto a mostrare come la teoria della computabilità sia strettamente in relazione con l'informatica, fino a potersi caratterizzare come fondamento teorico di quest'ultima. A questo scopo vengono introdotti i concetti informatici fondamentali (unità di *input* e di *output*, *CPU*, programma memorizzato, linguaggio di programmazione) e viene mostrato in che modo l'architettura di von Neumannn (sulla quale si basano i calcolatori contemporanei) possa essere messa in relazione con la macchina di Turing universale, di cui si accennava nel quinto capitolo. Infine, non manca un accenno ad alcune questioni di complessità computazionale.

La seconda parte del capitolo si occupa di mostrare in che rapporto stia la teoria della computabilità con la scienza cognitiva. Quest'ultima viene introdotta da un punto di vista storico e concettuale come una reazione al comportamentismo fondata sui concetti e sugli strumenti della teoria della computabilità: il funzionalismo che caratterizzava la prima impostazione cognitivista considerava possibile una descrizione dei processi mentali in termini di computazioni, facendo astrazione dal supporto fisico che realizza i processi mentali e concentrando l'attenzione sulle proprietà logiche, i rapporti reciproci e funzionali degli stati mentali. L'impostazione funzionalista ha lasciato il posto a nuovi sviluppi, più attenti alla dimensione neurobiologica della mente (gli studi sul cervello) e a quella ecologica (*embodied mind*, rapporto fra mente e ambiente), ma gli autori sostengono che i recenti sviluppi non siano necessariamente in contraddizione con un'impostazione, in fondo, computazionale, citando a sostegno alcune interessanti riflessioni metodologico dello scienziato cognitivo David Marr.

Riferimenti bibliografici

Boolos, George S., John P. Burgess e Richard C. Jeffrey (2007). *Computability and Logic*. Fifth Edition. Cambridge University Press.

Frixione, Marcello e Dario Palladino (2004). *Funzioni, Macchine, Algoritmi*. Carocci.

Turing, Alan M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society*.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Kripke's Modal Logic: A Historical Study

Melissa Antonelli

Abstract. In a very short time Saul Kripke provided a suitable and rigorous semantics for different axiomatic modal systems and established a series of related results. Many key ideas were already in the air in the late Fifties, but it was Kripkean articles' merit to systematically introduce comprehensive devices and solutions. Later on, the spreading of possible-worlds semantics massively changed the approach to modal logic, which enormously increased in popularity after that. Since Kripke's work in modal logic is central to the development of the discipline, the aim of this essay is to present the fundamental results published between 1959 and 1965. Indeed, it was in such a brief and early phase of his career that Kripke was able to conceive the main novelties that would become central to the subsequent academic debates about modality. Here, their presentation will follow the original historical progressive introduction. Particular attention will be given to the interconnection between articles, their similarities in structure and the unified analysis produced by means of them. It actually appears quite impressive that, already in 1959, Kripke seemed to have planned all the developments he would present, one after the other, in the following years. First, an overview of the background where Kripke's ideas start to rise is given. Then, each text's results are individually briefly analysed.

Keywords. Kripke, Modal Logic, Possible-Worlds Semantics, Completeness Proofs.

I wish to thank Sara Negri, Eugenio Orlandelli and the anonymous referees for their valuable comments and suggestions.

Copyright. 2018 Melissa Antonelli. Published in Italy. Some rights reserved.
Author. Melissa Antonelli, melissa.antonelli@studio.unibo.it.
Received. 11 August 2018. **Accepted**. 19 December 2018.

Introduction

The introduction of possible-worlds semantics completely revolutionised the study of modal logic after their first appearances from 1958–1959 on. Indeed, these, so-called, «marvellous years for possible world semantics» (Copeland 2002, p. 131) paved the way to an increasing interest in the subject. Although at that time some semantics for quantified modal logics and correspondent completeness proofs were being presented by other logicians,[1] it was Kripke who offered a unified tool to analyse different modal systems and to systematically obtain connected results. The enormous popularity of his publications stimulated remarkable steps forwards in many related fields, both under a formal and a philosophical viewpoint.[2]

Given the well-recognized role of Kripke's articles as a turning point in the development of modern modal logic, the aim of the essay is to provide an organic historical reconstruction of the main results published between 1959 and 1965, by particularly focusing on the original sources and by emphasizing the elements of novelty.[3] An overall coherence and connection among his publica-

[1] Particularly remarkable are 1959 Bayart's Henkin-style completeness proof for quantified **S5** and, even more, Hintikka's possible-worlds semantics and completeness proofs for quantified **T** (Kripke's **M**), **S4** and **S5**, presented in a series of seminars held in the Boston area in 1958–1959 (unluckily, the notes of the talks are currently lost). Actually, «it is not clear which of the two [Hintikka and Kripke] was in fact the first to produce a fully worked out completeness proof (it must have been a matter of a few months at most)» (Copeland 2002, p. 130).

[2] Considering, for example, the analysis of counterfactual conditionals, it seems that Prior (1968) and Lewis (1973) proposals have hardly been conceived without the reception of Kripke's semantics (Stalnaker explicitly refers to Kripke's work (Prior 1968, 103, fn. 6), while Lewis affirms to be inspired by the success of possible-worlds semantics, thanks to which «[i]n the last dozen years or so, our understanding of modality has been much improved» (Lewis 1973, p. 418)). Indeed, valuable observations about counterfactual statements have already been advanced by other authors (as Goodman) but it was only by moving the analysis to possible-worlds context that many problems could have been solved. On the other hand, the connection between some foundational problems in philosophy of language or in metaphysics and modal logic is evident.

[3] The title of my essay is too ambitious in two ways. Indeed, in presenting (now standard) Kripke's theories, not all the facets of his modal logic are actually taken into account, due to the vastness of the topic. First of all, my analysis is confined to Kripke's early contributions about formal aspects and, consequently, it is focused on a restricted period of time. Of course, Kripke's philosophy is much wider and other fields are inseparably related (see at least (Kripke 1972) and reflections concerning "paradoxes of identity statements" and names of non-existing entities (Kripke 1963b) and (Kripke 1971)). Although Seventies studies about metaphysics and philosophy of language seem to have born and grown in strict relation to the formal ones, because of their complexity, I will avoid to treat them, confining myself to technical logical aspects. This cut can be partially justified by Kripke's choice – between 1959 and 1965 – to omit, as much as possible, explicit philosophical investigations, see (Kripke 1959a, p. 2) (however, some interests in philosophical questions emerges in (Barcan Marcus 1963b) and (Kripke 1963b)). Secondly, because philosophical issues cannot be possibly split from the «labyrinth full of twist and problems» (Fitting 1999, p. 105) of quantified modal logic, which is the central topic of (Kripke 1963b), I have decided to mainly consider propositional results and to simply outline the content of (*ibid.*) (while treating more in detail (Kripke 1959a)). It must be stressed that some other - extremely relevant but very vast - topics have been completely

tions has interestingly emerged. While many authors have introduced important innovations and at the end of the Fifties other logicians are getting closer to conceiving a relational semantics,[4] what is striking in Kripke's work is the unitary and comprehensiveness of the analysis that, starting from a specific system (quantified **S5**), is extended (with the necessary adjustments) to many other logics. Already in (Kripke 1959b), Kripke consciously planned to deal with all the topics he would debate in the following six years. In a certain way, his articles are rather chapters or parts of a unique patchwork – aimed at systematically clarifying modal logic – and not individually-conceived products. Indeed, although almost each paper can be read independently from the others, the author himself often emphasizes the link with previous and future works.

1 Overview

Formal modal logic enormously expanded during the Twentieth century. Modern interest in it was mainly revived by MacColl's series of articles published in *Mind* from 1880 on and, even more, by C.I. Lewis' publications as starting from 1918. Both authors express their dissatisfaction towards the common notion of material implication.[5] While MacColl does not propose formal definitions or axiomatizations, Lewis begins to introduce various systems of strict implication. In particular, in Appendix II of (Lewis and Langford 1932), he presents the five axiomatic systems **S1-S5**, which rapidly become canonical in the subsequent studies in modal logic.[6] Lewis' original axiomatization does not separate, re-

omitted; in particular, the relationship between Kripke's and category-theoretic semantics has not been treated at all.

[4] In 1946, Carnap proposed a possible-world semantics for quantified **S5**, based on the idea of state-description. In 1947 McKinsey and Tarski gave algebraic characterization for **S4** and **S5**, while the work of Jonsson and Tarski is described by Kripke as the (unaware, see (Copeland 2002, p. 105)) «most surprising anticipation» (Kripke 1963a) of his own (the so-called, "algebraic tradition"). The importance of Kanger's contribute is debated, see (Copeland 2002, pp. 122-123). Hintikka also stated, without proving, soundness and completeness for **T**, **B**, **S4** and **S5** (Kripke quotes Hintikka's research in (Kripke 1959b, p. 324); (Kripke 1963a, 69, fn. 2) and (Kripke 1963b, 83, fn. 1); Hintikka maybe gave some completeness results during Boston seminars). In 1955 Smiley established completeness for **M**. In 1959 Bayart published in French, a Henkin-style completeness proof for **S5**.

[5] In *Symbolical Reasoning* (1897), MacColl claims that $P \supset Q$ and $\neg P \vee Q$ are not equivalent and he distinguishes between extensional and intensional readings of the connectives. However, his texts were not particularly popular (probably because of Russell's wrong interpretation and consequent critique). Anyway, he influences C.I. Lewis, who, in 1912, criticizes Russell and Whitehead's notion of material implication. The so-called "paradoxes" should show its inadequacy to represent the actual and ordinary meaning of implication. Thus, in a series of subsequent articles, Lewis proposes different axiomatic systems of strict implication.

[6] More precisely, Lewis alone wrote Appendix II (but he acknowledges his debts to other authors in (Lewis and Langford 1932, 492, fn. 1)). The systems are numbered in order of strength and weaker systems are contained in stronger ones: **S1** consists of axioms B1-B7, **S2** adds B8 to **S1** (both have been already presented in Ch. 6), **S3** corresponds to A1-A8 (also (Lewis 1918) system), **S4** contains B1-B7 and C10, **S5** contains B1-B7 and C11 (*ivi*, pp. 500-501).

spectively, propositional and modal axioms and rules. This improved presentation – then standard – was first introduced for **S4** by Gödel in the short note in 1933, (Gödel 1933).[7] Differently from (Lewis and Langford, 1932), Gödel opts for necessity (B or N), rather than possibility, as primitive operator.[8] Later on, different presentations and formal systems widespread, such as Fays **T** (1937), equivalent to von Wright **M** (1951), or Lemmon's normal and non-normal systems. It is unlikely that without this "syntactic tradition" Kripke's works would have been conceived in the way they were. Furthermore, such a variety of systems lead a search for more rigorous interpretations of modal notions.

2 A Completeness Theorem in Modal Logic (1959)

According to (Copeland 2002, p. 129), Kripke first becomes interested in modal logics in 1956, after reading (Prior 1956). However, Kripke's relational semantics was not introduced all at once. The main goal of (Kripke 1959a), submitted in 1958 and published in 1959, is to present semantics and completeness theorem for first-order **S5** with equality. Most important novelties introduced are (1) the notion of model as based on a domain **D** and constituted by a set **K** (conceivable worlds) and by the actual world **G** and (2) the definition of the necessary proposition as true in all possible worlds. The notion of validity is defined as disconnected from the one of necessity. Although this apparatus is the basis for subsequent semantics for systems weaker than **S5**, many elements are still missing. Indeed, in 1959 neither accessibility relation nor separate valuation function appears. Moreover, in (Kripke 1963b) some elements are modified.[9] **S5*=** completeness proof is given by Beth's semantical tableaux method. This fruitful application of (Beth 1955) technique will be repeated (with some modifications)

[7] In *Eine Interpretation des intuitionistischen Aussagenkalkülus* (presented in 1932 at the Vienna Mathematical Colloquium and published in 1933), Gödel adds a provability operator B ("beweisbar") to a propositional language in order to obtain an interpretation of Heyting's intuitionistic calculus as a logic of provability. He also observes that the given system *G* (propositional system plus axioms T, K, and 4 and rule of necessitation) comes out to be equivalent to Lewis **S4**. The importance of (Gödel 1933) for future (Kripkean and not) developments in modal logic is dual: it introduces the fruitful practice of axiomatizing systems by separating propositional and modal parts and it connects intuitionistic and modal logics. Actually, the idea of introducing a provability predicate was probably suggested him by von Neumann in 1932 (this result was presented in Jan von Plato's "Gödel detective" course, University of Helsinki, 2017).

[8] Goldblatt (Goldblatt 2006, p. 6) notices that, before introducing the diamond operator in (Lewis and Langford 1932), in (Lewis 1918, p. 292) Lewis employs the impossibility operator (~) to define strict implication. The box is devised by Fitch and it first appears in a 1946 paper of Barcan.

[9] In (Kripke 1959a) the model is defined in a domain **D**, while in (Kripke 1963b) the assignment function ψ can assign different domains to different worlds of the same model. For Ballarin (Ballarin 2005), this change is related to a technical problem of 1959 semantics. Indeed, although Prior seemed to have proved the Barcan formula and its converse for quantified **S5**, Kripke suggests that they are not actually derivable (Kripke 1963b), but in (Kripke 1959a, p. 9) he has employed Prior's alleged results (see also (*ivi*, p. 10)).

in the majority of the following articles. Despite the enormous success of this result, not all the passages of the proof are explicit and it sometimes lacks of rigour.

As already said, the propositional system was first introduced in (Lewis and Langford 1932, p. 501). In 1959, Kripke adopts a quantified axiomatization for first order predicate calculus with equality, taken from Rosser, supplemented with modal axioms and rules of inference obtained from (Prior 1956):[10]

A1: $\Box A \supset A$

A2: $\neg\Box A \supset \Box\neg\Box A$

A3: $\Box(A \supset B) \supset (\Box A \supset \Box B)$

R1: If $\vdash A$ and $\vdash A \supset B$, then $\vdash B$

R2: If $\vdash A$, then $\vdash \Box A$

Given a non-empty domain **D** and a formula A, a *complete assignments* for A is defined as a function which assigns an element of **D** to every free individual variable of A, either truth (T) or falsity (F) to every propositional variable of A and a set of ordered n-tuples of members of **D** to every n-adic predicate variable. A *model* of A in **D** is an ordered pair (**G**, **K**), where **K** is a set of complete assignments for A in **D**, **G** $\in$ **K**, and every member of **K** agrees with **G** on the assignment of free individual variables of A. The intuitive meaning of this definition is presented only later, in (Kripke 1959a, pp. 2-3). Being **H** a member of **K**, the evaluation of **H** for a compound formula B is inductively defined in the usual way, apart from the modal case: «$\Box$B is assigned T if every member of **K** assigns T to B» (*ibid.* p. 2).

[10] Prior takes as the starting point Lewis **S5** formulated by means of Gödel's separation between propositional and modal parts: «It has been shown by Gödel that a system equivalent to **S5** may be obtained if we add to any complete basis for the classical propositional calculus a pair of symbols for 'Necessarily' and 'Possibly', which here will be 'L' and 'M', the axioms G1. CLCpqCLpLq, G2. CLpp, G3. CNLpLNLp; the rule RL: If α is a thesis, so is Lα; and the definition Df. M : M = NLN» (Prior 1956, p. 60). Actually, G1 corresponds to Kripke's (1959) A3 (now-called axiom K) written in Polish notation; G2 to A1 (T) and G3 is A2 (E). However, differently from what Prior seems to argue, this is not the axiomatic system presented in (Gödel 1933). Indeed, the Gödelian 1933 system *G* corresponds to **S4**, not to **S5** (Gödel's 1 (Bp $\supset$ p) corresponds to T, 2 (Bp $\supset$ .B(p $\supset$ q) $\supset$ Bq) to K, and 3 (Bp $\supset$ BBp) to 4). In fact, Prior quote Gödel's text only indirectly, as «cited in R. Feys, *Les systemes formalises des modalities aristotéliciennes* [...] 16.1-16.24» (Prior 1956, 60, fn. 2). In (Feys 1950), Feys presents **S5** subsequently to **S4**: «**16.1** Le système de postulats suivants (Gödel) est un système postulats pour S4 [...] **16.14** :: C Lp LLp. **16.2** Le système obtenu en replaçant le postulat 16.14 par 16.24 ci-dessous est un système de postulats pour S5. **16.24** :: C NLp LNLp» (*ibid*, pp. 499-500). Clearly, Feys' 16.12, 16.13, 16.24 corresponds, respectively, to Prior's G2, G1 and G3. It is likely that the direct source of Prior for **S5** is (Feys 1950) and that he did not consult (Gödel 1933), where **S5** is absent (the first English translation of it seems to be (Hintikka 1969), see (Dawson 1983)). It is possible that, in reading the secondary French source, Prior attributed both systems (introduced one after the other) to Gödel, without guessing that the substitution of **16.14** with **16.24** to obtain **S5** was not another «système de postulats suivants (Gödel)», but an axiomatization provided by Feys himself.

Furthermore, various notions of validity are introduced. A formula A is said to be *valid* in a model (**G**, **K**) of A in **D** iff A is assigned T by **G**.[11] It is *satisfiable* in **D** iff there is some model of A in **D** in which A is valid. It is *valid* in **D** iff A is valid in every model of A in **D**. A is *universally valid* iff A is valid in every non-empty domain. These definitions are required for the completeness proof.

The central concept of the intuitive interpretation is the notion of "possible world" (which is not further analysed). Because in modal logic «we wish to know not only about the real world but about other conceivable worlds» (Kripke 1959a, p. 3), Kripke introduces *models* (**G**, **K**) as composed, not by a single assignment but by a set (**K**) of assignments, one of which (**G**) represents the actual world, and the others embody all the possible ones. Moreover, because, intuitively, a proposition is necessary iff it is true in all conceivable situations, it naturally follows that $\Box$B is defined as true iff B is true in all **K**s. Coherently, a proposition is true in the actual world if it is assigned T by **G**.

After giving possible-worlds semantics for **S5***$^{=}$, Kripke presents his completeness proof. He employs an adaptation to modal logic of the method of semantical tableaux introduced by (Beth 1955). A semantical tableau is «a device for testing whether or not a given formula is semantically entailed by other given formulas» (Kripke 1959a, p. 3) and a formula B is *semantically entailed* by a set of formulas $A_1, \ldots, A_n$ iff $A_1 \,\&\, \ldots \,\&\, A_n \supset B$ is universally valid. Consequently, $A_1, \ldots, A_n$ do not entail B in case there is a model in which $A_1, \ldots, A_n$ are true and B is not. This situation is represented by a tableau with $A_1 \,\&\, \ldots \,\&\, A_n$ in the left column and B in the right one. Then, the «test of semantical entailment» (Negri 2009, p. 239) proceeds as a systematic search for a countermodel, through the construction of a system of alternative sets of tableaux: the *main tableau* and the *auxiliary tableaux*. The construction is produced by applying usual Beth's rules supplemented by two modal rules, called Yl and Yr.[12] A tableau is *closed* iff either a formula occurs in both its columns or, for some a, $a = a$ occurs in the right column. A set of tableaux is closed iff one of its tableaux is closed.

Then, Kripke establishes a series of proofs in order to obtain, only at the end, completeness for **S5***$^{=}$ (Theorem 7): $\vdash$A in **S5***$^{=}$ iff A is universally valid. This result is obtained by summing Theorem 5 (completeness: if A is universally valid, then $\vdash$A in **S5***$^{=}$) and Theorem 6 (validity: if $\vdash$A in **S5***$^{=}$, then A is universally valid). Both of them require previous proof of Theorem 1, which states that B is seman-

[11] The terminological choice is not particularly satisfying, indeed, in (Kripke 1963a, p. 69, Kripke substitutes the analogous definition for this propositional notion of "*validity* in a model", with "*truth* in a model". This second choice, which for Kripke is «clearly an improvement» (*ivi*, p. 70) and that I will adopt from now on, is the one usually preferred in literature.

[12] See, (Kripke 1959a, p. 4). Kripkean rules are substantially the same as Beth's ones but Kripke presents two rules for each connective (left and right, with rule $\wedge$r the tableau splits and there is nor rule for Ir) and considers only negation, conjunction, universal quantifier and identity, whereas Beth enumerates ten rules and more connectives are employed (Beth 1955, pp. 20-21).

tically entailed by $A_1, \ldots, A_n$ iff the construction beginning with $A_1 \& \ldots \& A_n$ in the left column and B in the right one is closed. It is proved by means of lemma 1 (if a construction beginning with $A_1 \& \ldots \& A_n$ on the left and B on the right is closed, then B is semantically entailed by $A_1, \ldots, A_n$)[13] and lemma 2 (contraposedly to the "only if" part, if the construction beginning with $A_1 \& \ldots \& A_n$ on the left and B on the right is not closed, then B is not semantically entailed by $A_1, \ldots, A_n$, so a countermodel can be found).[14] After presenting Theorems 2 and 3, i.e. the «modal analogues of the Löwenheim-Skolem Theorem» **[22, pp. 6-7]**, and some other results, Kripke defines the *characteristic formula* of a particular stage of a given tableau as $A_1 \& \ldots \& A_m \& \neg B_1 \& \ldots \neg B_n$ (where $A_1, \ldots, \& A_n$ are the formulas on the left and $B_1, \ldots, \& B_n$ the ones on the right side at the specific stage of the considered tableau). The subsequent Lemma 4 states that if A is the characteristic formula of the initial stage of a construction, and B the characteristic formula of any stage, then $\vdash A \supset B$ in **S5***$^{=}$. Finally, Kripke establishes completeness (Theorem 5) by Theorem 1 and Lemma 4: If A is universally valid (so, by Theorem 1, the tableau construction beginning with A in the right column is closed), then $\vdash A$ in **S5***$^{=}$. Subsequently he proofs validity (Theorem 6) again by using Theorem 1.

Despite the success of the article, (Bayart 1966a) highlights a structural problem in Kripke's completeness proof. Indeed, in both Theorem 1 and 7, «at each step of the construction of a system of tableaux, several possibilities generally occur so that different end results can be reached» (*ivi*, p. 277), e.g., a construction starting with bx and $\neg bx$ in the left column and $(x)bx$ in the right, can either give a closed tableau (working on $\neg bx$) or a not closed tableau (working on $(x)bx$). Bayart suggests supplementing the procedure with a rule imposing a specific choice at each step.[15]

After considering the variety of the logical systems (five distinct ones «proposed in (Lewis and Langford 1932) alone» (Kripke 1959a, p. 13)), Kripke concludes his paper announcing his intention to analyse them in a sequel. Thus, even without introducing the definitive apparatus, already in 1958[16] he plans to extend his semantical approach to obtain completeness for other modal systems. Indeed, in the same year, he publishes an abstract, (Kripke 1959b), where

[13] The proof is a *reductio ad absurdum*. Extremely roughly, if the construction is closed and we assume that B is not semantically entailed by $A_1, \ldots, A_n$ (i.e. there exists a non-empty domain **D** and a model (**G**, **K**) such that $A_1 \& \ldots \& A_n \supset B$ is not valid in (**G**, **K**)), then, because of the previously defined tableaux rules, a contradiction follows (the same formula is both true and false in the same model or $a = a$ is false). For further details, see (Negri 2009).

[14] Roughly, for each non-closed construction, it is shown how to define a suitable countermodel, say (**G**, **K**), such that – it is inductively proved – $A_1 \& \ldots \& A_n$ are assigned T and B is assigned F (so $A_1 \& \ldots \& A_n$ do not semantically entail B). See (Negri 2009). However, in (Kripke 1963a, p. 77), Kripke partially emends this proof.

[15] An emended proof close to Kripke's original one is presented in (Negri 2009, pp. 257-263).

[16] (Kripke 1959a) was received on the 25^{th} of August 1958 (after Prior's revision), but it was probably submitted in March, see (Goldblatt 2006, p. 35).

he explicitly announces all the results of (at least) the subsequent six years. In it he anticipates not only (Kripke 1963a) – where propositional systems weaker than **S5** are modal-theoretically analysed and their completeness is established – but also his following studies of non-normal systems (Kripke 1965b), quantified extensions (Kripke 1963b), and intuitionistic semantics (Kripke 1965a). Interestingly, Kripke quotes Hintikka's preceding works about **S4**, **S5** and **M** (while underling the independence of its own results). Kripke also emphasizes the strict connection between formal analyses (in particular of logics with identity) and widespread issues in philosophy of language (point 3) (e.g. the morning star paradox), which he will deepen during the Seventies.[17]

3 The Undecidability of Monadic Modal Quantification Theory (1962)

The aim of (Kripke 1962) is to show the undecidability of a monadic fragment F of MQ, subsystem of **S5***. This result is defined by Kripke as «*prima facie* surprising», considering the well-known decidability of the monadic predicate calculus for first order logic.

First of all, Kripke presents MQ.[18] Acquaintance with (Kripke 1959a) is explicitly presupposed. Then, the proof of the undecidability of F – fragment of MQ consisting of monadic formulae in two predicate letters – is sketched. Kripke shows how to reduce the decision problem for F to the one of first-order dyadic predicate logic, which is known to be undecidabile. To do so, each closed formula A of extensional dyadic predicate logic is associated to a closed formula A* of F, such that A is valid iff A* is provable in F. Concretely, given any A containing just one dyadic predicate R(x,y), the correspondent modal formula A* in F replaces R(x,y) with $\Diamond$(P(x) & Q(y)). It is then established that (1) if A is a valid formula of the dyadic theory, A* is provable in MQ and, conversely, (2) if A is not valid, A* is not provable. Given that decision problem for extensional formulae is reduced to the one of the monadic fragment of MQ, Kripke can concludes that «[s]ince the former decision problem is unsolvable, so is the latter» (*ivi*, p. 115).

After presenting some other modal results, Kripke ends the paper with the general consideration that undecidability cannot be escaped by considering mo dal systems different from **S5***: «it seems unlikely that there will be a *reasonable* modal system in which some formal analogue of this argument could not be

[17] For further details, see (Kripke 1959b, p. 324). Acknowledgment of these topics, central in (Kripke 1971) and (Kripke 1972) is witnessed by Kripke's participation at the Boston Colloquium for the Philosophy of Science 1961/1962 (Barcan Marcus 1963b, 108ss.).

[18] Kripke defines MQ as a system such that: (1) the language is the same as **S5*** (actually, in (Kripke 1962) also $\Diamond$ is used); (2) if A contains only &, $\neg$ and the universal quantifier and is valid, then it is provable in MQ; (3) MQ contains the rule of substitution; (4) MQ is a subsystem of **S5***.

carried out. In the domain of modal logic, decidable monadic systems simply do not arise» (*ivi*, p. 116) (mine italics). In this way, Kripke links formal results to intuitive considerations. Moreover, future extensions of this analysis to intuitionistic predicate calculus are announced.[19]

4 Semantical Analysis of Modal Logic I. Normal Modal Propositional Calculus (1963)

1959 semantics is inadequate for systems weaker than **S5**, so, in order to extend his proposal, two major novelties enter the stages, when, in 1963, Kripke considers propositional systems **T**, **B**, **S4** and **S5**: an accessibility relation between words and an external function ϕ, which assign value to variables relative to worlds (which are no longer complete assignments). Consequently, «the "absolute" notion of possible world in (Kripke 1959a) (where every world was possible relative to every other) gives way to relative notion, of one world being possible relative to another» (Kripke 1963a, p. 70). These innovations will be applied to many other (quantified, intuitionistic, non-normal) logics.[20] Despite the changes, Kripke considers this semantical apparatus as a *generalization* of 1959 one (see, Kripke 1963a, p. 69) and his 1963 work as aimed «to extend the [previous] results» (*ivi*, p. 67). Also the arrangement of the two articles is extremely similar. Furthermore, Kripke announces upcoming treatments of quantificational and non-normal logics (the ones enumerated in (Kripke 1959b)). On ce again, the link with previous and subsequent works is explicit.

The main goal of (Kripke 1963a) is the establishment of completeness for the considered modal propositional systems (still by Beth's method) and of the correspondence between the characteristic axiom of each system and the specific frame property. In defining the *modal propositional calculus* (MPS), in **§1**, Kripke presents the distinction between normal and non-normal systems: a *normal* calculus contains axiom schemes A1 and A3 and rules R1 and R2 are admissible.[21] Starting from the so-defined basic **M**, other normal systems are ob-

[19] Actually, in (Kripke 1965b), Kripke presents a semantics for intuitionistic logic, completeness and decision procedure for the propositional case, but undecidability of monadic quantification logic is not established. It should have been treated in a sequel (Part II) (*ibid.* p. 92) but this has never been written.

[20] It is disputed whether these innovations reflect new philosophical interpretation of modal operators or not. Ballarin interprets these novelties as a technical innovation, so that «[t]here is absolutely no sense in which it is natural to think of such model theoretic constructions (vis-à-vis the 1959 M-models) as better suited to represent a non-semantic notion of metaphysical necessity» (Ballarin 2005, pp. 284-285).

[21] In his review, Kaplan adds to them A0: A, if A is a tautology (Kaplan 1966, p. 120). In (Kripke 1963b, p. 84), Kripke defines normal systems as in 1963, but adds A0: Truth-functional tautologies. Furthermore, in (*ibid.*), he writes that «all systems considered contain all tautologies of classical propositional logic as axiom; thus these axioms will not be listed explicitly» and adds that «[t]his

tained by adding specific axioms (the so-called «reduction axioms» (*ivi*, p. 70)). In a parenthesis, Kripke defines non-normal systems as not satisfying R2. The propositional systems treated in 1963 are **M**, **B**, **S4** and **S5**. It is worth noting that Kripke's basic normal system is **M** (or **T**) of Feys-von Wright (A1-A3 plus R1 and R2)[22] and not **K**, as in recent literature about modal logics.[23] Thus, given that axiom T (A1) corresponds to the frame property of reflexivity, Kripke basic normal modal system comes out as reflexive. System **S4** is obtained by adding the axiom scheme A4 $\vdash \Box A \supset \Box\Box A$ (now 4) to **M**. **B** corresponds to **M** supplemented by the Brouwersche axiom $\vdash A \supset \Box\Diamond A$ (now B).[24] **S5** is defined as in (Kripke 1959a): **M** plus A2: $\vdash \neg\Box A \supset \Box\neg\Box A$. Kripke also notices that **S4** + B is equivalent to **S5**.

In **§2** a *normal model structure*[25] (n.m.s.) is defined as an ordered triple (**G**, **K**, **R**), where **K** is a non-empty set, $\mathbf{G} \in \mathbf{K}$, and **R** is a reflexive relation defined on **K**. As already anticipated, the presence of **R** is the novelty that permits Kripke to deal with **M**, **B** and **S4**. Moreover, Kripke explicitly points out the reflexivity of **R** for normal model structure, its transitivity for the **S4**-m.s., its symmetry for Brouwesche-m.s. and its being an equivalence relation for **S5**-m.s. Then, a **M** (**S4**, **S5**, **B**) *model* for a wff A of **M** (**S4**, **S5**, **B**) is obtained by giving a binary function ϕ(P, H) associated with the model structure, from P (atomic subformula of A) and H (subset of **K**) to the set of truth values. The appearance of this evaluation function is the second substantial novelty of 1963 semantics. The inductive definition of ϕ(P, H) is standard for propositional connectives, while for modal formulas it is: $\phi(\Box B, H) = T$ iff $\phi(B, H') = T$ for every $H' \in K$ such that H**R**H'. This definition is a coherent extension of 1959 intuitive meaning of necessity. Furthermore, consistently with (Kripke 1959a), a formula A is *true* in a model (**G**, **K**, **R**, ϕ)[26] if it is true in the actual world (i.e. $\phi(A, \mathbf{G}) = T$), otherwise it is false. A is *valid* if it is true in all the models. A is *satisfiable* if true in, at least, one of them.

Similarly to (Kripke 1959a), after the presentation of the formal notions required for completeness proof (presented in **§4**), Kripke gives an informal motivation for its semantics (**§2.1**).[27] Then, he considers the correspondence be-

proviso was inadvertently negated in (Kripke 1963a)».

[22] It was introduced in 1937 as 't' by Feys, who obtained it by dropping 3 in (Gödel 1933), see (Hughes and Cresswell 1966, 50, fn. 7). It was first called **T** in 1953 by Sobocinski, who also showed its equivalence with **M**.

[23] The, now standard, name **K** was given to it by Lemmon and Scott in 1977, in honour of Kripke, and it does not appear in any of Kripke's paper, see (*ibid.*, p. 49, fn. 1). Kripke motivates the reflexivity of his basic system as «an intuitively natural requirement» (Kripke 1963b, p. 84). However, at the end of (Kripke 1963a, p. 95), he considers the possibility to drop it (T) in order to obtain a system for deontic logic.

[24] The formula $p \supset LMp$ appears in Becker (1930). The name of axiom (B) derives from Brouwer. For further details, see (Hughes and Cresswell 1966, 70-71, fn. 5).

[25] Kripke never uses the word *frame*. It was employed for the first time in 1968 by Segerberg, apparently under suggestion of Dana Scott. About this, see also (*ibid.* p. 50, fn. 3).

[26] Kripke never uses (G, K, **R**, ϕ) for model, but he writes «in a model ϕ associated with a m.s. (G, K, **R**)» (Kripke 1963a, p. 69).

[27] In this section Kripke himself presents the elements of continuity and discontinuity with (Kripke

tween the characteristic axiom of each system and the related property of **R**. First of all, for each system, «[i]t is clear that every world H is possible relative to itself» (Kripke 1963a, p. 70), so that not only he emphasizes, once again, the reflexivity of basic normal system, but also presents the correspondence between **M** and reflexive structure. Then, he shows (not extremely rigorously and straightforwardly)[28] the ones between, respectively, transitive structure and **S4**, symmetric structure and **B**, and structure where **R** is an equivalence relation and **S5**.[29]

As for (Kripke 1959a), completeness[30] is established by systematic searching for a countermodel. In **§3** Kripke, again, introduces the extended modal variant of Beth's tableaux method. The rules for **M** in (Kripke 1963a) are very close to 1959 ones.[31] As before, the falsification procedure for $A_1 \,\&\, \ldots \,\&\, A_m \supset B_1 \vee \ldots \vee B_n$ assumes $A_1, \ldots, A_m$ to be true and $B_1, \ldots, B_n$ to be false in the model, by putting $A_1, \ldots, A_m$ to the left and $B_1, \ldots, B_n$ to the right side of the main tableau. Then rules are applied. No order of application is specified (although more rigorous, Kripke considers it as superfluous). For systems **B**, **S4** and **S5**, it is assumed that the relation **R** is (respectively) symmetric, transitive or both. However, Kripke does not treat these rules as formal part of the syntax in tableaux construction (Negri 2009, p. 243).[32] A countermodel for a formula A is then searched by applying the given construction to a main tableau where A is put on the right side. If no countermodel exists, the formula is valid. Still a tableau is *closed* iff a formula appears on both sides (there is no identity in 1963 language); a set of tableaux is closed iff some tableau in it is and a system is closed iff each of its alternative sets is. The continuity between 1959 and 1963 articles is evident. The presentation of the method is followed by some examples of **S4** and **S5**-tabelau construction procedure (**§3.1**).

In **§3.2** Kripke proves the completeness of tableau procedure with respect to

1959a). Kripke defines relational semantics as a generalization of 1959 one, from which it differs for the auxiliary function ϕ and for the appearance of R, whose informal interpretation is given: $H_1 \mathbf{R} H_2$ expresses that H_2 is possible relative to H_1 or H_2 is related to H_1. The «"absolute" notion of possible world» for **S5**, is replaced by a «relative notion» (Kripke 1963a, p. 70) (see also (Kripke 1963b, p. 84)). The definition of necessity (and possibility) of A in a world also changes.

[28] The overall strategy to prove correspondence between a frame property and an axiom is to assume the validity of the axiom in the frame, and to show that this frame has the correspondent property. For example, given the Brouwersche axiom B, $A \supset \Box\Diamond A$, and the fact that for an arbitrary H_1, $H_1 \mathbf{R} H_2$ holds, Kripke proves $H_2 \mathbf{R} H_1$, so that the structure is symmetric.

[29] As an alternative, he proposes to simply abandon **R** and to use the model structure (**G**, **K**).

[30] In (Kripke 1963a, 69, fn. 2), Kripke presents a "status questionis" concerning completeness proof, showing his awareness about previous results and underlying the independence of his discoveries.

[31] Due to the presence of the reflexive relation **R**, the rules $\wedge$r (the one which splits) and the two modal rules change, see respectively (Kripke 1959a, p. 4) and (Kripke 1963a). Of course, propositional 1963 rules do not include $\prod$l, $\prod$r, Il and Ir.

[32] Kripke himself is aware that his presentation lacks in formal clearance. Indeed, in describing $\wedge$r he writes: «I hope this explanation makes the process clear intuitively; the formal statement is rather messy» (*ibid.*, p. 73). See also (Kaplan 1966) critique.

the semantics (for the four propositional systems, a construction for A is closed iff A is valid). This corresponds to 1959 Theorem 1, where semantically entailment substitutes the notion of A-validity.[33] The completeness theorem is developed in **§4**: A is provable in **M**, **B**, **S4** or **S5** iff it is true in the corresponding model. In **§4.1**, validity is proved: every provable formula is valid in the specific system. Kripke defines this constructive proof as an «easy mechanical task», obtained by simply verifying that every axiom is valid for the appropriate theory, and that the rules preserve validity. Then, in **§4.2**, completeness for each system is established.[34] In **§5.1** Kripke proves decidability for **M** and **B** and for **S4** and **S5** in, respectively, analogous ways. In **§6**, Kripke announces future extension of the semantical analysis to non-normal logic and quantification theory.

The legacy of these results is enormous. However, even if Kripke's papers pave the way to a rigorous analysis of modal logic, the most widespread presentation of modal completeness theorems is usually not the original one, based on tableaux method. Indeed, in his 1966 review, Kaplan judges 1959 proof as not clear enough. He alternatively proposes a sketch of a Henkin-style proof, considered «more rigorous» (Kaplan 1966, p. 121) since it avoids tableau technique and does not require any reader's geometrical intuition. Kaplan attributes the idea to Dana Scott. Actually, the first Henkin-style completeness proof for **S5** appeared is 1958 Bayart's paper.[35] Later, completeness proofs for various modal systems appear in Makinson (1966) and in Cresswell (1967). Despite the success of Kaplan's suggestion and the general adoption of Henkin's technique, Negri emphasizes that it hides some information that are instead explicit in Kripke's original proof (Negri 2009). Indeed, elegant Henkin's proof is based on a "trick" and it does not show how to obtain a countermodel for underivable propositions. So, Kripke's proof is more informative. In order to overcome both the lack of data in Henkin's proof and the lack of clearance in Kripke's one, Negri proposes the introduction of a labelled sequent calculus which permits a direct and rigorous completeness proof.[36]

[33] Kripke himself considers 1963 Lemma 1 and 2 corresponds to 1959 homonymous. Lemma 1 establishes validity. As for (Kripke 1959a), the proof is a *reductio ad absurdum*. Lemma 2 demonstrates completeness, by contraposition (and employs König's *Unendlichkeitslemma* is quoted). See, (Negri 2009).

[34] Obtained in **§3.2** the completeness of tableau procedure with respect to the semantics, he has to simply show that «if the construction for A is closed, then A is provable in the appropriate system» (Kripke 1963a, p. 82). He employs Lemma: If A_0 is the characteristic formula of the initial stage of a construction, and B_0 is the characteristic formula of any stage, then $\vdash A_0 \supset B_0$ (analogous to 1959 Lemma 4).

[35] It is striking that Bayart himself, in his review to (Kripke 1959a) does not mention his alternative approach for establishing **S5*** completeness, see (Bayart 1966b).

[36] In (Negri 2009), the labelled calculus is given through internalization of possible-worlds semantics within the syntax. Modal systems stronger than **K** are obtained by adding to G3K rules corresponding to the characteristic properties of the desired systems. This approach, which makes the accessibility relation an explicit part of the syntax and not an implicit property of the tree, simplifies

5 Semantical Considerations on Modal and Intuitionistic Logic (1963)

Kripke completes his formal analysis of normal modal logics in (Kripke 1963b) by considering quantified **M**, **B**, **S4** and **S5**. Despite some changes in model theory, the continuity with previous articles into a unified project is again evident.[37] Moreover, future treatment of non-normal ones is announced. First of all, previous definitions and results for normal **M** and its extensions (definition of model structure and completeness theorem) are summed up. Then, from p. 84 on, Kripke introduces quantifiers by defining a *quantificational model structure* (**G**, **K**, **R**) and a function ψ, which assigns to each H $\in$ **K** a set ψ(H) of all individuals existing in H, called the *domain* of H. ψ(H) needs not to be the same set for different Hs: intuitively, Pegasus does not exist in the real world but may appear in some other. This rises the problem of if and which truth-value to assign to ϕ(P(x), H) when x exists in the domain of some world H' but not in the one of H (Kripkean example is Sherlock Holmes). After comparing different historical proposals of solution to this (Frege-Strawson and Russell), Kripke concludes that «[f]or the purposes of modal logic we hold that different answers to this question represent alternative *conventions*. All are tenable» (*ivi*, pp. 85-86). Kripke opts for the (bivalent) solution «that a statement containing free variables has a truth-value in each world for every assignment to its free variables» (*ibid.*). Differently from previous works, where the content is mainly formal, here the link with general issues in philosophy of language emerges. Kripke also shows that with 1963 semantics the Barcan formula and its converse are not **S5***-valid.

The article concludes with some brief remarks on the "provability" interpretations for propositional modal logics: «Provability interpretations are based on a desire to adjoin a necessity operator to a formal system, say Peano arithmetic, in such a way that, for any formula A of the system, □A will be interpreted as true iff A is provable in the system» (*ivi*, p. 90). Thus, ϕ(□P, F) = T iff P is provable in **PA**. Kripke also deals with the mapping of intuitionistic logic into **S4** in order to get a model theory for intuitionistic predicate calculus, without giving its model theory and confining the study to propositional calculus. However, some central elements of (Kripke 1965a) are anticipated: ¬A is verified in **E** iff there is no consistent extension of **E** verifying A; A ⊃ B is verified in **E** iff every consistent extension **E'** of **E** verifying A also verifies B.

Kripke's tableau method. For further details about completeness proof and modal proof theory, see at least (Negri 2005, pp. 312-319), (Negri and von Plato 2001, pp. 81-86) and (Negri 2011).

[37] Apart from explicit and various references to them, the arrangement of the paper is analogous to the former ones (excluded completeness proofs, which are mostly suppressed).

6 Semantical Analysis of Intuitionistic Logic I (1965)

(Kripke 1965b) published in 1965 (but presented in 1963) does not directly concern modal logics but it is still strictly connected with it and to previously mentioned results. Indeed, Kripke writes that «the semantics for modal logic which we announced in (Kripke 1959b) and developed in (Kripke 1963a) and (Kripke 1963b), together with the known mapping of intuitionistic logic into the modal system **S4**, inspired the present semantics for intuitionistic logic» (*ivi*, p. 92).

In **§1** intuitionistic semantics is presented. An intuitionistic model structure is an ordered triple (**G**, **K**, **R**), with **R** reflexive and transitive. Kripke adds to the usual definition a condition to be satisfied: if ϕ(P, H) = T and H**R**H', then ϕ(P, H') = T (H, H' $\in$ **K**). The truth-value of a formula in a world (H) is defined in the standard way for & and $\vee$, while for negation and implication, it is: ϕ(A $\supset$ B, H) = T iff for all H' $\supset$ **K** such that H**R**H', ϕ(A, H') = F or ϕ(B, H') = T; ϕ($\neg$A, H) = T iff for all H' $\in$ **K** such that H**R**H', ϕ(A, H') = F. Then, the usual notion of validity and quantificational model are presented. Kripke explicates that **G** has to be interpreted as the "evidential situation". Given H to be any situation, we have H**R**H' if, as far as we know at the time H, we may later get enough information to advance to H'. Thus, «[t]he requirement that, for any A, if ϕ(A, H) = T and HRH', then ϕ(A, H') = T simply means that if we already have e proof of A in the situation H, then we can accept A as proved in any later situation H'» (Kripke 1965b, p. 99). The interpretation of the connectives in intuitionistic semantics is explained as well: «To assert $\neg$A intuitionistically in the situation H, we need to know at H not only that A has not been verified at H, but that it cannot possibly be verified at any later time, no matter how much more information is gained [...] Again, to assert A $\supset$ B in a situation H, we need to know that in any later situation H' where we get a proof of A, we also get a proof of B» (*ibid.*).

After presenting tableaux method for intuitionistic logic (**§2**), validity (**§3.1**) and completeness are established by Beth's method (**§3.2**).[38] Results concerning decidability (decision procedure for propositional intuitionistic logic and undecidability of monadic quantification theory) should have appeared in a sequel, Part II, but this has never been published.

[38]The analogy with 1963 proof is evident. The proof of Theorem 2 (the completeness of tableau procedure) is just sketched because it is «a routine variation of the proofs of the corresponding theorems in (Kripke 1963a) [Lemma 1 and 2]». Validity (Theorem 3) is the same «trivial» (Kripke 1965b, p. 214) task, as the analogous «mechanical» (Kripke 1963a, p. 82) proof presented in **§4.1** two years before. Also completeness proof (Theorem 4) is defined as similar to 1963 one.

7 Semantical Analysis of Modal Logic II. Non-normal Modal Propositional Calculi (1965)

Kripke also introduced a special kind of worlds, dubbed *non-normal worlds*, in order to provide a semantics for modal logics weaker than the basic **K** (called "non-normal" in their turn), such as C.I. Lewis **S2** and **S3**. Specifically, the non-normal modal systems at issue do not include the rule of necessitation: if $\vdash$A, then $\vdash \Box$A (Kripke 1963a, p. 67). For this reason, to Kripke they appear to be «intuitively somewhat unnatural» (*ivi*, p. 206). Nevertheless, he considers useful to propose an «elegant model theory» (*ibid.*) for them, where non-normal worlds appears to be technical devices to make necessitation fail (Berto 2013).

Their treatment in (Kripke 1965a) requires acquaintance of (Kripke 1963a) is presupposed.[39] In **§2**, Kripke defines some non-normal propositional systems[40] and, in **§3**, their semantics. He starts showing that in **E2**, **E3** no formula of the form $\Box$B is provable, not even $\Box$(A $\supset$ A) (although A $\supset$ A is true in every world). Moreover, in **E2**, $\vdash \Box$B $\supset \Box$(A $\supset$ A). Consequently $\vdash \neg\Box$(A $\supset$ A) $\supset \neg\Box$B. Thus, if A $\supset$ A is not necessary, nothing is. This leads Kripke to divide possible worlds in two classes: "normal worlds", where necessity is evaluated according to 1963 semantics, and "non-normal worlds", where $\Box$B is always false. Then, (**E2**) *model structure* is defined as a quadruple (**G**, **K**, **R**, **N**), where **K** is the set of worlds, **G** $\in$ **K**, **N** $\subset$ **K**, and **R** is reflexive on **N**. **N** represents the set of normal worlds.[41] A *model* is obtained by associating a valuation function ϕ(P, H) to the frame. ϕ(A, H) is inductively defined as in 1963 for propositional connectives (&, $\neg$), but it changes for modal formulas. Indeed, a modal formula is true in a normal world if it results necessary in it in the usual sense (otherwise it is false), but it is always false in non-normal worlds. Again, a formula A is true in a model (**G**, **K**, **R**, **N**, ϕ) iff ϕ(A, G) = T and A is valid iff it is true in every model. Differently from (Kripke 1963a), the distinguished **G** plays here an essential role in the definitions.[42]

[39] Kripke writes that (Kripke 1965a) «continues the investigations of (Kripke 1963a)» and that it «extends the results of (*ibid.*) to these and other systems. The results of this paper were announced in (*ibid.*), (Kripke 1959b)» (*ibid.*, p. 206). Titles himself emphasize the continuity between these two works.

[40] Lemmon **E2** is characterized by axioms A1 (T) and A3 (K) and rules R1 and (Eb) If $\vdash$A $\supset$ B, then $\vdash \Box$A $\supset \Box$B (Lemmon presents, in 1957, the first "Gödel-style" formulation of *non-normal* **S1-S3** and introduces **E1-E5** as the "epistemic" counterparts of **S1-S5**). **E3** is obtained from **E2** by replacing A3 with stronger (1) $\Box$(A $\supset$ B) $\supset \Box(\Box$A $\supset \Box$B). **E4** consists of axioms A1, A3 and A4 (4) and rules R1 and (Eb). The re-defined **E5** is constituted by A1, A3 and A2, i.e. $\Box$A $\supset (\neg\Box$B $\supset \Box \neg\Box$B) and rules R1 and (Eb) (original **E5** collapses into **S5**). Łukasiewicz Ł-modal system is axiomatized by A1, (4), i.e. (A $\supset$ B) $\supset (\Box$A $\supset \Box$B), plus R1.

[41] **E3**-m.s. are obtained when **R** is transitive; **S2** (**S3**)-m.s. are **E2** (**E3**) where **G** is normal.

[42] Indeed, for example, $\Box$(A $\supset$ A) is valid in **S2** (where **G** is normal), because $\phi(\Box$(A $\supset$ A), **G**) = T. Then, in **§3** and **§4** Kripke presents tableaux method and semantical rules for non-normal propositional logics, in **§5** completeness theorem and in **§7** (according to (Makinson 1970), somehow inaccurately) other non-normal modal systems widespread in literature.

References

Ballarin, Roberta (2005). "Validity and Necessity". In: *Journal of Philosophical Logic* 34.3, pp. 275–303.

— (2017). *Modern Origins of Modal Logic*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/entries/logic-modal-origins/.

Barcan Marcus, Ruth (1963a). "Discussion". In: *Boston Studies in the Philosophy of Science: Proceedings of the Boston Colloquium for the Philosophy of Science 1961/1962, vol. I*. Ed. by M. Wartofsky. Dordrecht: Reidel, pp. 105–116.

— (1963b). "Modal Logic I: Modalities and Intensional Languages". In: *Boston Studies in the Philosophy of Science: Proceedings of the Boston Colloquium for the Philosophy of Science 1961/1962, vol. I*. Ed. by M. Wartofsky. Dordrecht: Reidel, pp. 77–96.

Bayart, Arnauld (1966a). "Review: Saul A. Kripke, A Completeness Theorem in Modal Logic". In: *The Journal of Symbolic Logic* 31.2, pp. 267–277.

— (1966b). "Review: Saul A. Kripke, The Undecidability of Monadic Modal Quantification Theory". In: *The Journal of Symbolic Logic* 31.2, pp. 277–278.

Berto, Francesco (2013). *Impossible Worlds*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/entries/impossible-worlds/.

Beth, Evert Willem (1955). "Semantic Entailment and Formal Derivability". In: *The Philosophy of Mathematics*. Ed. by J. Hintikka. Oxford 1969: Oxford University Press, pp. 9–41.

Borghini, Andrea et al. (2010). *Il genio compreso: La filosofia di Saul Kripke*. Roma: Carocci.

Copeland, Brian Jack (2002). "The Genesis of Possible Worlds Semantics". In: *Journal of Philosophical Logic* 31.1, pp. 99–137.

Dawson, John William (1983). "The Published Work of Kurt Gödel: An Annotated Bibliography". In: *Notre Dame Journal of Formal Logic* 24.2, pp. 255–284.

Feys, Robert (1950). "Les systemes formalises des modalities aristotéliciennes". In: *Revue philosophique de Louvain* 20.1, pp. 478–509.

Fitting, Malvin (1999). "On Quantified Modal Logic". In: *Fundamenta Informaticae* 39.1, pp. 105–121.

Gabbay, Dov (1969). "Review: Saul A. Kripke, Semantical Considerations for Mo dal Logics". In: *The Journal of Symbolic Logic* 34.3, pp. 501–519.

Ghilardi, Silvio (1991). "Incompleteness results in Kripke semantics". In: *The Journal of Symbolic Logic* 56.2, pp. 517–538.

Goldblatt, Robert (2006). "Mathematical Modal Logic: A View of its Evolution". In: *Handbook of the History of Logic, vol. 7*. Ed. by M. Wartofsky. New York: Elsavier, pp. 1–97.

Gödel, Kurt (1933). "An Interpretation of Intuitionistic Sentential Logic". In: *The Philosophy of Mathematics*. Ed. by J. Hintikka. Oxford 1969: Oxford University Press, pp. 128–129.

— (2003). *Collected Works, Volume V Correspondence H-Z*. Oxford: Clarendon.

Hintikka, Jaakko (1969). *The Philosophy of Mathematics*. Oxford: Oxford University Press.

Hughes, George Edward and Maxwell John Cresswell (1966). *A New Introduction to Modal Logic*. London-New York: Routledge.

Kaplan, David (1966). "Review: Saul A. Kripke, Semantical Analysis of Modal Logic I. Normal Modal Propositional Calculi". In: *The Journal of Symbolic Logic* 31.1, pp. 120–122.

Kontchakov, Roman (2005). "Undecidability of first-order intuitionistic and modal logics with two variables". In: *Bullettin of Symbolic Logic* 11.3, pp. 428–438.

Kripke, Saul (1959a). "A Completeness Theorem in Modal Logic". In: *The Journal of Symbolic Logic* 24.1, pp. 1–14.

— (1959b). "Semantical Analysis of Modal Logic (abstract)". In: *The Journal of Symbolic Logic* 24.1, pp. 323–324.

— (1962). "The Undecidability of Monadic Modal Quantification Theory". In: *Zeitschritft für matheatische Logik und Grundlagen der Mathematik* 8.2, pp. 113–116.

— (1963a). "Semantical Analysis of Modal Logic I. Normal Modal Propositional Calculi". In: *Zeitschritft für matheatische Logik und Grundlagen der Mathematik* 9.5, pp. 67–96.

— (1963b). "Semantical Considerations on Modal Logic". In: *Acta Philosophica Fennica* 16.1, pp. 83–94.

— (1965a). "Semantical Analysis of Intuitionistic Logic I". In: *Formal Systems and Recursive Functions: Proceedings of the Eight Logic Colloquium (Oxford 1963)*. Ed. by J.N. Crossley and M. Dummett. Amsterdam: North Holland, pp. 93–130.

— (1965b). "Semantical Analysis of Modal Logic II. Non-Normal Modal Propositional Calculi". In: *Theory of Models*. Ed. by J. Addinson, L. Henkin, and A. Tarski. Amsterdam: North Holland, pp. 206–220.

Kripke, Saul (1971). "Identity and Necessity". In: *Identity and Individuation*. Ed. by M.K. Munitz. New York: New York University Press, pp. 135–164.

— (1972). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Lewis, C.I. and C.H. Langford (1932). *Symbolic Logic*. New York: Century Co.

Lewis, Clarence Irving (1918). *A Survey of Symbolic Logic*. Berkeley: University of California Press.

Lewis, David (1973). "Counterfactuals and Comparative Possibility". In: *Journal of Philosophical Logic* 2.4, pp. 418–446.

Lindström, Sten and Krister Segerberg (2007). "Modal Logic and Philosophy". In: *Handbook of Modal Logic*. Ed. by P. Blackburn. New York: Elsevier, pp. 1149–1214.

Makinson, David (1970). "Review: Saul A. Kripke, Semantical Analysis of Modal Logic II. Non-Normal Modal Propositional Calculi". In: *The Journal of Symbolic Logic* 35.1, pp. 135–136.

Mugnai, Massimo (2013). *Possibile/necessario*. Bologna: Il Mulino.

Negri, Sara (2005). "Proof analysis in modal logic". In: *Journal of Philosophical Logic* 34.1, pp. 507–544.

— (2009). "Kripke completeness revisited". In: *Acts of Knowledge – History, Philosophy and Logic*. Ed. by G. Primiero and S. Rahman. College Publications, pp. 233–266.

— (2011). "Proof theory for modal logic". In: *Philosophy Compass* 6.8, pp. 523–538.

Negri, Sara and Jan von Plato (2001). *Structural Proof Theory*. Cambridge: Cambridge University Press.

— (2011). *Proof Analysis: A Contribution to Hilbert's Last Problem*. Cambridge: Cambridge University Press.

Prior, Arthur Norman (1956). "Modality and Quantification in S5". In: *The Journal of Symbolic Logic* 21.1, pp. 60–62.

— (1968). "A Theory of Conditionals". In: *American Philosophical Quarterly* 2.1, pp. 98–112.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

A Formal Analysis of the Best System Account of Lawhood

Giovanni Cinà[1]

Abstract. In this work I attempt a reformulation of Lewis' Best System Account, explicitating the underlying formal conception of scientific theories and trying to define the concepts of simplicity, strength and balance. This essay is divided in three sections. In the first one I introduce the Best System Account of natural laws and formulate the need for its improvement. In the second section I outline a formal framework where the notions of deductive system and scientific theory can be defined precisely. In the last section the notions of simplicity, strength and balance are analyzed. To conclude I argue that the framework proposed does indeed provide the precision required. In addition, it also offers interesting insights on the plurality of concepts of simplicity, strength and balance, and on the general enterprise of formalizing scientific theories.

Keywords. Best System Account, Theory Choice, Formalization of Scientific Theories..

[1]I thank the anonymous reviewers for their helpful observations and suggestions.

Copyright. © 2018 Giovanni Cinà. Published in Italy. Some rights reserved.
First published on *Rivista Italiana di Filosofia Analitica Junior* 5:1 (2014), pp. 59-63.
Author. Giovanni Cinà, giovanni.cina88@gmail.com.

1 The Best System Account

The Best System Account, BSA hereafter, is an attempt to answer the philosophical question: "What are natural laws?". The three philosophers associated with this perspective on natural laws are J.S. Mill, F.P. Ramsey and D. Lewis, and for this reason BSA is also known as MRL account. Let us introduce BSA quoting the *locus classicus* of the latter author. In his 1973 book *Counterfactuals*, Lewis characterized BSA in the following terms:

> Whatever we may or may not ever come to know, there exist (as abstract objects) innumerable true deductive systems: deductively closed, axiomatizable sets of true sentences. Of these true deductive systems, some can be axiomatized more *simply* than others. Also, some of them have more *strength*, or *informational content*, than others. The virtues of simplicity and strength tend to conflict. Simplicity without strength can be had from pure logic, strength without simplicity from (the deductive closure of) an almanac. [...] What we value in a deductive system is a properly balanced combination of simplicity and strength - as much of both as truth and our way of balancing will permit. We can restate Ramsey's 1928 theory of lawhood as follows: a contingent generalization is a law of nature if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength. (Lewis 1973, p. 73, original italic)

We can immediately observe that Lewis reduces the problem of characterizing natural laws to the problem of theory choice: once we have selected the best system(s) we can determine if a statement is a natural law by checking if it is a theorem or an axiom of said system(s). It is worthwhile to remark that this procedure will fail if the systems we are considering are undecidable.

Lewis' conception itself was not monolithic. It was articulated and slightly modified during time in order to make it fit in Lewis' own philosophy, e.g. with Principal Principle, modal realism and natural properties.[1] In what follows, however, I won't analyze the development of Lewis ideas through time. My aim is to discuss, and possibly clarify, the four core notions of BSA, namely the notions of deductive system, simplicity, strength and balance. As can be seen from the last quotation, for Lewis simplicity and strength are binary relations such that:

- a system is simpler than another one if it has a simpler axiomatization;
- a system is stronger than another one if it has more informational content.

From other textual evidences it seems that for Lewis, given a deductive system, the addition of an assumption increases the strength and decreases the simplic-

[1] See (Lewis 1973, 1986, 1994, 1999).

ity of the deductive system. I therefore take the *number* of axioms (or hypotheses, as I will prefer to call them later) to be the Lewisian measure of the symplicity of a deductive system.[2]

This characterization is insufficient, as I will argue in what follows. Indeed, the necessity to pin down these concepts more precisely can be traced back to Lewis himself, as witnessed by the following quotations:

> In science we have standards - *vague* ones, to be sure - for assessing the combination of strength and simplicity offered by deductive systems. (Lewis 1973, pp. 73–74, emphasis mine)

and

> Of course, it remains an unsolved and difficult problem to say what simplicity of a formulation is. (See the 1983 article "New work for a theory of universals", reprinted in Lewis 1999, p. 42)[3]

In order to pursue the analysis of these notions I will stick to the 1973 formulation of BSA. This is, to the best of my knowledge, faithful enough to the version of BSA that was received in the literature on natural laws.[4]

1.1 The contemporary debate and the need for a more precise version of BSA

The contemporary literature on BSA addresses a wide range of issues, essentially accepting the 1973 formulation and its core notions. In general, we can identify roughly two attitudes towards the explicit definitions of simplicity, strength and balance. On one hand, the issue is ignored, in the sense that scholars rest content with Lewis' characterization or simply decide to postpone its analysis (among the others, the articles (Cohen and Callender 2009), (Jaeger 2002) and (Robert 1999) are, in different degrees, examples of this perspective). On the other hand, it is perceived as problematic (see for example (Psillos 2002, p. 152); (Bird 1998, p. 40); (Armstrong 1983, p. 67); (Mumford 2004, p. 44)). The clearest exposition of this second stance is Van Fraassen's:

> I have written here as if simplicity, strength and balance are as straightforward as a person's weight or height. Of course they are not, and the literature contains no account of them which it would be fruitful to discuss here. [...] To utilize these motions uncritically, as if they dealt with such well-understood triads as 'under five foot five, over

[2] The correctness of this interpretation is however not essential for the aim of this paper, namely providing an apt framework to specify the notions of simplicity, strength and balance.

[3] Where 'formulation' refers to the formulation of a deductive system.

[4] See for examples, among the recent papers, (Bird 2008, p. 74) and (Cohen and Callender 2009, p. 4).

> 200 pounds, overweight' may be unwarranted. (Van Fraassen 1989, pp. 41–42)

I agree with this concern and I take the insufficient precision of such notions as a drawback of BSA. The following section will be devoted to the (re)construction of a suitable frame for such tasks.

2 The Formal Framework

To attempt a clearer formulation of simplicity, strength and balance we have to use a toolkit of more precise, and possibly shared, definitions. According to Lewis, these notions are to be applied to scientific theories conceived as deductive systems. But what is a deductive system exactly? In his words a deductive system is a "deductively closed, axiomatizable sets of true sentences" (Lewis 1973, p. 73). However, a deductive system is usually understood as a purely syntactic object.[5] What is then the role of truth in a formal representation of scientific theories and what do we mean by deductive system? Given that BSA is essentially a formal account of lawhood, the notions of axiomatization, derivation and deductive system are crucial. But Lewis is not explicit in explaining how they enter the picture. I maintain that we need a more precise formal framework. This is not just a concern about tidiness: we need an improved version of BSA to evaluate BSA itself, its assumptions and its consequences. Questions like

- what conception of scientific theories is required by BSA?
- how do standards of simplicity and strength look like?
- how do we calculate the balance of a deductive system?

cannot be addressed employing the 1973 formulation of BSA. In what follows I will provide an aswer to the first two questions and suggest possible replies to the third one.

To this end in the rest of this section we will attempt a reconstruction of BSA. Assuming that scientific theories can be formalized, we treat them as theories in model-theoretic sense.[6] To add further generality, we abstract from a particular deductive system (in Model Theory it is usually first order classical logic) using a general theory of logical calculi such as the one developed in Abstract Algebraic Logic.[7] This latter step enables us to vary the inferential environment in which a scientific theory lives and study the consequences.

[5] See (Font, Jansana, and Pigozzi 2003, p. 5 and subsection 2.2).

[6] The founding fathers of this approach are, among the others, Tarski and Carnap, see (Tarski 1944, pp. 346–347), (Tarski 1994) and (Carnap 1937). For more recent considerations on this stance see (da Costa and French 2000), for a classic text of Model Theory see (Chang and Keisler 1990).

[7] See (Font, Jansana, and Pigozzi 2003).

2.1 Logical languages and formulas

Prior to outlining the definition of deductive system, let us define a formal language along the lines of Johnstone's presentation.[8] For the sake of simplicity I will stick to first order languages (for a definition of language appropriate for higher order logic see (Johnstone 2002, p. 940)). Each language can have non-logical symbols for basic sorts, functions and relations: these symbols constitute the signature of the language. A *signature* Σ is thus composed of:

1. A set Σ-Sort of *sorts*, symbols for kinds or families of objects.
2. A set Σ-Fun of *function symbols* together with a map assigning to each function symbol its *type*, a finite non empty list of sorts (where the last sort is the sort of the output). We write $f : A_1 \dots A_n \to B$ to indicate that f has type $A_1 \dots A_n B$ and call n the *arity* of f. If $n = 0$ f is called a *constant* of sort B.
3. A set Σ-Rel of *relation symbols* together with a map assigning to each relation symbols its type, a list of sorts as in the previous case. We write $R : A_1 \dots A_n$ to indicate that R has type $A_1 \dots A_n$ and call n the arity of R. If $n = 0$ R is called an *atomic proposition*.

For each sort A of Σ-Sort we assume to have a countably infinite number of variables of sort A. We now define the *terms* of a language and their sorts recursively (we write $t : A$ to indicate that t is a term of sort A):

1. $x : A$ if x is a variable of sort A.
2. $f(t_1, \dots, t_n) : B$ if $f : A_1 \dots A_n \to B$ and $t_1 : A_1, \dots, t_n : A_n$.

Note that for the second clause constants are terms. The terms are those collections of symbols of the language that stand for individuals (even though they do not always denote a specific one).

The next step is to define the formulas of the language, but to do that we first have to introduce the logical symbols. Roughly speaking[9], logical symbols are defined by a set Con of *quantifiers* and *connectives symbols* together with a map assigning to each connective symbol a natural number n corresponding to its arity. A *language* L is thus composed of a signature Σ, a set Con with the relative map and a set of auxiliary symbols (such as brackets). With the aid of logical symbols we can finally define the set of formulas Fm_L of the language L in the usual recursive fashion:

1. $R(t_1, \dots, t_n)$ belongs to Fm_L if R is a relation of type $A_1, \dots, A_n$ and $t_1 : A_1, \dots, t_n : A_n$.

[8] See (Johnstone 2002, p. 808).

[9] For the sake of brevity we avoid a precise discussion of free and bounded variables. This discussion is inessential for our purposes and these notions should be clear to anyone familiar with basic logic. See (Johnstone 2002, p. 809) for details.

2. $c(\phi_1, \ldots, \phi_n)$ belongs to Fm_L if c is an n-ary connective and $\phi_1, \ldots, \phi_n$ are formulas.

3. $qx.\phi(x)$ belongs to Fm_L if q is a quantifier and $\phi(x)$ is a formula with free variable x.

The formulas obtained via the first condition are called *atomic formulas*. By definition they are completely independent from the choice of connectives. The set Fm_L is thus generated combining atomic formulas by means of connectives and quantifiers. In general, formulas are assertions about individuals.

2.2 Deductive systems and theories

Now that we have all the linguistic notions in place, let us turn to the definition of deductive system. Following (Font, Jansana, and Pigozzi 2003), a *deductive system* or a *logic* in a language L is a pair $S = \langle Fm_L, \vdash_S \rangle$ where $\vdash_S$ is a substitution invariant *consequence relation* on Fm_L, i.e., a relation $\vdash_S \subseteq \wp(Fm_L) \times Fm_L$ satisfying:

1. if $\phi \in X$ then $X \vdash_S \phi$.

2. if $X \vdash_S \phi$ for all $\phi \in Y$ and $Y \vdash_S \psi$ then $X \vdash_S \psi$.

Intuitively $\vdash_S$ represents all the inferential procedures of a deductive system. When such relation holds between a set of formulas Γ and a formula ϕ we write $\Gamma \vdash_S \phi$ to mean that we can derive the formula ϕ, the conclusion, applying the inferential procedures of $\vdash_S$ to the formulas in Γ, the premises. In general, a deductive system is nothing more than a machinery to make proofs in a certain language, it is a purely syntactical inferential engine.

As this definition shows, a deductive system is dependent on the language, or, more precisely, on the set of formulas generated by a certain language. But there is, as we have seen, a distinction between logical and non-logical symbols, between the set Con and the signature of a language. The reason for this distinction is that a deductive system is dependent on the connectives and quantifiers but not on the signature. Logical symbols play an essential role in inferential processes, while the non-logical symbols are idle in this respect.

The *theorems* of S are the formulas ϕ such that $\emptyset \vdash_S \phi$, that is, the formulas that can be proved without any premise. There are different ways to present a deductive system: for example as an axiomatic calculus, as a natural deduction calculus or as a sequent calculus. Given that the issue of the number of axioms is important in Lewis' definition of the criterion of simplicity, let us spend a few words on the axiomatization of deductive systems (we will return to the problem in Subsection 3.1.1). A *Hilbert-style calculus* is a pair $P = \langle Ax, Ru \rangle$ consisting of a set of axioms and a set of inference rules, where by 'inference rule'we mean any

pair $\langle \Gamma, \phi \rangle$ and by axiom a rule of the form $\langle \emptyset, \phi \rangle$ (which is usually written simply as ϕ). In what follows we will use the term 'inference rule' to refer to inference rules stricto sensu, not to axioms.

A pair $\langle Ax, Ru \rangle$ is a *presentation* of a deductive system S if $\Gamma \vdash_S \phi$ iff ϕ is contained in the smallest set of formulas that includes Γ together with all substitution instances of the axioms of Ax, and is closed under direct derivability by the inference rules in Ru.

The same deductive system can have different presentations: given two presentations P_1 and P_2 in the same language, it is sufficient that the consequence relation $\vdash_1$ associated with P_1 is the same as the consequence relation $\vdash_2$ associated with P_2. This for example happens when, given the same inference rules and two different sets of axioms Ax_1 and Ax_2, we can derive all the axioms of Ax_1 from Ax_2 and vice versa.

We define an *S-theory* (or just a *theory*, when S is understood) as a set of formulas Γ closed under the consequence relation $\vdash_S$, i.e., such that if $\Gamma \vdash_S \phi$ then $\phi \in \Gamma$. In words, Γ is closed under the consequence relation if every formula that can be derived from the formulas in Γ is already in Γ. The smallest S-theory will be of course the set of theorems of S, and, as can be easily seen, the set of theorems of S is included in every S-theory. In what follows we will use the symbols $\mathbf{T}_1$, $\mathbf{T}_2$, etc to refer to theories, in order to distinguish them from ordinary sets of formulas.

A S-theory $\mathbf{T}$ is *generated* by a set of formulas Θ if, for all ϕ, $\phi \in \mathbf{T}$ iff $\Theta \vdash_S \phi$, that is to say, if we can derive any formula of $\mathbf{T}$ from Θ and no formula that can be derived from Θ is outside $\mathbf{T}$. Given any presentation $P = \langle Ax, Ru \rangle$ of S, the set of theorems of S is generated by (the substitution instances of) the statements in Ax. Given our previous characterization of the presentation of a deductive system, we will use the term 'axiom' only to indicate the statements used in a Hilbert-style presentation, and we will employ the term 'hypothesis' to denote the statements used to generate an S-theory different from the trivial one composed only of theorems. We can have different sets of hypotheses for the same S-theory, and these sets can be partially overlapping or completely disjoint. We will use the term 'presentation of theory $\mathbf{T}$' to refer to a set of hypotheses $\Theta^{\mathbf{T}}$ used to generate $\mathbf{T}$.

2.3 Old and new

How do these concepts relate to Lewis'? What we called deductive system has no counterpart in Lewis' account, probably because of the fact that he was considering only one logic, classical logic, and thus he had no need to introduce further distinctions. What Lewis terms 'deductive system' is, in our framework, an S-theory. An S-theory is then what corresponds to a scientific theory. By def-

inition, an S-theory **T** is deductively closed, every formulas that can be deduced from those in **T** is already contained in **T**.

Furthermore, an S-theory is axiomatizable in the sense that it can be generated by a set of hypotheses Θ. We have thus recovered most of Lewis' original idea of a deductive systems as "deductively closed, axiomatizable sets of true sentences". Is there a sense in which an S-theory is a set of *true* sentences?

The answer to this question is: no, unless we take some semantic considerations into account. These would add another layer to our framework. For the rest of this article we will remain at the level of the syntax, running the risk of oversimplification, and leave the semantic side to be developed in future work.

Let us summarize what we have defined in this section. In the framework here presented a scientific theory is composed of the following ingredients:

1. a language L, composed of a signature Σ, a set Con of connectives with the relative maps and some auxiliary symbols.

2. a deductive system S, defined by a consequence relation on the set of formulas generated by L.

3. a set of hypothesis Θ.

A concrete example of a scientific theory presented in a similar fashion can be found in "Axiomatic Foundations of Classical Particle Mechanics" by McKinsey, Sugar and Suppes (McKinsey, Sugar, and Suppes 1953).

2.4 The mathematical apparatus of scientific theories

I have so far ignored the mathematical apparatus employed by many scientific theories. How does mathematics fit into the picture just described? The answer is: we treat mathematical theories as theories in a model-theoretic sense and we add them to the other hypotheses. Therefore, if a scientific theory **T** is using a particular piece of mathematics, an axiomatization of the mathematical notions employed in **T** will be included in the set of hypotheses $\Theta^{\mathbf{T}}$. If, for example, a scientific theory uses real numbers to represent some parameters, we will insert in the mathematical hypotheses an axiomatization of the arithmetic of real numbers.[10]

In this respect it is worth noting that to be able to axiomatize certain mathematical theories we may require a language rich enough to formulate the axioms ('mathematical hypotheses' in our terminology) and a deductive system powerful enough to deduce the desired theorems (some mathematical theories may require second order logic, for example).[11] As a consequence, because of the

[10] As, for example, the one in (Tarski 1994, p. 205).

[11] For a thorough discussion of this matter see (Parsons 2010).

mathematics they employ, some scientific theories cannot even be formulated without assuming a core vocabulary and some kind of minimal deductive power.

The advantage of this account of mathematics is an extreme flexibility: we can tailor the mathematical notions to the need of a scientific theory and study what happens when we modify such notions or their axiomatization (see Subsections 3.1.1 and 3.2 for the implications for simplicity and strength). Moreover, without any specific commitment to the content of the mathematical and non-mathematical hypotheses, we can reasonably hope to describe both the highly-formalized scientific theories, where mathematics is pervasive and there are few non-mathematical hypotheses, and the non-formal scientific theories, where very few mathematical assumptions will be coupled with many non-mathematical hypotheses. Another point worth making is that in this account there is no syntactic characteristic to distinguish between mathematical and non-mathematical hypotheses, in the sense that both are treated as formal statements (maybe the former are more heavily formalized than the latter).

3 Redefining the Core Notions

Having defined a scientific theory as an $\mathcal{S}$-theory, I now turn to the discussion of simplicity and strength. Before analyzing how a theory can be simpler or stronger than another one, however, there is an important observation to make. The comparison between two theories is meaningful, I believe, only if these theories are about overlapping domains of events. To explain this with an example, if I am interested in the laws of nature governing the electromagnetic phenomena I will consider theories that model this kind of phenomena, not Population Biology. This means that at least a naive idea of the intended semantics of our theories is needed if we want to avoid useless comparisons between unrelated theories.

With this in mind and the aid of the framework just defined, let us now turn to simplicity, strength and balance, in this order. In what follows I will intend all the relations in their weak version, that is, we will use the terms 'subset' as short for 'subset or equal', 'less' for 'less or equal in number', and so on.

3.1 Simplicity

We start analyzing simplicity by having a closer look at Lewis' formulation.

3.1.1 Conceptual Simplicity

In the Lewisian model a theory $\mathbf{T}_1$ is simpler than a theory $\mathbf{T}_2$ if $\mathbf{T}_1$ has fewer 'axioms' than $\mathbf{T}_2$. In light of the previous discussion, I maintain that this statement

is too vague. Translating this definition into the new terminology one obtains two definitions:

1. $\mathbf{T}_1$ is simpler than $\mathbf{T}_2$ if $\mathbf{T}_1$ has fewer axioms than $\mathbf{T}_2$.
2. $\mathbf{T}_1$ is simpler than $\mathbf{T}_2$ if $\mathbf{T}_1$ has fewer hypotheses than $\mathbf{T}_2$.

depending on how one interprets Lewis' term 'axiom'. There are two observations to make. The first one is that 1 is arguably in contrast with other notions of simplicity. Consider two S-theories $\mathbf{T}_1$ and $\mathbf{T}_2$ with the same language, the same deductive system and the same hypotheses. The difference between $\mathbf{T}_1$ and $\mathbf{T}_2$ lies in the presentation of the deductive system S: the presentation in $\mathbf{T}_1$ has, say, 3 axioms and 2 inference rules; the presentation in $\mathbf{T}_2$ has 10 axioms and 2 inference rules. As can be easily inferred, the derivations of theorems in $\mathbf{T}_1$ will be generally longer than the derivations in $\mathbf{T}_2$, for the derivations in $\mathbf{T}_1$ will require multiple uses of the same axioms to obtain lemmas that can be easily derived in $\mathbf{T}_2$. If $\mathbf{T}_1$ has fewer axioms then derivations in $\mathbf{T}_1$ are more complicated from a computational point of view (see Subsection 3.1.3). The second observation concerns the second definition of simplicity, *conceptual simplicity* from now on. We can have two versions of conceptual simplicity, a *ceteris paribus* one and a general one:

Definition 1 (*Ceteris paribus conceptual simplicity* (CPCS)) *For every pair of theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *sharing the same language* L *and the same deductive system* S*, we define:*

$$\mathbf{T}_1 \textit{ is simpler}_{CPCS} \textit{ than } \mathbf{T}_2 \textit{ if } \mathbf{T}_1 \textit{ has fewer hypotheses than } \mathbf{T}_2.$$

Definition 2 (General conceptual simplicity (GCS)) *For every pair of theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$$\mathbf{T}_1 \textit{ is simpler}_{GCS} \textit{ than } \mathbf{T}_2 \textit{ if } \mathbf{T}_1 \textit{ has fewer hypotheses than } \mathbf{T}_2.$$

As can be easily seen, CPCS is just GCS restricted to theories sharing the same language and deductive system. In its domain of applicability CPCS is an effective measure of simplicity, but such domain is extremely narrow and CPCS cannot be regarded as more than a limiting case. GCS, on the other hand, is defined on every pair of theories, and it is probably the closest to (our interpretation of) Lewis' relation of simplicity. Notably, 'having fewer hypotheses' does not mean that the first set of hypotheses is included in the other as a subset. Substituting the condition of set-theoretical inclusion to the condition on the number of hypotheses one obtains two different relations: CPCS* and GCS*. By the definition of S-theory, CPCS* is nothing more than the relation of inclusion between different set of hypotheses generating the same theory (that is, it is applicable only if $\mathbf{T}_1$ and $\mathbf{T}_2$ coincide). An interesting version of GCS* is:

Definition 3 (Mathematical Simplicity (MS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$\mathbf{T}_1$ *is simpler*$_{MS}$ *than* $\mathbf{T}_2$ *if* $MH(\mathbf{T}_1))$ *is a subset of* $MH(\mathbf{T}_1)$.

where $MH(\mathbf{T}_1))$ and $MH(\mathbf{T}_2))$ denote the set of mathematical hypotheses of $\mathbf{T}_1$ and $\mathbf{T}_2$ respectively.

The reason why a small number of hypotheses is preferable is quite clear: a compact theory is easier to handle and to understand.

Nevertheless, somebody may wonder why the number of hypotheses should be an indicator of simplicity in the first place. The obvious objection is: given a language with conjunction, it is possible to conflate finitely many formulas into one by taking the conjunction of them (even infinitely many, if the language has infinitary conjunctions). This of course makes the counting of hypotheses an irrelevant matter. This however is not a problem, for two reasons. The first, of a pragmatic flavour, is simply that there are no theories with hypotheses where the conjunction is the main connective. The second is that, even if we want to avoid pragmatic considerations, it is possible to write a simple computer programs that, in counting the number of hypotheses, checks whether the hypotheses have a conjunction as outer connective. If this is the case, the program consider the subformulas as distinct hypotheses, and restarts the counting (and the check). As long as we have hypotheses made of finitely many symbols, the program will output the correct number of axioms, despite of conjunctions.

Nevertheless, the number of hypotheses is just one of the components of a theory, and we should also consider the role of languages and deductive systems.

3.1.2 Expressive simplicity

As far as the language is concerned, we can compare two theories in terms of the expressive power of their signatures, of their *expressive simplicity*. Let us explain this with an example. Consider two theories $\mathbf{T}_1$ and $\mathbf{T}_2$ such that in both their signatures there is a sort A. In the language of $\mathbf{T}_1$ there is a symbol for a constant of sort A, while in the language of $\mathbf{T}_2$ there is no such symbol, and thus to refer to the same object we have to use a paraphrase like "the object of sort A satisfying conditions x, y, etc". The same argument can be applied to every other symbol of the signature: to function symbols ("the function of type $A\ldots$ satisfying conditions x, y, etc") and to relation symbols ("the relation of type $A\ldots$ satisfying conditions x, y, etc").

The signature of $\mathbf{T}_2$ is simpler in the sense that it has less symbols and that some symbols of $\mathbf{T}_1$ can be substituted by a combination of symbols of $\mathbf{T}_2$. This feature can be important if we want to minimize the number of primitives for foundational purposes. The signature of $\mathbf{T}_1$ is simpler in the sense that is less cumbersome, instead of repeating a long list of symbols we can just employ a shorter expression. This can make the difference, for example, from a didactic

perspective or for computational complexity. We have here two conflicting notions of simplicity.

Definition 4 (Expressive Simplicity with Less Symbols (ESLS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$$\mathbf{T}_1 \textit{ is simpler}_{ESLS} \textit{ than } \mathbf{T}_2 \textit{ if } \mathbf{T}_1 \textit{ has less symbols than } \mathbf{T}_2$$

Definition 5 (Expressive Simplicity with More Symbols (ESMS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$$\mathbf{T}_1 \textit{ is simpler}_{ESMS} \textit{ than } \mathbf{T}_2 \textit{ if } \mathbf{T}_1 \textit{ has more symbols than } \mathbf{T}_2$$

Explicitating a particular kind of symbols in these definitions, respectively sort, function and relation symbols, we have three more specific versions of ESLS and ESMS. Along these lines, the importance of ESLS and ESMS can be weighed relatively to the symbols under examination: we may want a symbol with a pivotal role in our theory, say, the constant representing the speed of light, to be included in the signature, while a conceptually subordinate symbol may be defined in terms of others.

3.1.3 Computational Simplicity

It is also possible to find notions of simplicity connected with the deductive system of a theory. Consider for example the following case. Given a set of formulas Γ regarded as true, say, a set of formulas representing empirical observations or some important theorems, a theory $\mathbf{T}_1$ may be judged simpler than a theory $\mathbf{T}_2$ if the derivations of the formulas in Γ in $\mathbf{T}_1$ are 'simpler' than the corresponding derivations in $\mathbf{T}_2$.

But how can a derivation be simpler than another one? Before examining possible candidates of *computational simplicity*, one has to qualify two points. First, there are two variables to consider: which deductive system is used and how it is presented. A 'stronger' deductive system, one which is an extension of another one, for example, may produce simpler proofs (see below for examples of what this can mean). A more compact presentation, one employing fewer axioms or inference rules, will usually determine more complex derivations. Second, as far as computational simplicity is concerned, the choice of connectives has to be considered as part of the presentation of a deductive system. A wide set of connectives without the relative axioms or inference rules (say, having the symbol of conjunction but only axioms and inference rules for the symbol of entailment) cannot enhance the simplicity of derivations and, vice versa, axioms and inference rules can be used only in the presence of the relative connective symbol. This is why the choice of connective symbol is relevant for computational simplicity and not only for expressive simplicity. Here are two proposals for computational simplicity:

Definition 6 (Computational Simplicity in Length (CSL)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *and for every set of formulas* Γ*:*

$\mathbf{T}_1$ *is simpler*$^{\Gamma}_{CSL}$ *than* $\mathbf{T}_2$ *if all the derivations of the formulas in* Γ *in* $\mathbf{T}_1$ *are shorter than those in* $\mathbf{T}_2$

To be able to compare the lengths of proofs we have to introduce a measure of such length (usually the number of lines).

Definition 7 (Computational Simplicity in Time (CST)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*, for every set of formulas* Γ *and given a suitable automated theorem prover (a computer program that produces derivations), we define:*

$\mathbf{T}_1$ *is simpler*$^{\Gamma}_{CST}$ *than* $\mathbf{T}_2$ *if all the derivations of the formulas in* Γ *from the hypotheses in* $\mathbf{T}_1$ *take less time than those in* $\mathbf{T}_2$

Depending on the prover employed, this may require that $\mathbf{T}_1$ and $\mathbf{T}_2$ share the same deductive system. As long as Γ consists of a single formula, we can apply CSL and CST without worries. But if Γ contains two or more formulas one could have problems of applicability. Consider a case where $\mathbf{T}_1$ is simpler$^{\Gamma*}_{CSL}$ than $\mathbf{T}_2$ and $\mathbf{T}_2$ is simpler$^{\Gamma+}_{CSL}$ than $\mathbf{T}_1$, where $\Gamma*$ and $\Gamma+$ are two disjoint subsets of Γ. In this case CSL cannot be applied relatively to Γ (an analogous argument can be made for CST). To overcome this impasse and define a universally applicable version of CSL (CST respectively) we may define a total measure of length (respectively time) for the derivations of the formulas in Γ and then compare the total measure in $\mathbf{T}_1$ with the total measure in $\mathbf{T}_2$ instead of comparing derivations pairwise. This approach leads to a generalized version of CSL (respectively CST).

It remains to say why these notions of simplicity are interesting candidates. A common agument can be made for CSL and CST. It is essentially an optimization argument: given any application of a theory (for example checking whether some formulas follow from the theory or not) we prefer the theory that requires less effort to be used. Indeed, the fact that a theory is computationally expensive can be a reason to change or improve the theory.

3.2 Strength

For Lewis a theory is stronger than another if it has more informational content (Lewis 1999, p. 41). If we interpret the informational content of a theory $\mathbf{T}$ as all the formulas that can be derived from the hypotheses of $\mathbf{T}$ we have that, by definition of S-theory as a deductively closed set of formulas, the informational content of $\mathbf{T}$ coincides with $\mathbf{T}$. If one sticks to this interpretation it is possible to formulate strength as:

Definition 8 (General Strength (GS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$\mathbf{T}_1$ *is stronger*$_{GS}$ *than* $\mathbf{T}_2$ *if* $\mathbf{T}_2$ *is a subset of* $\mathbf{T}_1$

GS is interesting because it encodes the fact that we can reduce one theory to another, that is, we can prove all the statements of the first one inside the second one. There are cases, however, where GS cannot be applied. A sets of formulas can be included in another only if they share the same language, or the language of the bigger set is an extension of the other one. Another approach could be the following. Given a set of true formulas Γ, say, the formulas representing the observations made, the informational content of theory $\mathbf{T}$ is the portion of Γ that is derivable from the hypotheses of $\mathbf{T}$, that is, the intersection between Γ and $\mathbf{T}$. Of course the formulas in Γ have to refer to the shared part of the intended domain of interpretation, otherwise one of the two theories will be weaker a priori. We then have:

Definition 9 (Informational Strength (IS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*, and given a set of formulas* Γ*:*

$\mathbf{T}_1$ *is stronger*$^{\Gamma}_{IS}$ *than* $\mathbf{T}_2$ *if the informational content of* $\mathbf{T}_1$ *relative to* Γ *is bigger than that of* $\mathbf{T}_2$

where by bigger I mean cardinality-wise. One can of course restrict this notion substituting 'is bigger than' in the definition with 'includes' obtaining IS*. Obviously, IS* entails IS for every Γ.

We could also relate the notion of strength to that of deductive system:

Definition 10 (Computational Strength (CS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$\mathbf{T}_1$ *is stronger*$_{CS}$ *than* $\mathbf{T}_2$ *if* $\vdash_{\mathbf{T}_2}$ *is a subset of* $\vdash_{\mathbf{T}_1}$

In other words, the deductive system of $\mathbf{T}_1$ is stronger$_{CS}$ than that of $\mathbf{T}_2$ if in $\mathbf{T}_1$ we can derive every formula derivable in $\mathbf{T}_2$ and some more. Notably, if $\mathbf{T}_1$ and $\mathbf{T}_2$ share the same set of hypotheses then CS implies GS and IS* for every Γ. CS holds even though $\mathbf{T}_1$ and $\mathbf{T}_2$ do not share the language, as the language of, say, $\mathbf{T}_1$ can be an extension of that of $\mathbf{T}_2$.

Along the same lines of CS one can introduce a notion of strength connected with the mathematical apparatus of theories. A first option can be the inverse relation of MS:

Definition 11 (Mathematical Strength (MSt)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$\mathbf{T}_1$ *is stronger*$_{MSt}$ *than* $\mathbf{T}_2$ *if* $MH(\mathbf{T}_2)$ *is a subset of* $MH(\mathbf{T}_1)$

We have here a straightforward example of the conflict between a relation of strength and a relation of simplicity: if $\mathbf{T}_1$ is simpler$_{MS}$ than $\mathbf{T}_2$ then $\mathbf{T}_2$ is stronger$_{MSt}$ than $\mathbf{T}_1$. However, this is not the case in general for the notions that we defined, for example Expressive Simplicity is independent from Mathematical Strength. Hence the trade-off between simplicity and strength mentioned by Lewis is a consequence of particular selections of notions of simplicity and strength.

Alternatively, one can impose a further condition to have a more informative relation:

Definition 12 (Strict Mathematical Strength (SMS)) *For every two theories* $\mathbf{T}_1$ *and* $\mathbf{T}_2$*:*

$$\mathbf{T}_1 \textit{ is stronger}_{SMS} \textit{ than } \mathbf{T}_2 \textit{ if } MH(\mathbf{T}_2) \textit{ is a proper subset of } MH(\mathbf{T}_1)$$

This last relation might be appealing if we think that a particular mathematical theory is essential to model a certain class of phenomena, say, Hilbert spaces to model Quantum phenomena, and we want to draw a distinction between theories that employ such mathematical machinery and theories that do not.

3.3 Balance

Depending on the notions of simplicity and strength adopted, we can define balance in many ways. Following the characterization of simplicity and strength as binary relations, I will treat balance as a binary relation as well, that is to say, I will consider relative balance. In the presence of some absolute measures of simplicity and strength, absent in the present work, one may attempt a definition of the absolute balance of a theory.

As can be easily checked, apart from SMS all the relations defined are preorders in their respective domain of applicability, that is, they are reflexive and transitive. With this in mind, let us sketch two general procedures to define the balance. Suppose we have a set of theories to evaluate and a collection of relations of simplicity and strength.

The first procedure, of a qualitative nature, consists of aggregating the orderings of the set of theories produced by the chosen relations. Formally, this means that given n orderings $R_1, \ldots, R_n$ we want to have a procedure to obtain a single ordering R. The top theory/theories according to this last relation will be the best system(s). Of course, depending on how we aggregate these orderings we will obtain different outcomes. One first question to pose in this respect is: are all orderings equally relevant or do we regard some criteria as privileged?

A mathematical environment where such an aggregation procedure can be studied is provided by Social Choice.[12] To make an example, in this framework the condition encoding the idea that all orderings must be equally relevant is called anonimity (invariance of the aggregator under the permutations of the input orderings). In this context, given two theories $\mathbf{T}_1$ and $\mathbf{T}_2$ and k relations corresponding to the equally relevant selection criteria, we may say that $\mathbf{T}_1$ is better than $\mathbf{T}_2$ if $\mathbf{T}_1$ is preferable according to $k/2 + 1$ relations. The extent to which results and techniques of Social Choice can be applied to the present case will be explored in future work.

[12] For a standard reference in the field see (Gaertner 2009).

The second procedure involves the definition of quantitative measures relative to the chosen relations. If, say, $\mathbf{T}_1$ is simpler$_{GCS}$ than $\mathbf{T}_2$ we could take the difference between the number of hypotheses in $\mathbf{T}_2$ and the number of hypotheses in $\mathbf{T}_1$ as a number representing how much simpler $\mathbf{T}_1$ is compared to $\mathbf{T}_2$. By similar methods, counting or using percentages, one may associate a function to each relation in order to evaluate the relative degree of simplicity or strength. If this attempt succeeds one can then use these functions to construct an algorithm able to analyze the set of theories, apply such functions and combine the results to find the theories that score the best combination according to the chosen relations of simplicity and strength. To continue the example above, we could assign weights $n_1, \ldots, n_k$ to the k relations and say that the score of $\mathbf{T}_1$ is the sum of the weights of the relations in which $\mathbf{T}_1$ is preferable over $\mathbf{T}_2$. We could then conclude that $\mathbf{T}_1$ is better than $\mathbf{T}_2$ if $\mathbf{T}_1$ has a higher score.

Before concluding, we make three final remarks. The first is that the choice of the collection of relations of simplicity and strength does not influence the balance function just by changing the arity of its input. In the second methodology a particular choice of relations might change the internal structure of the algorithm. For example, if we employ General Strength we might want the algorithm to check this relation first, to know whether one theory is reducible to the other. The second remark is that in both cases if the chosen relations cannot be applied to the set of theories, because theories do not share enough features for the relations to be applied, we could not find any best system. The third remark concerns the viability of the two methodologies. Both of them are applicable only if the chosen relations are decidable. If they are not, then in the first case we might not get the orderings at all, and in the second case the algorithm may not terminate.

3.4 Conclusion

Let us draw some conclusions. In light of the formal analysis outlined and of the examples offered, I argue that the aforementioned framework is appropriate for a precise characterization of the notions of simplicity, strength and balance. Moreover, I believe that the plurality of definable notions of simplicity (respectively, strength and balance) casts doubt on Lewis' reliance on a single concept and demands for a more comprehensive discussion. Simplicity, strength and balance are, I think, multifaceted ideas, and the search for a unique characterization could be misleading. This of course does not imply that such notions have to be vague, as the present work showed.

Indeed we have alternative versions of BSA depending on

1. which relations of simplicity and strength we use
2. how do we aggregate them to obtain the balance

It is already hard to reach a consensus on the first item. For an experimental physicist, interested in testability and implementations, theories may be compared with an eye for their computational features. A philosopher, on the other hand, could think that the best theory is one with few primitives.

The advantage of our framework, as long as it is considered tenable, is that now we can look at specific, well defined candidates for relations of simplicity and strength. Likewise, we can design and analyze procedures to obtain the balance. This means that the discussion about item 1 and 2, although still philosophical in nature, is now more formally grounded.

References

Armstrong, David M. (1983). *What is a Law of Nature?* Cambridge: Cambridge University Press.

Bird, Alexander (1998). *Philosophy of Science*. London: UCL Press.

— (2008). "The Epistemological Argument against Lewis's Regularity View of Laws". In: *Philosophical Studies* 138, pp. 73–89.

Carnap, Rudolf (1937). *The Logical Sintax of Language*. Trans. by A. Smeaton. London: Routledge 2001.

Chang, Chen Chung and H. Jerome Keisler (1990). *Model Theory*. 3rd ed. Studies in Logic and the Foundations of Mathematics 73. Amsterdam: North-Holland.

Cohen, Jonathan and Craig Callender (2009). "A Better Best System Account of Lawhood". In: *Philosophical Studies* 145, pp. 1–34.

Da Costa, Newton and Steven French (2000). "Models, Theories, and Structures: Thirty Years On". In: *Philosophy of Science (Proceedings)* 67, S116–S127.

Font, Josep M., Ramon Jansana, and Don Pigozzi (2003). "Survey of Abstract Algebraic Logic". In: *Studia Logica* 74 (Special issue on Abstract Algebraic Logic, Part II), pp. 13–97. With an update in 2009, 91: 125-130.

Gaertner, Wulf (2009). *A Primer in Social Choice Theory*. Oxford: Oxford University Press.

Jaeger, Lydia (2002). "Humean Supervenience and Best-System Laws". In: *International Studies in the Philosophy of Science* 16.2, pp. 141–155.

Johnstone, Peter T. (2002). *Sketches of an Elephant: A Topos Theory Compendium*. Vol. 2. Oxford: Oxford University Press.

Lewis, David (1973). *Counterfactuals*. Oxford: Blackwell.

— (1986). *Philosophical Papers*. Vol. 2. New York: Oxford University Press.

— (1994). "Humean Supervenience Debugged". In: *Mind* 103, pp. 473–490.

— (1999). *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

McKinsey, John C.C., A.C. Sugar, and Patrick Suppes (1953). "Axiomatic Foundations of Classical Particle Mechanics". In: *Journal of Rational Mechanics and Analysis* 2.2, pp. 253–272.

Mumford, Stephen (2004). *Laws in Mature*. London: Routledge.

Parsons, Charles (2010). "Some Consequences of the Entanglement of Logic and Mathematics". In: *Reference and Intentionality: Themes from Føllesdal*. Ed. by W.K. Essler and M. Frauchiger. Frankfurt: Ontos Verlag.

Psillos, Stathis (2002). *Causation and Explanation*. Chesham: Acumen.

Robert, John (1999). ""Laws of Nature" as an Indexical Term: A Reinterpretation of Lewis's Best-System Analysis". In: *Philosophy of Science* 66 (Proceedings), S502–S511.

Tarski, Alfred (1944). "The Semantic Conception of Truth and the Foundations of Semantics". In: *Philosophy and Phenomenological Research* 4.3, pp. 341–376.

— (1994). *Introduction to Logic and to the Methodology of Deductive Sciences*. 4th ed. Oxford: Oxford University Press.

Van Fraassen, Bas (1989). *Laws and Symmetry*. Oxford: Clarendon Press.

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

Una teoria della razionalità: il modello BDI

Costanza Larese

Abstract. In quest'articolo propongo un'analisi di una teoria della razionalità, il modello *Belief-Desire-Intention* (BDI), con l'obiettivo di stabilirne la fecondità teoretica. Interpreto il modello come il risultato dell'indebolimento di alcuni principi cardine della teoria della scelta razionale: se questa è di natura normativa e considera agenti altamente idealizzati, il modello BDI è invece motivato dallo scopo di dare una caratterizzazione cognitivamente plausibile delle azioni degli individui e inserisce nella definizione di razionalità aspetti non normativi. Per questa ragione, la teoria BDI introduce il concetto di intenzione e complica la propria ontologia: le intenzioni pongono dei vincoli di consistenza sulla componente motivazionale dell'individuo e fungono da filtro di ammissibilità sulla selezione di altre intenzioni (Bratman 1987). Presento ed analizzo di seguito due formalizzazioni, sviluppatesi in due diverse aree di ricerca (logica e intelligenza artificiale), dei principi filosofici della teoria: il sistema BDICTL*-W3 (Georgeff e Rao 1998) ed un esempio di *Agent Control Loop* (Wooldridge 2000). La discussione vuole rilevare le peculiarità dei vari approcci alla teoria in oggetto, individuare i nodi concettuali comuni ma anche le specificità di ciascun apporto. Concludo quindi con alcune osservazioni di carattere epistemologico sui vantaggi di un approccio plurale.

Keywords. Razionalità, Modello BDI, Intenzione, Plausibilità cognitiva, Normatività.

Copyright. © 2018 Costanza Larese. Pubblicato in Italia. Alcuni diritti riservati. Prima pubblicazione su *Rivista Italiana di Filosofia Analitica Junior* 4:1 (2013), pp. 12-28.
Autore. Costanza Larese, costanza.larese@sns.it.

Introduzione

Il modello BDI è una teoria della razionalità che studia il ragionamento di individui razionali. Prima di cominciare l'analisi del modello, propongo subito due esempi del tipo di ragionamento che intendo discutere:

Esempio 1 (Bratman 1990, p. 23) *Il terrorista e l'attentatore strategico hanno entrambi lo scopo di promuovere azioni militari per danneggiare il nemico. Entrambi intendono realizzare il proposito tramite bombardamenti. Il piano del terrorista prevede di bombardare la scuola nel territorio nemico, uccidendo i bambini, terrorizzando la popolazione, costringendo così il nemico alla capitolazione. Il piano dell'attentatore strategico prevede invece di colpire il magazzino delle munizioni nemiche, minando gli sforzi bellici dell'avversario. Tuttavia, l'attentatore strategico sa anche che accanto al magazzino di munizioni vi è una scuola; pur essendosi preoccupato dell'effetto disumano dell'azione, l'attentatore ritiene che il guadagno per l'esito della guerra dato dalla distruzione delle munizioni nemiche sia superiore al costo dell'operazione.*

Esempio 2 (Georgeff e Rao 1998, p. 298) *Phil occupa un seggio alla Camera dei Rappresentanti e, in vista delle prossime elezioni, crede di avere le seguenti possibilità: candidarsi nuovamente per il seggio alla Camera, candidarsi per un posto al Senato oppure ritirarsi dalla politica. Non considera seriamente l'opzione di ritirarsi dalla vita politica, mentre è certo di poter mantenere il seggio alla Camera. Phil deve decidere se indire o meno un sondaggio, il cui esito sarà il consenso oppure il dissenso della maggioranza riguardo al suo passaggio al Senato. Sulla base del risultato del sondaggio, Phil deciderà se candidarsi alla Camera oppure al Senato.*

Nel primo esempio, i piani elaborati da entrambi gli agenti prevedono come effetto la strage dei bambini: tuttavia, mentre il terrorista intende colpire la scuola per danneggiare il nemico, l'attentatore strategico non intende uccidere i bambini, ma giudica la conseguenza del suo piano un mero effetto collaterale. Si vedrà come il potere espressivo del modello BDI permetta di rappresentare la distinzione tra gli effetti intesi e gli effetti collaterali presenti nello scenario possibile selezionato dall'agente. Poichè intuitivamente si può pensare ad una identificazione di terrorismo e irrazionalità, è opportuno notare subito che il modello BDI, come la teoria classica della scelta razionale, assume il Principio di Neutralità: la razionalità nella teoria è indipendente dal contenuto delle intenzioni.

Il secondo esempio analizza il ragionamento di un agente in condizione epistemica di incertezza: nel contesto accadono degli eventi su cui l'agente non ha un controllo diretto e a cui può solo assegnare una probabilità soggettiva che esprime il suo grado di convinzione del loro eventuale verificarsi. Phil non sa se

la maggioranza approverà o meno il suo passaggio al Senato e neppure conosce l'esito delle elezioni.

1 Due modelli di razionalità

1.1 Teoria della scelta razionale: idealizzazione e normatività

Analizzo ora i principi cardine della teoria classica della scelta razionale: mi riferirò alle decisioni individuali in condizioni di certezza poichè i modelli di decisione razionale per problemi più complessi non fanno che estendere tali fondamenti.

Fissati un insieme A di alternative reali e un insieme E di esiti, la teoria considera il comportamento di scelta di un agente dotato di preferenze personali e capace di codificare informazioni. Dato l'insieme non vuoto di alternative reali A, una *funzione di scelta* S ne restituisce un sottoinsieme non vuoto $S(A)$, detto l'insieme scelto: $\emptyset \neq S(A) \subseteq A$. Le *preferenze* di un agente i sono espresse da una relazione binaria sull'insieme non vuoto degli esiti E, $\succ_i \subseteq E^2$: posti $e, f \in E$, si scrive $e \succ_i f$ per indicare che l'agente i preferisce l'esito e a quello f. In condizioni di certezza, ogni alternativa reale $a \in A$ determina in modo univoco un esito $e \in E$. L'idea fondamentale della teoria classica della scelta razionale è data dal Principio di Massimizzazione, secondo cui un agente è razionale se e solo se si comporta in modo massimizzante rispetto agli esiti in E. Questo principio è normativo, perchè da un lato stabilisce una norma di decisione a cui gli agenti devono uniformarsi, dall'altro fissa un criterio con cui determina quali sono i comportamenti irrazionali.

A partire da una relazione di preferenza che rispetti determinati vincoli, si può definire una funzione di scelta che soddisfi il Principio di Massimizzazione. Il modello impone che $\succ_i$ sia una relazione d'ordine, cioè che rispetti i vincoli di asimmetria, completezza e transitività e di conseguenza esclude tutti quegli individui che, violando tali assiomi, sono detti irrazionali. In questo modo la teoria caratterizza come irrazionali, e dunque esclude, la gran parte degli individui concreti: questi infatti esibiscono spesso preferenze incomplete oppure preferenze acicliche ma non transitive. Si consideri un semplice esempio: Phil deve scegliere una bustina di tè fra tre possibilità, Earl Grey (G), tè verde (V) e tè Jasmine (J). L'unico criterio rilevante ai fini della sua scelta è costituito dall'aroma di ciascun infuso. Phil non preferisce V a J nè J a V: del resto, il tè Jasmine è preparato con tè verdi, per cui Phil non apprezza la differenza tra i due aromi. Tuttavia, Phil preferisce J a G e G a V. Gli attori della teoria classica della scelta razionale sono perciò agenti ideali.

Si chiama elemento ottimale rispetto all'ordine di preferenza l'esito $opt_E = \{e^* \in E | e^* \succ e, \forall e \in E\}$, cioè quell'esito preferito a tutti gli altri. Si dimostra

che $S(E) = opt_E$ è una funzione di scelta che, come conseguenza immediata, soddisfa il Principio di Massimizzazione. Si noti come la consistenza delle preferenze sia una condizione essenziale per la soddisfazione del Principio di Massimizzazione: imporre come norma di decisione la selezione dell'esito ottimale richiede che la relazione di preferenza sugli esiti sia consistente. Il Teorema Fondamentale dell'Utilità stabilisce che se $\succ$ è un ordine, allora esiste una funzione $u : E \longrightarrow \mathbb{R}$ tale che $e \succ f \Leftrightarrow u(e) > u(f)$: in altre parole, i vincoli di consistenza imposti sulle preferenze sono sufficienti a garantirne una rappresentazione numerica, che si interpreta come utilità. Questo Teorema permette di catturare formalmente la definizione di decisione razionale: un individuo decide razionalmente se e solo se massimizza la propria funzione di utilità individuale, cioè sceglie quell'alternativa reale a^* che determina l'esito ottimale e^*, cioè l'argomento massimo della funzione di utilità.

In conclusione, la teoria assume che le preferenze dell'agente siano consistenti ed impone che la scelta sia massimizzante rispetto agli esiti: con la prima assunzione, il modello stabilisce di considerare soltanto *agenti ideali*; con la seconda, essa determina il proprio *status normativo*. I concetti di preferenza consistente e di scelta come massimizzazione sono i principi fondamentali di questo modello di razionalità: un agente che violi questi vincoli è caratterizzato come irrazionale ed è escluso dalla teoria. Come si è visto, la consistenza delle preferenze è una condizione necessaria ma non sufficiente affinchè il Principio di Massimizzazione possa essere soddisfatto: analogamente, l'idealizzazione dell'agente è una condizione necessaria ma non sufficiente per la normatività dell'approccio epistemologico del modello.

1.2 Per una teoria dell'azione cognitivamente plausibile

Il modello BDI è una teoria dell'azione razionale che considera *agenti cognitivamente plausibili* a differenza di quelli completamente idealizzati della teoria classica della scelta razionale. Proprio perché l'agente non è completamente idealizzato, l'approccio epistemologico del modello non può essere rigorosamente normativo: insieme ad una riduzione dell'idealizzazione, BDI inserisce nella definizione di razionalità alcuni principi di natura *descrittiva*. Abbassando il grado di idealizzazione dell'agente e inserendo caratterizzazioni descrittive della razionalità, BDI complica l'ontologia della teoria della scelta razionale creando un'ontologia dotata di maggior potere espressivo: nello specifico, analizza ulteriormente la componente motivazionale dei suoi attori in cui introduce il concetto di intenzione. La componente informativa di un individuo resta invece sostanzialmente invariata: come l'agente della teoria della scelta esprime le proprie informazioni sul contesto tramite un'assegnazione di probabilità sugli esiti, così l'agente BDI esprime le proprie convinzioni sul mondo.

Nella teoria della scelta razionale, il processo di idealizzazione dell'agente consiste nell'imposizione di determinati vincoli sulla relazione di preferenza. Viceversa, la plausibilità cognitiva del modello BDI permette di classificare come razionali agenti con *desideri inconsistenti*. Proprio perchè l'insieme delle alternative reali A è interpretato come un insieme di azioni A e un'azione richiede per il suo compimento una determinata quantità di risorse (tempo, informazioni, denaro, energie, memoria, benzina, ...), BDI assume il principio di natura descrittiva per cui gli agenti hanno *limitazioni di risorse*.

Sulla base dell'osservazione secondo cui gli agenti hanno il bisogno di *coordinare azioni* presenti e future tra di esse e queste con le attività degli altri individui, il modello BDI assume che gli agenti elaborino dei *piani*. I piani collegano le diverse intenzioni dell'agente e strutturano le relazioni tra di esse: «*Plans are intentions writ large*» (Bratman 1987, p. 8). Il fatto che l'agente non abbia risorse infinite determina per Bratman due conseguenze sull'architettura dei piani: questi sono parziali e hanno una struttura gerarchica, nel senso che i piani che riguardano i fini contengono quelli sui mezzi per raggiungerli e gli obiettivi più generali vincolano quelli più specifici. Da un lato infatti piani molto dettagliati potrebbero risultare inutili nel momento in cui si dovessero verificare dei cambiamenti nel contesto e sarebbero difficili da formulare data la condizione di incertezza epistemica propria di un agente limitato; dall'altro lato, la gerarchizzazione dei piani permette di considerare intenzioni più specifiche tenendo ferme quelle più generali e quindi di instaurare rapporti di priorità tra i vari obiettivi. Nonostante la necessità di coordinazione inter e intrapersonale sia un'esigenza propria di attori cognitivamente plausibili, l'approccio epistemologico del modello non è descrittivo, perché assume che gli agenti strutturino dei piani ed impone normativamente dei vincoli di consistenza su questi. I piani di un agente razionale devono essere consistenti al loro interno: l'attentatore che intende indebolire l'avversario non può avere l'intenzione di sganciare una bomba e contemporaneamente non voler far uso di armi di distruzione di massa; inoltre, i piani devono essere consistenti rispetto alle convinzioni: l'attentatore non può avere l'intenzione di colpire il magazzino di munizioni nemiche pur essendo convinto che l'avversario non abbia un magazzino di munizioni; infine, i piani non devono rivelare un'incoerenza tra mezzi e fini: in altre parole, le intenzioni devono poter essere concretizzate attraverso la formulazione di piani riguardo i mezzi impiegabili.

Per poter esprimere sia l'inconsistenza dei desideri (aspetto descrittivo), che la consistenza dei piani (aspetto normativo), il modello BDI articola l'analisi della componente motivazionale dell'agente, complicando l'ontologia della teoria della scelta razionale, in cui le preferenze consistenti esprimono le motivazioni dell'individuo. L'ontologia BDI assume quindi che un agente razionale sia dotato non soltanto di *desideri* e *convinzioni*, ma anche di *intenzioni*. L'introduzione

dell'intenzione nell'analisi della razionalità migliora il potere espressivo del modello con cui diventa possibile considerare agenti cognitivamente plausibili. In ciò che segue, si mostrerà come le intenzioni di un agente razionale svolgano un ruolo centrale nella produzione di azioni e come la componente intenzionale di un individuo sia strutturalmente connessa a desideri e convinzioni.

2 Il modello BDI: relazioni tra convinzioni, desideri, intenzioni

***Intention is Choice with Commitment*: intenzioni, desideri, scelte.** Sia le intenzioni che i desideri esprimono le motivazioni di un individuo; tuttavia se le prime devono essere internamente consistenti, i secondi possono essere, e spesso lo sono, inconsistenti. Il modello impone normativamente che le intenzioni siano un sottoinsieme consistente dei desideri dell'agente.

Bratman (1990) nota che i desideri influenzano soltanto *potenzialmente* l'azione, al contrario dell'intenzione che invece controlla effettivamente la condotta successiva, interrompendo l'agente nel soppesare i pro e contro di un'opzione. Per esempio, il desiderio dell'attentatore di un cioccolatino durante le manovre militari interessa soltanto potenzialmente la sua condotta: sebbene questi possa adottare come intenzione l'acquisto di un cioccolatino, verosimilmente egli non agirà in conseguenza di questo suo desiderio, stabilendo come intenzioni delle motivazioni più rilevanti per orientare l'azione. Questa considerazione rafforza l'idea del ruolo fondamentale dell'intenzione nella produzione di azioni, tanto da indurre autori come Georgeff e Rao (1998) ad affermare che l'intenzione rappresenta lo stato *deliberativo* e non semplicemente quello motivazionale di un sistema.

Da queste due osservazioni, segue che le intenzioni sono il risultato di una *scelta* dell'agente nel dominio dei suoi desideri. Il modello impone normativamente che la scelta dell'agente restituisca un insieme consistente di desideri e che l'agente *si impegni* a realizzarli. A partire da questa analisi, Cohen e Levesque (1990, p. 220) stabiliscono un concetto di intenzione che non può prescindere dall'interazione di questo con scelte, desideri ed impegno:

> *Intention is choice with commitment.* Intention will be modeled as a composite concept specifying what the agent has chosen and how the agent is committed to that choice.

E' però importante notare che un individuo non intende tutto ciò che sceglie. Si consideri quindi un agente che ha scelto di realizzare un desiderio e, così facendo, ha scelto di raggiungere un certo stato di cose: se questi è convinto che le proprie azioni determinino certi effetti, selezionando quell'intenzione, egli ha scelto anche le conseguenze dei propri gesti. L'agente sceglie la globalità di uno

tra gli scenari possibili. Tuttavia, egli non intende qualunque cosa in tale scenario, specie effetti indesiderati (ma previsti) oppure mezzi consapevolmente poco graditi. Nel primo esempio, se il terrorista intende colpire la scuola per raggiungere la vittoria, l'attentatore strategico non intende uccidere i bambini, ma giudica tale conseguenza un mero effetto collaterale di un piano più ampio. Come suggeriscono Cohen e Levesque (1990, p. 219), «*Expected side-effects are chosen, but not intended*». Bratman analizza il problema della distinzione tra effetti collaterali ed effetti intesi, chiamato *The Problem of the Package Deal*, affermando che effetti collaterali e intenzioni non hanno lo stesso ruolo nella pianificazione delle azioni di un agente: ad esempio, se l'individuo non dovesse realizzare gli effetti indesiderati presenti nello scenario che ha scelto, non tenterà di eseguire un nuovo piano per concretizzarli.

***The Asymmetry Thesis*: intenzioni e convinzioni.** Dal momento che l'agente BDI è collocato in un contesto e gli scopi che si propone sono da questo dipendenti, è essenziale che l'individuo abbia delle informazioni sullo stato del mondo. In altre parole, il sistema è situato in un ambiente da cui acquisisce dei dati (input) per produrre azioni (output) che si ripercuotono sull'ambiente stesso. Data la caratteristica limitazione di risorse di un agente cognitivamente plausibile, le sue informazioni riguardo all'ambiente non saranno certezze, ma saranno soltanto convinzioni parziali, verosimili e orientative. Un aspetto fondativo del modello BDI, in cui si è visto che le intenzioni fungono da filtro di ammissibilità per la selezione di altre intenzioni, sarà quindi la relazione tra convinzioni ed intenzioni. L'analisi di Bratman (1987, p. 37), parte da questa idea centrale:

> There is a defeasible demand that one's intentions be consistent with one's beliefs. Violation of this demand is, other things equal, a form of criticizable irrationality.

Questo principio, eventualmente contraddetto da dati empirici sfavorevoli (*defeasible*), stabilisce normativamente un vincolo di consistenza tra intenzioni e convinzioni: esso fissa una regola a cui gli agenti devono uniformarsi per non essere caratterizzati come irrazionali ed esclusi dalla teoria. Tuttavia questa norma non è sufficiente a garantire che l'intenzione di un agente di a implichi la sua convinzione che a. Ad esempio, Phil intende fermarsi in libreria tornando a casa, ma, sapendo che mentre guida è spesso sovrappensiero, teme si dimenticherà di realizzare la sua intenzione e perciò non crede che si fermerà. Tuttavia, casi del genere non provano neppure che l'intenzione di compiere l'azione a non richieda la convinzione che a sia vera. Per questo motivo, Bratman non assume né che l'intenzione di compiere l'azione a implichi la convinzione che a sia vera, né la negazione di questa proposizione. Ciò che invece presenta è un'analisi più accurata della questione, l'*Asymmetry Thesis*, che consta di due argomen-

ti. Con il primo Bratman (1987, p. 38) *ammette l'incompletezza di intenzioni e convinzioni*, con il secondo *respinge l'inconsistenza tra intenzioni e convinzioni*:

> [1.] An intention to *a* normally provides the agent with support for a belief that he will *a*. But there need be no irrationality in intending to *a* and yet still not believing one will.
> [2.] In contrast, there will normally be irrationality in intending to *a* and believing one will not *a*; for there is a defeasible demand that one's intentions be consistent with one's beliefs.

Supponiamo che l'attentatore intenda indebolire il nemico, ma che non creda all'efficacia dei suoi tentativi, perché l'avversario è molto potente: non è certo di fallire, ma neppure crede nella riuscita del suo intento. In questo caso, il terrorista intende nuocere al suo antagonista, ma non crede che riuscirà ad indebolirlo. Tuttavia l'attentatore sarebbe irrazionale se, data la sua intenzione, fosse convinto di non indebolire l'avversario: intendere qualcosa che si crede impossibile impedisce all'intenzione di svolgere la sua funzione di filtro di ammissibilità nei confronti delle selezioni successive. Il terrorista convinto dell'impossibilità della propria intenzione non procederà neppure al ragionamento mezzi-fini per trovare una strategia adatta ad indebolire l'avversario, esibendo un comportamento irrazionale. Con l'*Asymmetry Thesis*, Bratman stabilisce che se un agente intende compiere *a*, allora: 1) l'agente crede che realizzare *a* sia possibile, 2) l'agente non crede che non riuscirà a realizzare *a*, 3) l'agente crede che realizzerà *a* sotto determinate circostanze.

3 Una formalizzazione logica: BDICTL*

Una prima formalizzazione logica delle nozioni di impegno ed intenzioni è data da Cohen e Levesque (1990), che adottano una struttura a mondi possibili in cui ciascun mondo è una struttura temporale lineare ed introducono le modalità di convinzione, scopo, scopo persistente e intenzione, analizzandone le relazioni. Un altro esempio di *famiglia* di logiche BDI è dato dal modello formulato da Georgeff e Rao (1998), BDICTL*, che combina una *logica temporale ramificata* (CTL*, Computational Tree Logic) con una *logica multi-modale* (dove gli operatori modali Bel, Des e Int rappresentano rispettivamente convinzioni, desideri e intenzioni di agenti). La semantica delle modalità BDI è data dalle strutture di Kripke. Inoltre si assume che i mondi stessi siano strutture temporali ramificate: ciascun mondo può essere visto come una struttura di Kripke per una logica temporale ramificata. Wooldridge (2000) estende il modello BDICTL* per definire $\mathcal{LORA}$ ($\mathcal{L}$ogic $\mathcal{O}$f $\mathcal{R}$ational $\mathcal{A}$gents), nel cui linguaggio confluiscono il modello BDI, la logica classica del prim'ordine, la logica temporale ramificata e una logica dell'azione. Presento e discuto ora il modello BDICTL*, seguendo le espo-

$(\mathsf{Bel}_i\varphi)$	L'agente i crede φ
$(\mathsf{Des}_i\varphi)$	L'agente i desidera φ
$(\mathsf{Int}_i\varphi)$	L'agente i intende φ
$\bigcirc\varphi$	φ è soddisfatta nel punto temporale successivo
$\Diamond\varphi$	φ è soddisfatta "adesso" o in un punto temporale successivo
$\Box\varphi$	φ è sempre soddisfatta
$\varphi\mathcal{U}\chi$	φ è soddisfatta fino a quando χ è soddisfatta
$\varphi\mathcal{W}\chi$	φ è soddisfatta a meno che χ è soddisfatta
$\mathbf{A}\varphi$	φ è soddisfatta in ogni cammino
$\mathbf{E}\varphi$	φ è soddisfatta in qualche cammino

Tabella 1: Denotazione di alcuni elementi dell'alfabeto di BDICTL*

sizioni di Georgeff e Rao (1998), Wooldridge (2000) e Van der Hoek e Wooldridge (2003).

3.1 Linguaggio e semantica

L'alfabeto di BDICTL* (Cfr. Tabella 1) è costituito da un insieme non vuoto Φ di lettere proposizionali; dai connettivi proprosizionali $\wedge$ e $\neg$; dagli operatori modali Bel, Des e Int, per convinzioni, desideri ed intenzioni degli agenti; dai connettivi temporali $\bigcirc$, $\Diamond$, $\Box$, $\mathcal{U}$, $\mathcal{W}$, **A** e **E**; e da variabili e costanti individuali per rappresentare gli agenti. Ci sono due tipi di formule ben formate: le *formule di stato*, che sono vere in determinati mondi in determinati punti temporali e le *formule di cammino*, che sono vere in determinati mondi lungo determinati cammini (ossia sequenze di transizioni di punti temporali). Le formule di stato sono definite ricorsivamente, per cui: ogni proposizione atomica φ è una formula di stato; se φ e χ sono formule di stato anche $\neg\varphi$ e $\varphi\wedge\chi$ sono formule di stato; se φ è una formula di cammino $\mathbf{A}\varphi$ e $\mathbf{E}\varphi$ sono formule di stato; se φ è una formula di stato allora $(\mathsf{Bel}_i\varphi)$, $(\mathsf{Des}_i\varphi)$, $(\mathsf{Int}_i\varphi)$ sono formule di stato, dove i è un termine (variabile o costante) che indica un agente. Anche le formule di cammino sono definite ricorsivamente, in questo modo: ogni formula di stato è anche una formula di cammino; se φ e χ sono formule di cammino allora anche $\neg\varphi$ e $\varphi\wedge\chi$ sono formule di cammino; se φ e χ sono formule di cammino allora anche $\bigcirc\varphi$, $\Diamond\varphi$, $\Box\varphi$, $\varphi\mathcal{U}\chi$, $\varphi\mathcal{W}\chi$ sono formule di cammino.
Una struttura di Kripke per BDICTL* è definita dalla settupla $\mathcal{M}$:

$$\mathcal{M} = \langle W, \{T_w : w \in W\}, \{R_w : w \in W\}, L, \mathcal{B}, \mathcal{D}, \mathcal{I}\rangle$$

- W è l'insieme dei mondi, T è l'insieme dei punti temporali, $R \subseteq T \times T$ è una relazione *totale* $(\forall t \in T, \exists t' | t' \in T\ e\ (t, t') \in R)$ che rappresenta tutte le possibili evoluzioni del sistema.

- Un mondo $w \in W$ su T e R è una coppia $\langle T_w, R_w \rangle$, dove $T_w \subseteq T$ e $R_w \subseteq R$. Si noti che nel modello in analisi, i mondi non sono stati istantanei, ma strutture temporali ramificate: l'intuizione è che tali strutture rappresentino l'incertezza di un agente non solo sullo stato presente del mondo, ma anche sulla possibile evoluzione di questo.

- L è una valutazione proposizionale classica per ciascun mondo $w \in W$ in ciascun punto temporale $t \in T$: $L\langle w, t\rangle : \Phi \longrightarrow \{0, 1\}$.

- $\mathcal{B}$ è una funzione che assegna ad ogni agente una *relazione di accessibilità per le convinzioni*, ovvero una relazione su mondi e punti temporali, come segue:

$$\mathcal{B}: Agenti \longrightarrow \wp(W \times T \times W)$$

  Con un abuso di linguaggio si dirà che $\mathcal{B}$ è una relazione di accessibilità per le convinzioni. Si scrive $\mathcal{B}_t^w(i)$ per indicare l'insieme dei mondi accessibili all'agente i a partire dal mondo w al tempo t. Dal momento che $\mathcal{B}$ dipende da un mondo w e un tempo t determinati, il risultato dell'applicazione di $\mathcal{B}$ su di una situazione differente può essere diverso: in questo modo si esprime il fatto che l'agente può cambiare le proprie convinzioni a proposito delle opzioni disponibili. Formalmente, $\mathcal{B}_t^w(i) = \{w' | \langle w, t, w'\rangle \in \mathcal{B}(i)\}$. $\mathcal{D}$ e $\mathcal{I}$ sono definite analogamente. Intuitivamente, i mondi $\mathcal{B}$-$\mathcal{D}$-$\mathcal{I}$-accessibili sono rispettivamente quelli che l'agente crede che siano possibili, che desidera realizzare e che intente concretizzare.

La soddisfacibilità di una formula è definita rispetto ad una struttura $\mathcal{M}$, un mondo w e un punto temporale t. L'espressione $\mathcal{M}, \langle w, t\rangle \vDash \varphi$ si legge "la struttura $\mathcal{M}$ nel mondo w e nel punto temporale t soddisfa φ". Un cammino $(t_0, t_1, \dots)$ in un mondo w si scrive $(\langle w, t_0\rangle, \langle w, t_1\rangle, \langle w, \dots\rangle)$.

Esempio 2* Formalizzo l'Esempio 2 relativamente al linguaggio e alle strutture semantiche fondamentali. Siano: *Sen*= concorrere per un seggio al Senato; *Sen(Win)*= vincere un seggio al Senato; *Rep*= mantenere il seggio alla Camera dei Rappresentanti; *Rit* = ritirarsi dalla politica; *Poll* = indire il sondaggio; *Poll(Yes)* = la maggioranza approva il passaggio al Senato.

Seguono alcuni esempi di formule espresse nel linguaggio BDICTL* con relativa interpretazione (l'indice P denota l'agente, Phil):

- $\mathsf{Bel}_P(Sen \mathcal{W} Rit)$: Phil crede di concorrere per un seggio al Senato a meno che non si ritiri dalla politica;

- $\mathsf{Bel}_P(\bigcirc Sen(Loss))$: Phil crede che in seguito non vincerà un seggio al Senato;

- $\mathsf{Bel}_P(Poll \rightarrow \mathbf{E}Poll(Yes))$: Phil crede che se indice il sondaggio allora è possibile che la maggioranza approvi il passaggio al Senato;
- $\mathsf{Bel}_P(\mathbf{E}Poll(Yes))$: Phil crede che sia possibile che la maggioranza approvi il passaggio al senato;
- $\mathsf{Bel}_P(\mathbf{A}\Diamond Rep)$: Phil crede che in ogni caso potrà mantenere il seggio alla Camera dei Rappresentati;
- $\neg\mathsf{Des}_P Rit$: Phil non desidera ritirarsi dalla politica;
- $\mathsf{Int}_P Poll$: Phil intende indire il sondaggio;
- $\neg\mathsf{Int}_P(\neg Poll)$: Phil non intende non indire il sondaggio.

Nella Tabella 2 è rappresentato l'insieme W degli otto mondi possibili. La Tabella 3 rappresenta invece le relazioni di accessibilità. La prima colonna mostra i quattro mondi $\mathcal{B}$-accessibili di Phil: questi corrispondono alla vittoria o meno del seggio al Senato sulla base dell'esito del sondaggio. Nella seconda colonna sono riportati i mondi $\mathcal{D}$-accessibili: si noti come l'opzione di ritirarsi dalla politica non sia presente nei mondi desiderati, ma soltanto in quelli creduti (Phil crede che ritirarsi dalla politica sia un'opzione, ma non la considera seriamente). Infine la terza colonna mostra i mondi $\mathcal{I}$-accessibili: questi sottomondi dei precedenti rappresentano la scelta di Phil e il suo impegno a realizzarla (Phil intende indire il sondaggio).

Seguono alcuni esempi di formule soddisfatte e di formule non soddisfatte:

- $\mathcal{M}, \langle w_3, t_0\rangle \vDash \mathbf{E}Rit$
  La struttura $\mathcal{M}$ nel mondo w_3 e nel punto temporale t_0 soddisfa $\mathbf{E}Rit$ perchè esiste un cammino $(\langle w_3, t_0\rangle, \langle w_3, t_1\rangle, \langle w_3, t_{13}\rangle)$ in cui la struttura $\mathcal{M}$ soddisfa Rit: $\mathcal{M}, (\langle w_3, t_0\rangle, \langle w_3, t_1\rangle, \langle w_3, t_{13}\rangle) \vDash Rit$.
- $\mathcal{M}, \langle w_3, t_0\rangle \nvDash \mathbf{A}Rit$
  La struttura $\mathcal{M}$ nel mondo w_3 e nel punto temporale t_0 non soddisfa $\mathbf{A}Rit$ perchè non è vero che in ogni cammino del mondo la struttura soddisfa Rit: ad esempio, $\mathcal{M}, (\langle w_3, t_0\rangle, \langle w_3, t_1\rangle, \langle w_3, t_2\rangle) \nvDash Rit$.
- $\mathcal{M}, \langle w_3, t_0\rangle \vDash \mathsf{Bel}_P\mathbf{E}Rit$
  La struttura $\mathcal{M}$ nel mondo w_3 e nel punto temporale t_0 soddisfa $\mathsf{Bel}_P\mathbf{E}Rit$ perchè in ogni mondo $\mathcal{B}$-accessibile a partire dal mondo w_3 e dal tempo t_0 la struttura soddisfa $\mathbf{E}Rit$; in altre parole, in ogni mondo che Phil nella situazione $\langle w_3, t_0\rangle$ crede che sia possibile esiste un cammino in cui la struttura soddisfa Rit. Formalmente, $\forall v \in \mathcal{B}^{w_3}_{t_0}$, $\mathcal{M}, \langle v, t_0\rangle \vDash \mathbf{E}Rit$.

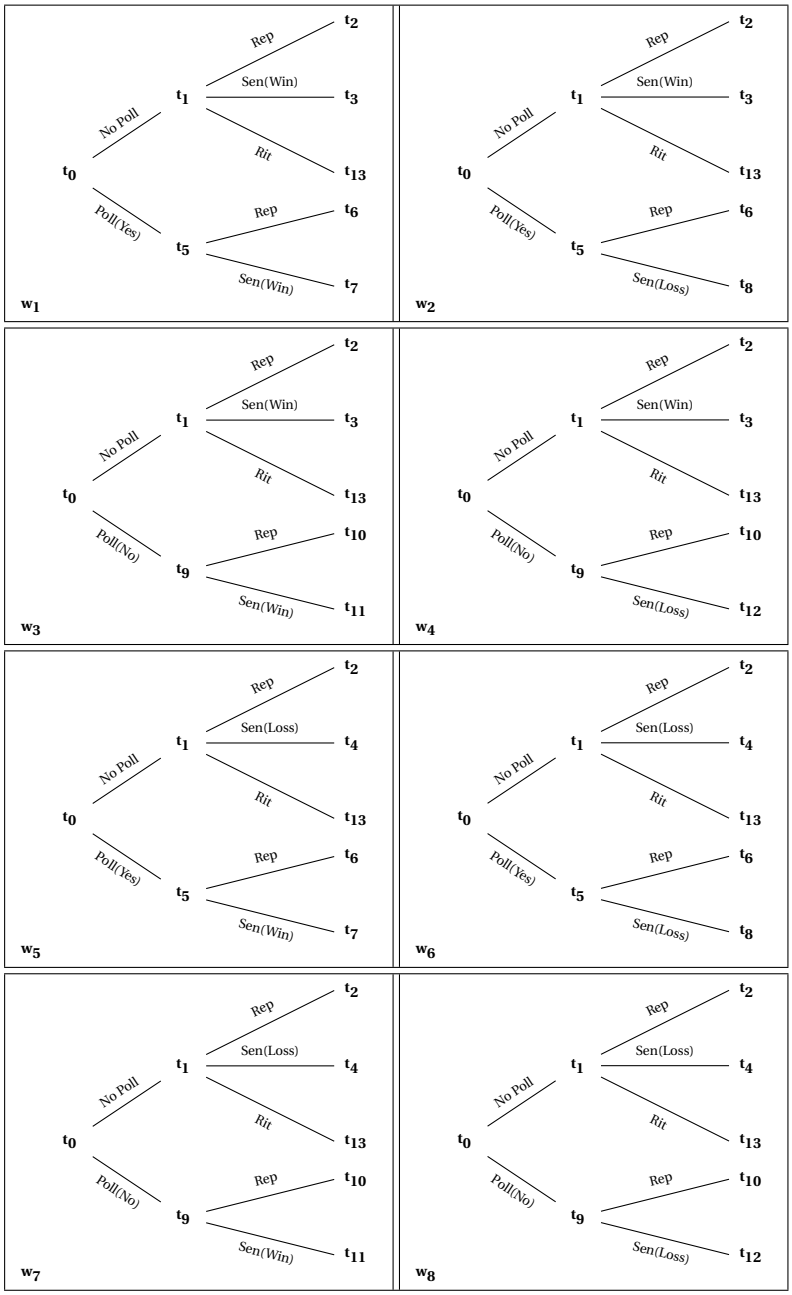


Tabella 2: L'insieme W dei mondi possibili dell'Esempio 2*.

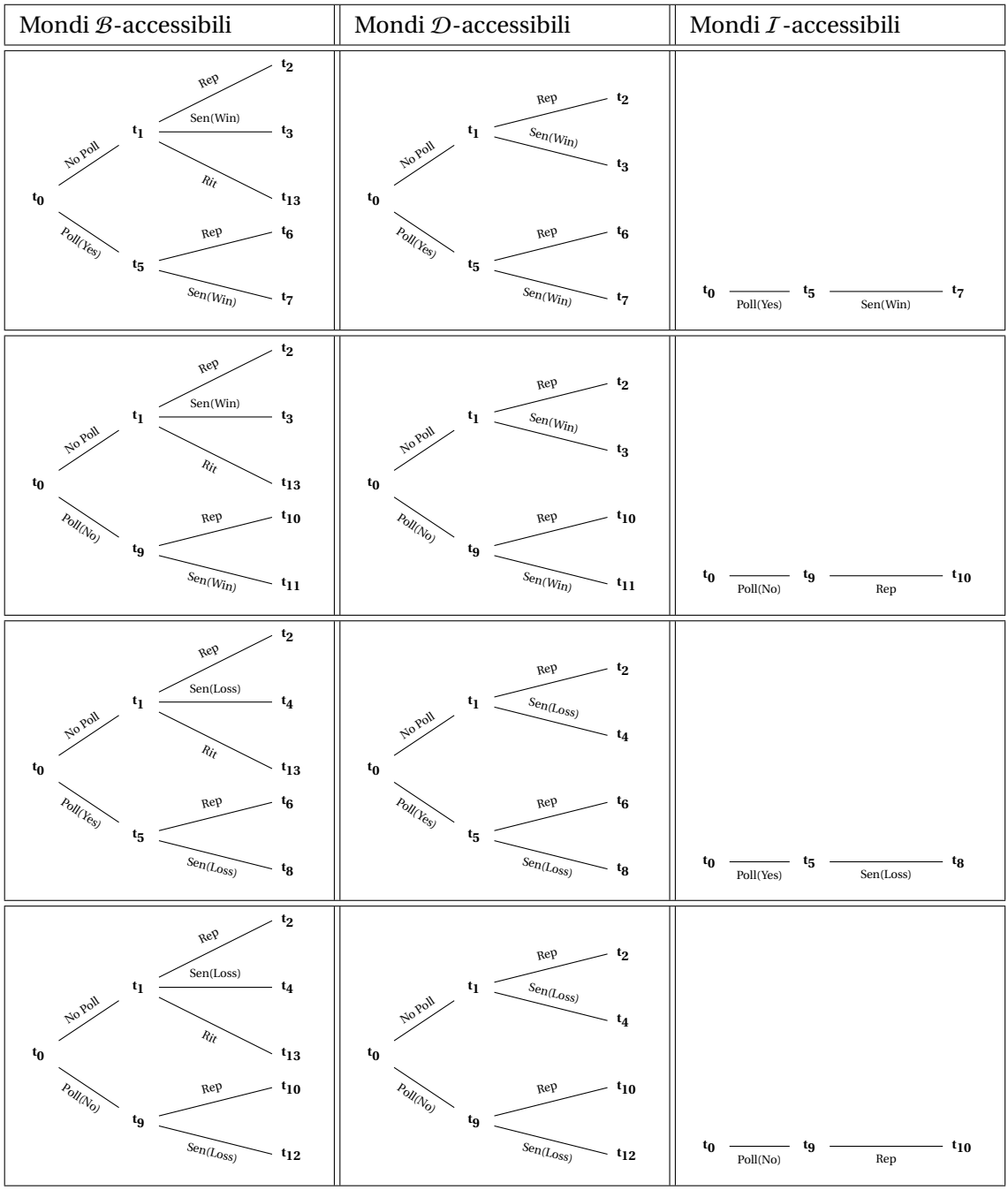


Tabella 3: I mondi $\mathcal{B}$-$\mathcal{D}$-$\mathcal{I}$-accessibili dell'Esempio 2*.

3.2 Assiomatizzazione

Assiomi modali. Presento e discuto gli assiomi modali della famiglia di logiche BDICTL* e, tramite la teoria della corrispondenza, le condizioni imposte sulle relazioni di accessibilità. Gli assiomi modali, che considerano individualmente ciascuna delle tre modalità, esprimono alcune tra le norme di razionalità analizzate dall'indagine filosofica.

- **C**$_B$: Se $w' \in \mathcal{B}_t^w(i)$, allora $t \in w$ e $t \in w'$
  Gen$_B$: Se $\vdash \varphi$, allora $\vdash (\mathsf{Bel}_i\varphi)$
  K$_B$: $(\mathsf{Bel}_i\varphi) \wedge (\mathsf{Bel}_i(\varphi \rightarrow \chi)) \rightarrow (\mathsf{Bel}_i\chi)$
  D$_B$: $(\mathsf{Bel}_i\varphi) \rightarrow \neg(\mathsf{Bel}_i(\neg\varphi))$
  4$_B$: $(\mathsf{Bel}_i\varphi) \rightarrow \mathsf{Bel}_i(\mathsf{Bel}_i\varphi)$
  5$_B$: $\neg(\mathsf{Bel}_i\varphi) \rightarrow \mathsf{Bel}_i(\neg\mathsf{Bel}_i\varphi)$
  Questi sei assiomi equivalgono ad imporre che la semantica dell'operatore di convinzione corrisponda al sistema modale KD45. Infatti:
  D$_B \Leftrightarrow \mathcal{B}$ è *seriale* (per ogni $\langle w, t\rangle$, esiste un w' tale che $w' \in \mathcal{B}_t^w(i)$)
  4$_B \Leftrightarrow \mathcal{B}$ è *transitiva* (se $w' \in \mathcal{B}_t^w(i)$ e $w'' \in \mathcal{B}_t^{w'}(i)$, allora $w'' \in \mathcal{B}_t^w(i)$)
  5$_B \Leftrightarrow \mathcal{B}$ è *euclidea* (se $w' \in \mathcal{B}_t^w(i)$ e $w'' \in \mathcal{B}_t^w(i)$, allora $w' \in \mathcal{B}_t^{w''}(i)$).

- **C**$_D$: Se $w' \in \mathcal{D}_t^w(i)$, allora $t \in w$ e $t \in w'$
  Gen$_D$: Se $\vdash \varphi$, allora $\vdash (\mathsf{Des}_i\varphi)$
  K$_D$: $(\mathsf{Des}_i\varphi) \wedge (\mathsf{Des}_i(\varphi \rightarrow \chi)) \rightarrow (\mathsf{Des}_i\chi)$
  D$_D$: $(\mathsf{Des}_i\varphi) \rightarrow \neg(\mathsf{Des}_i(\neg\varphi))$.
  Questi quattro assiomi equivalgono ad imporre che la semantica dell'operatore di desiderio corrisponda al sistema modale KD. Infatti:
  D$_D \Leftrightarrow \mathcal{D}$ è *seriale* (per ogni $\langle w, t\rangle$, esiste un w' tale che $w' \in \mathcal{D}_t^w(i)$).

- **C**$_I$: Se $w' \in \mathcal{I}_t^w(i)$, allora $t \in w$ e $t \in w'$
  Gen$_I$: Se $\vdash \varphi$, allora $\vdash (\mathsf{Int}_i\varphi)$
  K$_I$: $(\mathsf{Int}_i\varphi) \wedge (\mathsf{Int}_i(\varphi \rightarrow \chi)) \rightarrow (\mathsf{Int}_i\chi)$
  D$_I$: $(\mathsf{Int}_i\varphi) \rightarrow \neg(\mathsf{Int}_i(\neg\varphi))$.
  Questi quattro assiomi equivalgono ad imporre che la semantica dell'operatore di intenzione corrisponda al sistema modale KD. Infatti:
  D$_I \Leftrightarrow \mathcal{I}$ è *seriale* (per ogni $\langle w, t\rangle$, esiste un w' tale che $w' \in \mathcal{I}_t^w(i)$).

C, noto come assioma di compatibilità spazio-temporale, impone che se un mondo w' è accessibile per un agente a partire dalla situazione $\langle w, t\rangle$, allora t è un punto temporale sia in w che in w'. **C** è necessario poiché i mondi sono strutture temporali e le relazioni di accessibilità dipendono da una situazione determinata. **Gen**, nota anche come regola di necessitazione o di generalizzazione, impone che ogni formula valida sia creduta, desiderata o intesa, mentre **K**, detto assioma distributivo, è richiesto per ogni minimo sistema modale. **D** pone un vincolo di consistenza: come si è visto, convinzioni ed intenzioni di un

agente razionale devono essere consistenti e non contraddittorie. Infatti se l'attentatore crede di indebolire il nemico, allora non crede di non indebolirlo e se intende sganciare una bomba non intende non sganciare una bomba. Consideriamo i desideri: mentre dall'analisi filosofica era emerso che i desideri possono essere inconsistenti, $\mathbf{D}_D$ impone che questi siano logicamente consistenti, cioè che se un agente desidera qualcosa allora non può desiderare la sua negazione. Il dissenso tra principi filosofici e formalizzazione logica è solo apparente: data la struttura temporale ramificata, desideri contrastanti possono portare l'agente lungo diversi cammini che non possono essere percorsi insieme. Nonostante i desideri siano logicamente consistenti, questi possono non essere tutti realizzabili, dal momento che un agente può eseguire soltanto un cammino tra le strutture ramificate delle esecuzioni possibili. Infine, $\mathbf{4}_B$ e $\mathbf{5}_B$ sono gli assiomi di introspezione positiva e negativa: insieme stabiliscono che, mentre un agente può avere soltanto convinzioni imperfette su ciò che crede vero nel mondo, egli ha convinzioni perfette riguardo le proprie convinzioni.

Assiomi intermodali. Come l'analisi filosofica studia le relazioni tra convinzioni, desideri e intenzioni, così la formalizzazione studia i nessi tra le tre modalità rappresentandoli come assiomi intermodali, generati cioè dai rapporti tra le relazioni di accessibilità. Georgeff e Rao (1998) individuano gli assiomi intermodali tramite un'analisi puramente combinatoria di rapporti, quali quelli di sottoinsieme o intersezione, tra le relazioni di accessibilità: ciascuna logica appartenente alla famiglia BDICTL* differisce dalle altre proprio per la selezione degli assiomi intermodali, i quali corrispondono ad una relazione tra $\mathcal{B}$, $\mathcal{D}$, $\mathcal{I}$. Per questo motivo, diversi tra questi assiomi intermodali non esprimono affatto i principi normativi del modello BDI e di conseguenza molti tra sistemi esposti dagli autori non formalizzano questo modello di razionalità. In ciò che segue, discuto gli assiomi intermodali della logica BDICTL*-W3, indicando come questi esprimano o permettano di derivare i principi normativi del modello.

I tre assiomi intermodali di BDICTL*-W3 equivalgono ad imporre che le seguenti intersezione tra le relazioni di accessibilità siano non vuote:

- **Ass1** : $(\mathsf{Des}_i\,\chi \Rightarrow \neg\mathsf{Int}_i\neg\chi) \Leftrightarrow \mathcal{I}_t^w(i) \cap \mathcal{D}_t^w(i) \neq \emptyset$
- **Ass2** : $(\mathsf{Bel}_i\,\chi \Rightarrow \neg\mathsf{Des}_i\neg\chi) \Leftrightarrow \mathcal{D}_t^w(i) \cap \mathcal{B}_t^w(i) \neq \emptyset$
- **Ass3** : $(\mathsf{Bel}_i\,\chi \Rightarrow \neg\mathsf{Int}_i\neg\chi) \Leftrightarrow \mathcal{I}_t^w(i) \cap \mathcal{B}_t^w(i) \neq \emptyset$

Ass1, noto come assioma di consistenza di desideri e intenzioni, affermando che un agente non intende la negazione di ciò che desidera, è in accordo con il principio per cui le intenzioni sono un sottoinsieme consistente dei desideri: la scelta dell'agente rispetta le proprie attitudini motivazionali. **Ass2** richiede che l'agente non desideri ciò che crede impossibile: questo vincolo impone

che l'individuo non abbia desideri irrealizzabili, cioè che sia concreto. Come si è visto per l'assioma modale $\mathbf{D}_D$, anche **Ass2** non esclude che l'agente abbia desideri che lo portino lungo cammini divergenti e quindi non possa realizzarli tutti. Da **Ass3** discendono due importanti proprietà della teoria di Bratman: l'*Asymmetry Thesis,* di cui **Ass3** è la contrappositiva, e la soluzione a *The Problem of the Package Deal.* Si può dimostrare (Georgeff e Rao 1998) come dagli assiomi di BDICTL*-W3 discendono entrambi gli argomenti, così formalizzati:

- *The Asymmetry Thesis*:
  - **AT1**, *Consistenza di intenzioni e desideri*:
    $\vDash (\mathsf{Int}_i\varphi) \Rightarrow (\neg\mathsf{Bel}_i\neg\varphi) \;\Leftrightarrow \nvDash (\mathsf{Int}_i\varphi) \wedge (\mathsf{Bel}_i\neg\varphi)$
  - **AT2**, *Incompletezza di intenzioni e desideri*:
    $\nvDash (\mathsf{Int}_i\varphi) \Rightarrow (\mathsf{Bel}_i\varphi) \;\Leftrightarrow \vDash (\mathsf{Int}_i\varphi) \wedge (\neg\mathsf{Bel}_i\varphi)$
  - **AT3**, *Incompletezza di desideri e intenzioni*:
    $\nvDash (\mathsf{Bel}_i\varphi) \Rightarrow (\mathsf{Int}_i\varphi) \;\Leftrightarrow \vDash (\mathsf{Bel}_i\varphi) \wedge (\neg\mathsf{Int}_i\varphi)$
- *Solution to The Problem of the Package Deal*:
  - **PD**: $\vDash (\mathsf{Int}_i\varphi) \wedge (\mathsf{Bel}_i(\varphi \rightarrow \chi)) \wedge (\neg\mathsf{Int}_i\chi)$
    $\Leftrightarrow \nvDash (\mathsf{Int}_i\varphi \wedge \mathsf{Bel}_i(\varphi \rightarrow \chi)) \;\Rightarrow\; (\mathsf{Int}_i\chi)$

4 Un esempio di *Agent Control Loop*

La teoria BDI è un modello interdisciplinare, dove i contributi forniti da diverse aree di ricerca interagiscono apportando prospettive differenti sullo stesso problema. Si considera di seguito la componente informatica di programmazione di sistemi BDI capaci di agire in ambienti dinamici: nello specifico, esamino un esempio di *Agent Control Loop* (Wooldridge 2000), cioè di un segmento di processo operato da un sistema. Una delle ragioni che giustificano i formalismi che seguono è la possibilità di numerose implementazioni: le applicazioni nel mondo reale spaziano infatti dagli assistenti automatici in rete al controllo del traffico aereo e alla gestione delle telecomunicazioni. La plausibilità cognitiva del modello BDI è confermata anche dalla costruzione di sistemi che, programmati secondo queste procedure, trovano numerosi impieghi nella realtà. La descrizione algoritmica delle diverse procedure di ragionamento di un agente razionale mette perciò in rilievo l'interesse pratico computazionale della teoria BDI.

Notazioni. Siano *Bel* l'insieme di tutte le convinzioni e *B* l'insieme delle convinzioni correnti di un agente; analogamente, siano *Des* l'insieme di tutti i desideri e *D* l'insieme dei desideri correnti di un agente. Infine, siano *Int* l'insieme di tutte le intenzioni possibili e *I* l'insieme delle intenzioni correnti di un agente.

La *percezione* di un agente, ovvero le informazioni disponibili riguardo all'ambiente, è rappresentata da *impulsi percettivi* o *percepiti*. Si usa ρ', ρ, ρ_1, ... e *Per* per indicare, rispettivamente, gli impulsi percettivi e l'insieme di tutti i percepiti. Siano π', π, π_1, ... e *Plan*, rispettivamente, piani e l'insieme di tutti i piani. *execute*(π) è una procedura che prende come input un piano e lo esegue senza fermarsi; eseguire un piano significa eseguire ogni azione contenuta nel piano.

Procedure. A questo punto si possono discutere le formalizzazioni di tre tra le diverse procedure che portano un agente a compiere un'azione. Il *processo di aggiornamento delle convinzioni* è modellato dalla *funzione di revisione di convinzioni*, definita come:

$$bfr\colon \wp(Bel) \times Per \longrightarrow \wp(Bel)$$

A partire dalle convinzioni correnti e dagli impulsi percettivi, la funzione di revisione di convinzioni stabilisce un nuovo insieme di convinzioni. Anche in BDICTL* l'agente può aggiornare le proprie convinzioni: il ruolo dinamico di *bfr* è sostituito dalla struttura temporale ramificata e dalla definizione di $\mathcal{B}$, il cui risultato dipende da precise coordinate temporali espresse dalla situazione.

Il *processo deliberativo* di un agente è descritto dalla funzione,

$$deliberate\colon \wp(Bel) \longrightarrow \wp(Int)$$

che da un insieme di convinzioni, restituisce un insieme di intenzioni, quelle che l'agente vuole realizzare sulla base delle proprie convinzioni. Il processo di deliberazione è costituito da due fasi: nella prima, che chiamiamo *generazione di opzioni*, l'agente cerca di capire quali siano le opzioni disponibili; nella seconda, che indichiamo come *filtraggio*, l'agente sceglie una o più tra le opzioni appena selezionate, e si impegna a realizzarle. Formalmente:

$$options\colon \wp(Bel) \times \wp(Int) \longrightarrow \wp(Des)$$
$$filter\colon \wp(Bel) \times \wp(Des) \times \wp(Int) \longrightarrow \wp(Int)$$

La funzione *options* a partire da convinzioni e intenzioni correnti di un agente, determina un insieme di opzioni, ovvero stabilisce un insieme di possibilità che, date le convinzioni sul mondo, si configura come adatto a raggiungere le proprie intenzioni. Queste opzioni saranno chiamate desideri, per sottolineare l'interpretazione intuitiva di un desiderio secondo cui, in un mondo ideale, un agente vorrebbe che tutti i propri desideri fossero realizzati. Tuttavia è possibile che l'agente non sia grado di realizzare tutti i propri desideri, questo perchè spesso i desideri sono mutualmente esclusivi: perciò l'agente deve scegliere e la funzione *filter* rappresenta la selezione di un'opzione, quella cioè che l'agente si impegna a realizzare. La funzione *filter* interpreta il ruolo centrale delle intenzioni nel modello BDI emerso dall'analisi filosofica: queste condizionano e

fungono da filtro di ammissibilità per le scelte successive dell'agente. Infatti *filter* esprime la relazione tra desideri, intenzioni e scelte secondo cui l'intenzione sarebbe ciò che è stato scelto con l'impegno per la sua realizzazione.

Il *ragionamento mezzi-fini* di un agente è rappresentato dalla funzione:

$$plan\colon \wp(Bel) \times \wp(Int) \longrightarrow Plan$$

che sulla base delle convinzioni e intenzioni correnti, seleziona il piano opportuno. Si vedrà, nell'*Agent Control Loop*, come *plan* segua *filter*: l'idea è la stessa esplicitata dal ragionamento di Bratman intorno ai piani come gerarchizzati, grazie ai quali quelli che mirano a fini più ampi condizionano quelli sui mezzi.

Agent Control Loop. Alla luce delle considerazioni precedenti esamino un esempio di *Agent Control Loop* (Wooldridge 2000, p. 31), cioè un fragmento di un processo operativo.

```
Algorithm:  Agent Control Loop
1.
2.    B := B0;
3.    I := I0;
4.    while true do
5.        get next percept ρ;
6.        B := bfr(B, ρ);
7.        D := options(B, I);
8.        I := filter(B, D, I);
9.        π := plan(B, I);
10.       execute(π);
11.   end-while
```

A partire da due insiemi rispettivamente di convinzioni (2) ed intenzioni iniziali (3), l'agente esamina l'ambiente in cui è immerso, tramite i sensi di cui è fornito, ricavandone un'osservazione (5). Da questo impulso percettivo è indotto a compiere una revisione delle proprie convinzioni rispetto al mondo: l'esito di questo processo può essere un insieme differente di informazioni sul contesto (6). A partire da queste nuove convinzioni, l'agente avvia il processo di deliberazione: innanzitutto vaglia le opzioni disponibili (7) e in un secondo istante decide quali di queste impegnarsi a realizzare (8). Quindi ragiona circa i mezzi per realizzare l'intenzione selezionata, stabilendo un piano tra quelli di cui dispone (9), ed esegue il piano (10).

5 Conclusione

Si è visto come il modello BDI, per fornire una caratterizzazione cognitivamente plausibile delle azioni degli agenti, riduca il grado di idealizzazione sugli individui proprio della teoria della scelta razionale. Questo passaggio determina due conseguenze. In primo luogo l'ontologia del modello, per migliorare il proprio potere espressivo, inserisce il concetto di intenzione e ne analizza i rapporti con desideri e convinzioni. In secondo luogo l'approccio epistemologico del modello BDI presenta alcuni aspetti descrittivamente plausibili, come l'inconsistenza dei desideri, insieme ad altri di natura normativa, come i vincoli di consistenza sulle intenzioni.

L'analisi di questa duplice natura epistemologica della teoria è rafforzata dai contributi emersi dalle due formalizzazioni. Da un lato, infatti, il sistema logico che ho presentato sviluppa la componente normativa della teoria, già articolata nei termini di vincoli di consistenza sugli elementi dell'ontologia. I requisiti normativi per la razionalità dell'agente sono espressi dagli assiomi modali e intermodali: in altre parole, l'assiomatizzazione logica rappresenta e definisce le relazioni tra le norme di comportamento individuate dall'analisi filosofica. Dall'altro lato, l'implementazione del modello, attraverso una descrizione algoritmica della procedure di ragionamento di un agente razionale, ne dichiara l'interesse pratico computazionale e ne approfondisce gli aspetti di plausibilità cognitiva.

Si può quindi concludere affermando che logica teorica e implementazione reale approfondiscono i diversi aspetti dei principi filosofici del modello BDI: l'interazione e la pluralità di metodi di analisi potenzia la fecondità teoretica di questa concezione della razionalità.

Riferimenti bibliografici

Bratman, Michael E. (1987). *Intention, Plans, and Practical Reason*. Cambridge MA: Harvard University Press.

— (1990). "What is intention?" In: *Intentions in Communication*. Cambridge MA: The MIT Press, pp. 15–31.

Cohen, Philip R. e Hector J. Levesque (1990). "Intention is Choice with Commitment". In: *Artificial Intelligence* 42, pp. 213–261–339.

Georgeff, Michael P. e Anand S. Rao (1998). "Decision Procedures for BDI Logics". In: *Journal of Logic and Computation* 8.3, pp. 293–344.

Van der Hoek, Wiebe e Michael Wooldridge (2003). "Towards a Logic of Rational Agency". In: *Logic Journal of the IGPL* 11.2, pp. 135–159.

Wooldridge, Michael (2000). *Reasoning about Rational Agents*. Cambridge (MA): The MIT Press.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Rivista Italiana di Filosofia Analitica Junior 9:2 (2018)

Some proposals for the set-theoretic foundations of category theory

Lorenzo Malatesta

Abstract. The problem of finding proper set-theoretic foundations for category theory has challenged mathematician since the very beginning. In this paper we give an analysis of some of the standard approaches that have been proposed in the past 70 years. By means of the central notions of class and universe we suggest a possible conceptual recasting of these proposals. We focus on the intended semantics for the (problematic) notion of large category in each proposed foundation. Following Feferman (2006) we give a comparison and evaluation of their expressive power.

Keywords. Category Theory, Set Theory, Foundations..

Copyright. 2018 Lorenzo Malatesta. Published in Italy. Some rights reserved. First published on *Rivista Italiana di Filosofia Analitica Junior* 2:2 (2011), *Philosophy of Mathematics*, pp. 41-58.
Author. Lorenzo Malatesta, lorenzo.malatesta@gmail.com.

1 A problem of size

> [...] Thus, category theory is not just another field whose set-theoretic foundation can be left as an exercise. An interaction between category theory and set theory arises because there is a real question: what is the appropiate set-theoretic foundation of category theory?
>
> Andreas Blass[1]

It is common to date the birth of category theory to the publication of Eilenberg and Mac Lane's paper,[2] *A general theory of natural equivalences*. Already in this pioneering work, we can find a first analysis of some foundational issues concerning the raising theory. In fact Eilenberg and Mac Lane dedicate an entire paragraph to discuss some foundational problems of the set-theoretical interpretation of their theory. Here is the beginning of this paragraph:[3]

> We remarked in §3 that such examples as the "category of all sets", the "category of all groups" are illegitimate. The difficulties and antinomies here involved are exactly those of ordinary intuitive Mengenlehre; no essentially new paradoxes are apparently involved. Any rigorous foundation capable of supporting the ordinary theory of classes would equally well support our theory. Hence we have chosen to adopt the intuitive standpoint, leaving the reader free whatever type of logical foundation (or absence thereof) he may prefer.

The two authors immediately recognise the peculiarity of the constructions involved in their theory and offer a first simple diagnosis: since there is nothing new under the sun, just old well-known paradoxes, it is sufficient to give back these issues to the field they belong, i.e. set theory. Despite the apparent haste to dismiss the matter, what follows the above mentioned paragraph can be seen as the first concrete attempt to solve the problem: after having discussed some technical issues, the two mathematicians suggest a possible development of category theory inside the framework of the theory of sets and classes in the style of von Neumann, Bernays and Gödel's set theory (NBG). Before entering into the details of this and other proposals, it is important to focus on what is the problem. A good starting point is given by a critical analysis of the role played by the notion of **size** in category theory. Indeed, with the exception of set theory, it is difficult to find other mathematical fields where the notion of size plays such

[2] S. Eilenberg (1945).
[3] S. Eilenberg (1945), p. 246.

a central role. On the other hand, in category theory the distinction between **small categories** and **large categories** represents an important and inescapable dichotomy raised at the very beginning of any reasonable introduction to the subject. Nevertheless it is usual to get rid of this question as soon as possible and the working mathematician who uses category theory is therefore reluctant to deepen the analysis of the foundations of the theory. The following dialogue[4] is intended to parody this situation:

Dialogue 1.

TORTOISE: Hi Achilles, how are you? You have disappeared for a while, what have you been up to?

ACHILLES: My dear little Tortoise, you won't believe it, but I started studying some *abstract nonsense*. And, let me say that I found in it much more sense than is usually said.

TORTOISE: Good Achilles, I see you are not losing the habit to challenge your mind. I also have tried to give meaning to that bunch of arrows some time ago…now, I can just remember the definition of a category. Let me take the opportunity to ask you something that has bothered me since that time. Can you tell me what people mean with the term *large category*?

ACHILLES: Oh, my sweet little Tortoise, I know what you are driving at…you want to cheat me with the old story of the barber undecided if he shaves himself or not…this time I won't fell for it. The matter is simple: a large category is one whose collection of morphisms is a *proper class*.

TORTOISE: Then, let me bother you with my usual reasoning. The natural question to pose now is: what do you mean by proper class?

ACHILLES: Well, I'll be polite and I won't escape your innocent inquisition. I will call a proper class a collection which is not a set.

TORTOISE: It's not exactly a definition, but I'll give you that. I believe you already know what I am going to ask next…

ACHILLES: Let's see. Usually you don't have so much imagination. The only new term I introduced in our dialogue is set. I hope you don't want to ask me what is a set…

[4]The characters of this invented dialogue have been inspired by the dialogues in Hofstadter (1979).

TORTOISE: Exactly Achilles: less fantasy and more pedantry is the recipe of my philosophy...

ACHILLES: Ok. Let me surprise you. I have a new definition: a set is an object of the category Set, whose object are sets and whose morphisms are functions.

TORTOISE: Mmh..., you are right Achilles, you always surprise me...I am afraid you lost your way in an abstract nonsense...

Clearly positions like Achilles' one are unsatisfying from every possible point of view: mathematical, logical and philosophical. A proper category theorist, probably, would have preferred to answer Tortoise's question, "what is a set", saying "it's an object of a well-pointed topos with a natural number object and which satisfies the axiom of choice". Since this answer costs much more effort than trying to understand the problem, it is important to clarify what we mean by the problem of set-theoretic foundations of category theory, in such a way that also Achilles can understand why his position is not defensible.

It is an empirical fact that, to a great extent, mathematics can be formalized in set theory: a rather common choice for this set-theoretic "codification" is represented by the axioms of Zermelo Fraenkel's set theory with the axiom of choice (ZFC). For example, we can imagine to present group theory, algebraic topology or functional anaysis with the language only of set theory: objects of these theories can be described as sets whose properties can be derived from set-theoretic axioms. Following Blass,[5] it is therefore natural to ask in what sense category theory is an exception to this phenomenon. Why can't we leave this codification as a routine exercise?

As we have already observed, at the root of category theory lies the important *small/large distinction*. When doing category theory some of the objects and constructions that we deal with are (and have to be) essentially large. One of the first problem we meet if we regard this object from a set- theoretic perspective is to find an adequate encoding for large categories such as the category of all sets (*Set*) or the category of all groups (*Grp*). These categories are built having in mind essentially large collections and cannot be treated simply as sets.[6] This is not the only problem. The following list resumes some of the main issues that are essential to develop category theory.[7] In every reasonable foundational framework[8] we want to be able to:

[5] See the quotation at the beginning of this section.

[6] The argument is well known. A possible way to present it is the following: if the collection of all sets, V, was a set, then the collection of all its subsets, $\wp V$, would be a set included in V, contradicting Cantor's theorem.

[7] Compare with Feferman (2006) pp. 2–3.

[8] We use framework as synonym of metatheory or foundational system.

(A) form the category of every structure of a given type. Some elementary examples are: *Set*, *Grp* and *Top*;

(B) perform some basic set-theoretic constructions over an arbitrary category;

(C) form the category of all the functors between two arbitrary categories.

If we are specifically interested in set-theoretic foundations for category theory we would also like to be able to

(D) decide the consistency of these systems with respect to some accepted system of set theory.

It is worth mentioning that, beyond the concept of "large category" (requirement A), there are several different notions that rely on the same concept (*locally small category*, *small limits*, etc.). The frameworks should be expressive enough to make sense of each of these.

As we will see the choice of a specific foundational system will affect substantially the fulfilment of these requirements.

The next section gives an overview of the foundational proposals that we will consider in the rest of the paper.

2 Set-theoretic and other proposals: a retrospective.

As already noted, debates about foundations of category theory started with the very introduction of the notion of category. The rapid development of the theory and the ubiquity of categorical notions in different mathematical fields have brought these foundational issues to the attention of several mathematicians.

In the sequel we will consider some standard set-theoretic approaches to the problem of foundations of category theory. It is important to keep in mind that set theory is just *one* possible approach. Even among the set-theoretic frameworks, we won't be able to cover exhaustively all those proposed in the past, for example, Feferman's proposal to use Quine's set theory, *New Foundations*.[9] The question of what the *proper* set-theoretic foundation of category theory is can be misleading. We could argue that category theory, as any other mathematical subject, does not need any foundation either for its own internal development, or for understanding it. Nevertheless, once we have raised the question, we find that different solutions are at our disposal. As we will see none of the set-theoretical proposals we will consider definitely solve the problem. However, we stress that to a great extent each of these proposals is expressive enough to cover most of the cases of interest for the "working mathematician".

An important part in the debate of foundations of category theory that deserves a treatment on its own, is the possibility to regard category theory itself

[9]The interested reader should consult Feferman (2006).

as a foundational theory. The idea to consider category theory as a universal language capable of interpreting the entire mathematical edifice has been firstly proposed by Lawvere in the mid 60s. His research has led to a purely categorical description of the category *Set*. Nowadays, after his influential paper,[10] it is common to refer to these axioms with the acronym ETCS: Elementary Theory of the Category of Sets.

From a philosophical perspective, the project of Lawvere is intertwined with what has been called *categorical structuralism*.[11] As recent debates have shown, progress is impossible without a preliminarily agreed understanding of what is meant by the use of adjectives "structural" and "foundational" in this context.[12] Close to categorical structuralism, but not coinciding with Lawvere's position, is the idea to regard category theory as a foundation because of its *unifying character*. This position emerges for example in Marquis[13] and has recently been supported by some novel results discovered in topos theory.[14]

We finally mention a recent area of research that investigates set theory from a novel categorical perspective: Algebraic Set Theory (AST). The main goal of AST is to give a uniform categorical description for set-theoretical formal systems. Without addressing directly any foundational issues, AST focuses on bringing to light algebraic aspects of these systems by means of category theory.[15]

We can now focus on the organisation of the foundational proposals that we consider. The frameworks we will treat are the following:

- an approach internal to ZFC,
- NBG and MK,[16]
- Grothendieck's universes,
- Mac Lane's proposal,
- Feferman's proposal.

The first two set theories have in common the idea of using the notion of *class* to interpret the notion of *size* arising in category theory. The other three, instead, make use (in a more or less explicit way) of the notion of *universe* in order to better approximate the distinction *small/large*. Inspired by Shulman (2008), we suggest a possible recast of these proposals by means of these central notions.

[10] Lawvere (2005)
[11] See for example Awodey (1996), McLarty (2004).
[12] The interested reader should consult Hellman (2003) and Awodey (2004).
[13] See Marquis (2009).
[14] See Caramello (2010).
[15] A standard reference for AST is the book Joyal 1995. For a complete bibliography the reader should visit http://www.phil.cmu.edu/projects/ast/.
[16] MK is the acronym for Morse-Kelley set theory.

Before giving the details of these possible solutions we recall, in the next section, some specific examples of theorems "sensitive to the mathematical framework".

3 Examples

To give an idea of the ubiquity of notions of size in category theory we recall some basic results and definitions where these concepts play an important role[17]

Definition 1 (locally small category). *A category $\mathbb{C}$ is called locally small if, given two objects, a and b, the collection of morphisms between them, $Hom_{\mathbb{C}}(a, b)$, is small.*

If a category $\mathbb{C}$ is locally small then there exists the Hom-functor:

$$Hom_C : C^{op} \times \mathbb{C} \to Set.$$

Examples of locally small categories are: *Set*, *Grp* and in general all the categories built from "sets-with-structure". Given two locally small categories $\mathbb{C}$ and $\mathbb{D}$, the category of functors between them, $\mathbb{D}^{\mathbb{C}}$, is not in general, locally small.

A central notion in category theory is that of complete category: also in this case instances of the notion of size are explicitly involved.

Definition 2 (complete category). *A category $\mathbb{C}$ is said to be **complete** if every functor $F : J \to \mathbb{C}$, whose domain is a small category J, has limit.*

Examples of complete categories are: *Set*, *Grp*, *Rng*, *Comp Haus*. When the category is both small and complete, then it is just a preorder. Actually something stronger holds:

Theorem 2. *If a category $\mathbb{C}$ admits limits for any functor $F : \mathbb{D} \to \mathbb{C}$, with $\mathbb{D}$ any discrete category, then $\mathbb{C}$ is a preorder.*

For the proof see Borceaux (1994), proposition 2.7.1. This theorem explains why, in order to have a notion of *completeness* which makes sense for all categories, it is reasonable to ask for limits just for those functors whose domain is a small category J.

Another important theorem which is usually quoted when debating foundational issue in category theory is *Freyd's adjoint theorem*. We briefly recall some definitions which occur in the body of this theorem.

[17] Most of the examples here and in the rest of the paper can be found in Shulman (2008).

Definition 3. *A category is said to be* ***well-powered*** *if each of its objects admits a poset of subobjects.*

Definition 4. *A family Q of objects in a category $\mathbb{C}$ is called* ***cogenerating*** *if, given two parallel distinct morphisms, $f \neq g : a \to b$, there is a morphism $h : b \to q$ with $q \in Q$ such that $hf \neq hg$.*

Theorem 3. *Given a locally small, complete, well-powered, category $\mathbb{C}$ endowed with a cogenerating set, and a category $\mathbb{D}$, locally small, a functor $G : \mathbb{C} \to \mathbb{D}$ has a left adjoint if and only if it preserves small limits.*

For the proof see Lane (1998) ch. 5, par. 8. Note that this theorem relies essentially on some notion of size. If the theorem is expressed just for small categories we obtain the following.

Corollary 4. *Given a complete lattice, $\mathbb{C}$, a lattice morphism, $G : \mathbb{C} \to \mathbb{D}$, which preserves infima has a left adjoint.*

Clearly this corollary is just a shadow of Freyd's adjoint theorem. The significance of this latter can be appreciated if we think that in some cases this result represents the only device to build an adjunction.

4 Large categories and classes

small = "set" / large = "class"

Classes (more precisely proper classes) arise in classical set theory (ZFC) as those logical formulas without proper citizenship in models of set theory. They are built from set-theoretical formulas by means of unrestricted comprehension, and, even without a proper ontology,[18] they are commonly introduced as a useful device for manipulating formulas they abbreviate. As we are going to see in the next paragraphs, classes represent possible candidates to interpret large categories in a set-theoretical framework.

4.1 An approach internal to ZFC

A possible choice to give meaning to the notion of **large** categories is suggested by the usual convention adopted to introduce **classes** in ZFC. A class in the lan-

[18] They don't have proper ontolgy since they are outside the domain of discourse described by the axioms. Following Quine we can say that classes "do not have being" since they are not values of bound variables.

guage of ZFC is a formal expression of the form $x|\phi(x)$ where ϕ is a formula of the language of ZFC. Every set can be seen as a class (of its elements) but, by Russell's paradox, the converse is not true. We say that a class is a **proper class** if it is not a set.

Example 5. *The class of all sets, V, is defined by the formula*

$$V := \{x|x = x\}.$$

Another well-known proper class is the collection Ω of all ordinals. By the Burali-Forti paradox it cannot be a set.

The idea of this foundational recipe is very simple: we call a category *large* when the collection of its objects is a *proper class*.

One virtue of this approach is to work internally to ZFC: even if we cannot directly manipulate large objects we are still able to work with the properties (logical formulas in the language of ZFC) which define them. In this way we still have the possibility to perform simple basic constructions over large categories: for example if ϕ and ψ are formulas of ZFC we can still form the class of pairs whose first element satisfy ϕ and whose second element satisfy ψ, i.e. we can form the cartesian product of the two categories corresponding to ϕ and ψ.

The real problem of this approach is that ZFC cannot quantify over classes: theorems saying "there is a category such that…" or "for every category…" cannot even be stated in ZFC (one example is given by Freyd's adjoint theorem). Therefore, if we choose this foundational framework, we are led to reformulate most of our theorems as meta-theorems, which seems quite unpleasant from a foundational perspective.

4.2 NBG and MK

The most common set theory which introduces an ontology both for classes and sets is von Neumann, Bernays and Gödel's set theory (NBG). We briefly recall the axioms

(i) *axioms in common with ZFC*: pair, union, infinity, powerset;

(ii) *axioms both for sets and classes*: extensionality, foundation;

(iii) *axiom of limitation of size*: a class is a set if and only if it is not in bijection with the class of all sets V.

(iv) *axiom schema of comprehension*: for every property $\varphi(x)$, without quantifiers over classes, there exists the class $\{x|\varphi(x)\}$.

The system NBG is a conservative extension of ZFC: every sentence, *relative to sets*, which is provable in NBG, is already provable in ZFC. Therefore having

NBG as a foundation does not imply any particular ontological commitment. The differences with ZFC are mainly at a stylistic level.[19] As we mentioned in the first paragraph the use of NBG as a possible foundation for category theory trace back to the original paper of Eilenberg and Mac Lane.[20] The advantage of NBG with respect to ZFC consists essentially in the explicit treatment of classes: several constructions become easier, and, moreover, it is legitimate to quantify over classes. As suggested in Shulman (2008), another interesting feature of NBG consists in the possibility of adopting a form of **global choice**. This, surprisingly, is an easy consequence of the axioms. Consider the following observation due to von Neumann:

Theorem 6. *In NBG, the class of all sets, V, is well-orderable.*

Proof. The class Ω of all ordinals is a proper class and it is well-ordered. By the axiom of limitation of size this class is in bijection with V. This bijection induces a well order on V. □

The fact that V is well-orderable is one of the possible formulations of the axiom of choice for classes; in category theory the possibility to have this large choice is sometimes essential. In fact we are generally assuming it when choosing representatives of universal constructions over large categories. Despite these good points, and the several advantages over the approach internal to ZFC, NBG still presents some problems as a possible foundational framework for category theory: one, for example, is the use of comprehension restricted to formulas not involving classes.[21] A possible solution is then to strengthen the axioms of NBG by allowing for arbitrary quantification in the formulas involved. The resulting theory is known as Morse-Kelley set theory (MK). In this case, however, we have lost conservativity over ZFC, and the theory we end up with is genuinely stronger than ZFC.

In all the cases examined so far, a central problem has still to be overcome: none of these systems allow for the construction of the category of functors between two arbitrary categories. We can form the category of functors from a small category to an arbitrary one,[22] but this construction still remains illegiti-

[19] Historically the interest in this system have been motivated by the search for an equivalent system to ZFC which was finitely axiomatizable.

[20] S. Eilenberg (1945).

[21] When proving a statement $\varphi(n)$ by induction in ZFC or NBG we usually form the set $\{n \in N | \neg\varphi(n)\}$ and then use the fact that N is well-ordered. Since this argument involves an istance of comprehension, it can be carried on into these systems just in case φ does not involve quantifiers over classes.

[22] This is allowed in all the cases examined so far: for example in ZFC a functor $F : \mathbb{C} \to \mathbb{D}$, where $\mathbb{C}$ is a small category, is itself a set by replacement, and therefore the collection of all these functors form a class.

mate when the domain of these functors is a large category. However, to a great extent all these systems are expressive enough to cope with the cases of interest: even if the functor category seems a perfectly reasonable construction which can be performed regardless of size issues, most of category theory can be developed confining our attention to those functors whose domain is a small category. This limitation is consistent with the one on completeness.[23]

In summary, the foundational frameworks considered so far fulfil (with different degrees of approximation) the requirements (*A*) and (*B*) (p. 4), but none of them manages to satisfy (*C*) in its full generality. To sum up relative consistency of these systems (*D*) we can say that V_α models ZFC if and only if $(V_\alpha, Def(V_\alpha))$[24] models NBG. If α is an inaccessible than $(V_\alpha, V_{\alpha+1})$ models both NBG and MK.

5 Large categories and Universes

$small = \text{“}\in U\text{”} \; / large = \text{“}\in U\text{”}$

It is difficult to trace back to the first appearance of the concept of a universe. Essentially, it captures the idea of a collection closed under certain operations. But why introduce universes in the context of set-theoretic foundations of category theory? As Shulman (2008) suggests, we can reason as follows: on an *informal* level what we need for freely manipulating large categories seems to be a theory of classes which resembles closely ZFC; in practice it should be enough to have two copies of the axioms of ZFC, once for sets, once for classes. On a *formal* level, **universes** are introduced as a more elegant (and economic) solution to the same problem: instead of rewriting twice the axioms of ZFC we identify specific sets in our system as good candidates to interpret *large* collections.

5.1 Grothendieck's universes

As the name of this subsection suggests, the use of *universes* as foundational recipe for category theory goes back to Grothendieck. The purpose of his project, closely related to Bourbaki, was to justify the use of category theory in mathematical practice (and in Grothendick's perspective specifically in Algebraic Geometry).

Here is the definition of universe:[25]

[23] See here definition 2.

[24] $Def(X)$ denotes the set of all the subsets definable from element of X, i.e. sets of the form $\{x \in X | \varphi(x)\}$, where $\phi(x)$ can contain parameters from X and all its quantifiers range only over elements of X.

[25] In the original presentation (Bourbaki (1972), p. 185) the definition of universe also includes closure under ordered pairs which are a primitive notion in Bourbaki's presentation.

Definition 5. *A set U is a **Grothendieck universe** if the following conditions hold:*

(i) *if $y \in x \in U$, then $y \in U$;*

(ii) *if $x, y \in U$, then $\{x, y\} \in U$;*

(iii) *if $x \in U$, then $\wp(x) \in U$;*

(iv) *if $(x_i)_{i \in I}$ is a family of elements of U, and $I \in U$, then $\bigcup_{i \in I} x_i \in U$.*

In words, the definition says that U is a Grothendick universe if it is a transitive set closed under pairs,[26] power set and union of elements of U indexed by an element of U. In a more set-theoretical flavour, we can describe this definition as requiring $U = V_\kappa$ for some inaccessible cardinal κ (under the added hypothesis that U is uncountable[27]). Since inaccessible cardinals cannot be proved to exist in ZFC,[28] asserting the existence of a Grothendieck universe is a genuine strengthening of ZFC's axioms.

For a fixed Grothendieck universe U, we can rephrase our dichotomy between small and large by calling a category *large* whenever its collection of objects is a set *not belonging to* U. In case the universe is uncountable this is equivalent to assert that a category is small if and only if its collection of objects has *rank* less than κ, where κ is inaccessible.

This third approach, does not just give a satisfactory solution to conditions (A) and (B) (page 4), but it also allows for the construction of the category of functors between arbitrary categories (requirement C). In addition, it gives a more expressive semantics for the term *large*: we do not collapse every large collection to the size of V, as it happens in NBG, but we can retain a more careful distinction between *small*, *large* and *even larger* categories.

The following example gives an idea of the expressive power that we reach when introducing universes in the metatheory.

Example 7. *Every large category $\mathbb{C}$ has a category of presheaves $Set^{\mathbb{C}^{op}}$, and, if $\mathbb{C}$ is locally small*[29] *we can consider the Yoneda embedding $y : \mathbb{C} \hookrightarrow Set^{\mathbb{C}^{op}}$.*

Nevertheless we might want to be able to *encode* more abstract nonsense, and not satisfied by a single universe, we would like to have at our disposal a bigger universe U' (i.e. another inaccessible cardinal $\lambda > \kappa$), and then one other

[26] We do not assume as primitive the notion of ordered pair but, as usual, we define them à la Kuratowski: $(x, y) = \{\{x\}, \{x, y\}\}$.

[27] If we do not make any condition on the cardinality of U, also the emptyset, $\emptyset$, and ω are Grothendieck universes.

[28] A simple argument is the following: since V_κ, for κ inaccessible, represents a model of ZFC, if it was possible to prove the existence of such a cardinal in ZFC, then ZFC would also prove its own consistency, contradicting Gödel's second incompletness theorem.

[29] See table on page 14 for the definition of a locally small category in presence of a universe.

above.[30] For this reason Grothendieck's initial proposal consisted of adding not just a single universe but an abundance. Formally we can express this by adding to the usual axioms of ZFC the following:

Grothendick's axiom. *Every set is contained in some universe.*

This axiom guarantees the existence of sufficient large sets where every possible category we can meet is included.[31] Clearly, we have moved far from the strength of ZFC: the system obtained by adding Grothendieck's axiom to ZFC has the same consistency as ZFC + "there exist inaccessible cardinals of arbitrary size". As noticed by Mac Lane[32] this axiom does not solve definitely all the problems. We do not have any a priori certainty that changing universe does not affect the construction of our categories, or preserves all the properties of a specific object. Consider the following example

Example 8. *Let us assume that we have proved, for some property ϕ, the existence of a group G such that $\phi(G, H)$ is true for every small group H (for example ϕ could tell us that G is the limit of some diagram in* Grp*). The same argument still holds if we interpret the notion of largeness with some specific inaccessible κ, but there is no guarantee that the group G will be the same under all the possible interpretations.*

As kindly pointed out by one of the anonymous referees, in order to obtain this stronger property we should ask for the universe U to be an elementary substructure of V. For this, stronger axioms of infinity are needed, namely we have to ask for the cardinality of the universe to be at least a Mahlo number. The introduction of such large cardinals can be related to a general reflection principle for ZFC.[33] Even if the existence of these cardinals are given by axioms stronger than the one asserting the existence of a single inaccessible, and also stronger than Grothendieck's axiom, these axioms are still quite "weak" if compared to current large cardinal axioms used by set theorists. A similar approach based on a general reflection principle has been sketched by Engeler and Röhrl (1969). The following quotation concludes the paragraph where the two authors describe their proposal:[34]

> [...] However, the main objection to this approach is quite indepen-

[30] One possible reason is that we do not want just to consider the category of all small categories but also the category of all large ones, or of all locally small ones...

[31] As Shulman (2008) notes, this axiom asserts the possibility to enlarge the universe, more than asserting the existence of multiple stratified universes.

[32] See Lane 1969, p. 2.

[33] The interested reader should consult Lévy (1960). We will come back on a much weaker formulation of the reflection principle in section 5.3.

[34] See E. Engler (1969), p. 62.

> dent of the strength and questionability of the additional assumptions creating universes. We believe that it is a faulty to make a procrustes bed of set theory and try, bend or break, to fit all mathematical structures into it. This does injustice, in particular to category theory, as it denies the autonomous role that such theories play in mathematics.

To conclude our survey of the use of universes as foundation for category theory, we can sum up the situation with the following table:[35]

small	$Morph(\mathbb{C}) \in U$
locally small	$\forall c, d \in Obj(\mathbb{C}) \in U$ $Hom_{\mathbb{C}}(c, d) \in U$
large	$Morph(\mathbb{C}) \subseteq U$, $Morph(\mathbb{C}) \notin U$
enormous	$Morph(\mathbb{C}) \not\subseteq U$

5.2 Mac Lane's proposal

> [...] It turns out that a flexible and effective formulation of the present notions of category theory can be given with a more modest addition to the standard axiomatic set theory: the assumption that there is **one** universe.
>
> Saunders Lane (1969), p. 193.

As we have already mentioned in the last paragraph, one of the first mathematician who highlighted some problems of the foundational approach proposed by the French school of Grothendieck was Saunders Mac Lane, one of the founders of category theory.

In 1969 Mac Lane published a paper with a meaningful title: *One universe for the foundations of category theory.* In this work he argues that the existence of a single universe in ZFC is sufficient to have a foundational framework for

[35] Observe that we can always identify objects of $\mathbb{C}$ with identity morphisms. In this table we indicate with *Morph(*$\mathbb{C}$*)* the collection of morphisms of a category $\mathbb{C}$, and with $Hom_{\mathbb{C}}$*(c,d)* the set of morphisms between two given objects c, d of $\mathbb{C}$.

category theory. His proposal essentially consists in weakening Grothendieck's axiom asking, not for an abundance of universes, but just one.

Mac Lane defines a universe as follows:

Definition 6. *A set U is called a universe if:*

(1) $x \in y \in U implies x \in U$;

(2) $\omega \in U$;

(3) $x \in U implies \wp(x) \in U$;

(4) $x \in U implies \bigcup x \in U$;

(5) if $f : x \to y$ *is a surjective function such that* $x \in U$ *and* $y \subset U$, *then* $y \in U$.

As Mac Lane notices, the conjunction of condition (4) and (5) is equivalent to condition (*iv*) in definition 5. Apart from this and the requirement that U is uncountable (condition (2) and (3)), the definition is the same as that given by Bourbaki.[36]

In this framework the systematization of the dichotomy *small/large* is essentially the same as that of Grothendieck's school (See table on page 14). The restriction to a single universe allows for a (almost[37]) complete treatment of category theory and, at the same time, allows us to escape from the "jungle" of multiple universes.[38]

Finally we remark that consistency of Mac Lane's proposal amounts to the consistency of ZFC + "there exists a strong inaccessible cardinal".

5.3 Feferman's proposal

Foundations of category theory have represented a problem of major interest for Solomon Feferman, who came back to this topic several times during the last forty years. He dedicated four papers[39] to this issue, proposing more than a single solution. Here we confine ourselves to the analysis of his first proposal.

The first paper where Feferman addresses the question was published in 1969.[40] In this work he proposes an alternative to the solution of adopting new axioms for universes. Feferman's idea consists in using a well-known principle of set theory, namely the *reflection principle*.

[36] We also recall the treatment of ordered pairs as a primitive entity, characteristic of Bourbaki's approach.

[37] This approach does not allow for the construction of the category of all large categories.

[38] As remarked by one of the anonymous referee, the request of a single universe U inside V could be seen as a kind of opprobrium from the point of view of a set theoretician. An alternative solution to the universe juggling has been mentioned at the end of the last section: see for example E. Engler (1969).

[39] Namely Feferman (1969), Feferman (1977), Feferman (2006), Feferman (2004).

[40] See Feferman (1969).

Feferman's system, which we indicate as ZFC/s,[41] consists, in the first instance, in adding a new constant symbol s to the usual language of ZFC. Secondly we add to the axioms of ZFC further axioms in order to describe (the interpretation of) s as a natural model of ZFC.[42]

Before giving the axioms we recall that if φ is a formula of the language of ZFC, its relativization to s, denoted by φ^s , is given when all the quantifiers that occur in φ are bounded by s.[43]

Definition 7. *The system ZFC/s is given in the language $\mathcal{L}$ of ZFC extended with the constant symbol s by the following axioms:*

(1) Axioms of ZFC: extensionality, emptyset, pairs, union, powerset, infinity, foundation, replacement, choice.

(2) s is not empty:

$$\exists x(x \in s)$$

(3) s is transitive:

$$\forall x, y(y \in x \wedge x \in s \rightarrow y \in s)$$

(4) s is closed under subsets:

$$\forall x, y(x \in s \wedge \forall z(z \in y \rightarrow z \in x) \rightarrow y \in s)$$

(5) reflection axioms: for every formula φ with free variable $x_1, \ldots, x_n$:

$$\forall x_1 \ldots \forall x_n(\varphi(x_1, \ldots, x_n) \leftrightarrow \varphi^s(x_1, \ldots, x_n))$$

The axiom schema (5) can be read in model-theoretic terms as follows: let (M, ϵ, S) be a model of ZFC/s,[44] call M_s the set $\{x \in M | x \epsilon S\}$, and ϵ_s the restriction of ϵ to M_s,[45] , then (M, ϵ) is an elementary extension of (M_s, ϵ_s). In other words the two models satisfy the same formula in the language $\mathcal{L}$.

As we mentioned, this axiom schema, is based on the reflection principle. A specific istance of this principle can be suitably reformulated as a theorem of ZFC. It might be helpful to highlight the common point this *reflection theorem* shares with the downward Löwenheim-Skolem theorem. The proof of the latter shows, given a model N of a theory T and an infinite subset $S \subset N$, how to

[41] the symbol s stands for *smallness*.

[42] A natural model of ZFC, (M, ϵ), is a transitive model of ZFC, closed under subsets: $x \subset y \in M$ implies $x \in M$.

[43] For example, the relativization to s of the formula $\forall a \exists b \forall x(x \in b \leftrightarrow \psi(x, a))$ is

$$\forall a \in s \exists b \in s \forall x \in s(x \in b \leftrightarrow \psi s(x, a)).$$

[44] We indicate with S the element of M which interprets the constant symbol s.

[45] i.e. $x \epsilon_s y$ iff $x, y \in M_s$ and $x \epsilon y$.

build a model M, such that $M \prec N$ (M is an elementary substructure of N) and $|N| = |S|$. In order to obtain the model M we build a sequence of sets M_n in this way: starting from $M_0 = S$ every M_{n+1} is obtained from M_n by adding a witness $b \in N$ for every existential sentence $\exists y \phi(y, x_1, \ldots, x_n)$ and every n-tuple of elements $a_1, \ldots, a_n \in M_n$ such that $\exists y \in N \phi N(y, a_1, \ldots, a_n)$ is a true sentence. M is then obtained as

$$M = \bigcup_{n \in \omega} M_n.$$

Since there are just a countable amount of sentences ϕ, the cardinality of the various M_n never increases. Finally, the countable union of countable sets is still countable from which it follows that $|M| = |S|$. This construction can be rearranged to be carried out on the cumulative hierarchy of V_α's. Even if this method enables us "to build models of ZFC", this does not violate Gödel's second incompletness theorem. Indeed even if we can reflect every finite conjunction of sentences of ZFC, we are not able to reflect at once a single infinite conjunction of sentences expressing that V_κ is a natural model of ZFC for a specific κ.

One of the main advantages of this "logical" approach consists exactly in this: the "formal description" of s as a "natural model of ZFC" is not sufficient to prove in ZFC that (the interpretation of) s is a natural model of ZFC. This, in fact, allows Feferman to prove the following important result in Feferman (1977):

Theorem 9. *ZFC/s is a conservative extension of ZFC.*

This result guarantees that we have not really strengthen our starting set theory; in particular, in categorical terms this means that all that can be proved in ZFC/s about *small* objects, even using *large categories*, can already be proved in ZFC. Now it should be sufficiently clear that interpreting *small* as "element of S" and *large* as "set not necessarily in S", what we have is an appropriate foundational framework where it is possible to interpret definitions and theorems of category theory.

As in the other cases we can evaluate the expressive power of Feferman's system using the conditions on page 4. While we can check that ZFC/s easily meet (A), (B), (D), the problem with functor categories noticed with other systems is also complicated in this case: we do not only have a limitation of size for the domain of the functors, but we also have to confine ourselves to consider those functor categories whose objects are S-definable. This is a consequence of the relativization of the replacement axioms to s, which can be considered to express the inaccessibility of s under all functions definable in $\mathcal{L}$.[46] In other words,

[46] See Feferman (1969), p. 208.

if we read the replacement axioms as saying that the image of a set under a class-function is still a set, their relativization can be rephrased as stating that the image of a set under a function-class *which is definable from elements of S* is still a set. This restriction, even if apparently innocuous, can have annoying consequences: for example we should change the notion of completeness (definition 2) in ZFC/s, requiring limits for "all small functors" (i.e functors definable from S) rather then for all functors with small domain.

5.4 Some comments on Feferman's proposal

Feferman has been one of the first mathematical-logicians to get interested in foundations of category theory: his motivation has been primarily to fill the gap between the rapid development of category theory and its proper systematization inside the mathematical edifice.

Initially a careful attitude led him to investigate the foundations of this theory with different systems of set theory. Only later he turned his attention to a critical analysis of a categorical foundation of all mathematics.[47]

The system proposed by Feferman in Feferman 1969 and the conservativity result over ZFC are of particular interest for a foundational analysis of category theory. The relevance of Feferman's contribution is well expressed in the words of Blass[48]

> This approach developed in Feferman (1969), has two advantages. First, the assumptions guarantee that, if we prove a theorem about small sets by using large categories, then the same theorem holds for arbitrary sets; [...]. Second, the assumptions do not really go beyond ZFC; any assertion in the first-order language, not mentioning κ,[49] that can be proved using these assumption can also be proved without them.

The second feature of Feferman's system mentioned by Blass is the conservativity result of the previous paragraph (theorem 9). This theorem highlights the "conventional" use of inaccessible cardinals when discussing set-theoretic foundation of category theory. As Shulman notes:

> [...] Thus we obtain a precise version of our intuition that the use of inaccessibles in category theory is merely for convenience: since many categorical proofs stated using inaccessibles can be formalized

[47] The main argument of his criticism for a possible categorical foundation of all math was firstly formulated in his '77 paper Feferman (1977). He then came back to the same argument in his successive works.

[48] Blass (1984), page 8.

[49] Blass uses κ to indicate the level of the cumulative hierarchy which corresponds to the interpretation of the added constant symbol s in Feferman's system.

> in ZFC/s, any *consequence* of such a theorem not referring explicitly to inaccessibles is also provable purely in ZFC.

Even if inaccessible cardinals, and in general stronger axioms of infinity, have became part of modern mathematical research, their use in foundational contexts remains dubious. Again, in the words of the philosopher Marquis:[50]

> Any reference to inaccessibles is simply removed. This is an exact formulation of the conviction that questions of size are only used to justify certain general construction and they do not bear on the real mathematical content of the constructions and its consequences. [...] Feferman's results[51] are important for they can be interpreted as showing that *as far as set theory is concerned*, category theory does not raise *new* foundational problem.

[50] Marquis (2009), pp. 183-184.
[51] The reference is to Feferman (1969).

References

Awodey, S. (1996). "Structuralism in mathematics and logic". In: *Philosophia Mathematica* 3.4.

— (2004). "An Answer to G. Hellman's Question "Does Category Theory Provide a Framework for Mathematical Structuralism?"" In: *Philosophia Mathematica* 1.12.

Blass, A. (1984). "The interaction between category theory and set theory". In: *Mathematical Application of Category Theory*. Ed. by J. Grey. Contemporary Mathematics, vol. 30.

Borceaux, F. (1994). *Handbook of categorical algebra 1. Basic Category Theory*. Cambridge University Press.

Bourbaki, N. (1972). "Univers. Théorie des topos et cohomologie étale des schémas. Tome 1: Théorie des topos". In: *Seminare de Geometrie Algebrique du Bois-Marie 1963-1964 (SGA 4)*. Berlin: Springer-Verlag, pp. 185–217.

Caramello, O. (2010). "The Unification of Mathematics via Topos Theory". In: DOI: arXiv:math.CT/1006.3930.

E. Engler, H. Röhrl (1969). "On the Problem of Foundations of Category Theory". In: *Dialectica* 3.1.

Feferman, S. (1969). "Set-Theoretical Foundations of Category Theory". In: *Reports of the Midwest Category Seminar III, Lecture Notes in Mathematics* 106, pp. 201–247.

— (1977). "Categorial foundations and foundations of category theory". In: *Logic, foundations of mathematics and computability theory*. Fifth Internat. Congr. Logic, Methodology and Philos. of Sci. Dordrecht: Reidel, pp. 149–169.

— (2004). "Typical ambiguity: trying to have your cake and eat it too". In: *One hundred years of Russell's paradox*. Ed. by G. Link. Berlin.

— (2006). "Enriched stratified system for the Foundation of Category Theory". In: *What is category theory?* Ed. by G. Sica. Milano: Polimetrica.

Hellman, G. (2003). "Does Category Theory Provide a Framework for Mathematical Structuralism?" In: *Philosophia Mathematica* 11.2.

Hofstadter, D. (1979). *Gödel, Escher and Bach. An eternal Golden braid*. New York: Basic Book Inc. Publisher.

Joyal, A. (1995). *Algebraic set theory*. cambridge University Press.

Lane, S. Mac (1969). "One Universe as a foundation for Category Theory". In: *Reports of the Midwest Category Seminar, III, Lecture Notes in Mathematics*. Ed. by S. Mac Lane. Vol. 106. New York: Springer-Verlag, pp. 192–200.

Lane, S. Mac (1998). "Categories for the working mathematician". In: *Graduate Texts in Mathematics, second edition*. Vol. 5. Springer.

Lawvere, F. W. (2005). "An elementary theory of the category of sets (long version)". In: *Reprints in Theory and Application of Categories*. Vol. 11.

Lévy, A. (1960). "Axiom schemata of strong infinity in axiomatic set theory". In: *Pacific Journal of Mathematics* 10, pp. 223–238.

Marquis, J. P. (2009). "From a geometrical point of view. A study of the History and Philosophy of Category Theory". In: *Series: Logic, Epistemology, and te Unity of Science*. Vol. 14. Springer.

McLarty, C. (2004). "Exploring Categorical Structuralism". In: *Philosophia Mathematica* 3.12.

S. Eilenberg, S. Mac Lane (1945). "A General Theory of Natural Equivalences". In: 58.2, pp. 231–294.

Shulman, M. A. (2008). "Set theory for Category Theory". In: DOI: arXiv:0810.1279v2[math.CT].

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

[entry]nyt/global/

Epistemic Logic and the Problem of Epistemic Closure

Davide Quadrellaro[1]

Abstract. This paper argues that propositional modal logics based on Kripke-structures cannot be accepted by epistemologists as a minimal framework to describe propositional knowledge. In fact, many authors have raised doubts over the validity of the so-called *principle of epistemic closure*, which is always valid in normal modal logics. This paper examines how this principle might be criticized and discusses one possible way to obtain a modal logic where it does not hold, namely through the introduction of impossible worlds.

Keywords. Epistemology, Epistemic Logic, Epistemic Closure, Rantala Semantics, Logic of Knowledge, Impossible Worlds.

[1]I wish to thank Greg Sax, Gianmarco Defendi and three anonymous referees for comments and suggestions.

Copyright. © BY $ 2018 Davide Quadrellaro. Published in Italy. Some rights reserved. First published on *Rivista Italiana di Filosofia Analitica Junior* 7:2 (2016), pp. 3-16.
Author. Davide Quadrellaro, davide.quadrellaro@gmail.com.

1 Introduction

The purpose of this article is to describe a minimal logic of knowledge which can be used by epistemologists with different philosophical orientations. A first way to proceed is describing a modal logic based on a Kripke-semantics, specifying how the accessibility relation should be restricted in order to represent knowledge. However, it is not difficult to prove that this standard formal epistemological analysis implies the validity of the principle of epistemic closure, namely of the fact that, if one both knows that p and that if p then q, then he/she also knows that q. This principle, however, has been object of criticism and objections by some epistemologists. Therefore, if we are looking for a general modal logical framework that can be used by philosophers with different orientations, we have to construct a formal system where the closure principle does not hold. An interesting way to proceed is working with the semantics which has been developed by logicians to account for the paradox of the logical omniscience. In fact, if we introduce the "impossible worlds" and we construct a Rantala-semantics based on them, we obtain a weaker logic where the closure principle does not hold.

In the first part of this article I present the modal logic **T**, which is generally considered the minimal formal system for the logic of knowledge. Firstly I introduce the syntax and the semantics of modal logic, secondly I characterize how the accessibility relation R_a has to be restricted in order to obtain the logic **T**. In the second part I prove that the principle of epistemic closure follows from **T** and I try to underline some critical aspects of it. In the third part I introduce an alternative logic for knowledge where the closure principle does not hold, namely a modal logic with impossible worlds and a Rantala-semantics. Finally, in the fourth part, I evaluate this proposal, trying to underline both upsides and downsides of it.

2 The standard logic of knowledge

A first way to give a formal account to epistemological concepts such as belief and knowledge is to adopt the language of modal logic. Even if the modal operators $\Diamond$ and $\Box$ are usually read as possibility and necessity, we can also adopt an epistemic interpretation of them. On this alternative reading we will translate a logical formula like $\Box p$ not as "it is necessary that p" but rather as "it is known that p", "it is believed that p" or "it is certain that p". Following each of these interpretations we can formulate a different modal logic, in order to formalize the specific features of the considered epistemic operator. In what follows I will be interested exclusively in the former of these alternatives and I will focus my attention on the *logic of knowledge*.

Working with an epistemological interpretation of modal logic, it is worth

Rivista Italiana di Filosofia Analitica Junior 9:2 (2018)

specifying who is the subject of the knowledge we are speaking about. If we read $\Box p$ simply as "it is known that p", the meaning of this operator remains not clear enough. What does it mean, in fact, that something is known? Does it mean that someone knows it? Or does it mean that everyone knows it? Therefore, in order to be as clear as possible, we should adopt a more intuitive terminology and make explicit the fact that we are working with a *propositional notion of knowledge* and within a logic of *individual agents*. The box operator will be substituted by a K (for "knowledge"), followed by a letter that indicates who is the agent that knows the considered proposition. Modal formulas will look, thus, like $K_a p$ and $K_b p$ and they will be read as "the agent a knows that p" and "the agent b knows that p". In what follows, we will be interested in formal systems with only one agent, but it is important to keep in mind that we can introduce many K-operators, in order to map the knowledge of more than one subject[1].

Let us now move, after these introductory remarks, to give a precise definition of the *syntax* of the propositional modal logic for knowledge. We proceed extending the alphabet of classical propositional logic with a knowledge operator K_a.

Definition 2.1 (Alphabet of Propositional Modal Logic for Knowledge)**.** An alphabet for propositional modal logic for knowledge is defined as the union of the following disjointed sets:

- A denumerable set of *atomic propositional variables* $\mathcal{P} = \{p_0, p_1, ...\}$.
- The set of the *logical connectives* $C = \{\neg, \wedge, \rightarrow\}$.
- The set of the *knowledge operator* $O = \{K_a\}$.
- The set of *auxiliary symbols* $\mathcal{A} = \{(,)\}$.

Given the alphabet, it is possible to define inductively the set of the formulas of the logic of knowledge.

Definition 2.2 (Formulas of Propositional Modal Logic of Knowledge)**.** The formulas of the modal logic of knowledge are given by the following definition by induction:

1. If φ is an atomic propositional variable, then φ is a formula.
2. If φ is a formula, then also its negation $\neg\varphi$ is a formula.
3. If φ and χ are formulas, then also their conjunction $(\varphi \wedge \chi)$ is a formula.
4. If φ and χ are formulas, then also the conditional $(\varphi \rightarrow \chi)$ is a formula.

[1] For the introduction of multiple agents see both ´ family=Hendricks, familyi=H., given=Vincent, giveni=V., ´ family=Symons, familyi=S., given=John, giveni=J., (2015, pp. 9-11) and ´ family=Holliday, familyi=H., given=Wesley H., giveni=W. H., (forthcoming, pp. 5-7).

5. If φ is a formula, then also $K_a\varphi$ is a formula.

6. Nothing else is a formula.

The *semantics* of the logic of knowledge is provided by a Kripke-structure, which is the standard way to interpret modal languages.

Definition 2.3 (*Kripke-structures*)**.** Given a propositional modal logic of knowledge, a *Kripke-structure* $\mathcal{M}$ is a triple $\langle W, R_a, V \rangle$, where:

1. W is a non-empty set. Intuitively, W is a set of "possible worlds" or "possible scenarios".

2. R_a is a binary relation over W, i.e. a subset of $W \times W$. Intuitively, we read vR_aw as "the possible world w is epistemically accessible from the possible world v by the agent a".

3. V is a function that assigns to every atomic propositional formula a subset of W. Intuitively, V specifies in which possible worlds each atomic formula is true.

Given the Kripke-structures, we can define the notion of truth in a world:

Definition 2.4 (*Truth in a world*)**.** Given a propositional modal logic for knowledge, a Kripke-structure $\mathcal{M}$ and a world w, the notion $\mathcal{M} \vDash_w \varphi$ of being true in a world is defined as follows:

1. when φ is atomic, then $\mathcal{M} \vDash_w \varphi$ iff $w \in V(\varphi)$;

2. when φ has the form $\neg\chi$, then $\mathcal{M} \vDash_w \varphi$ iff $\mathcal{M} \nvDash_w \chi$;

3. when φ has the form $(\chi \wedge \psi)$, then $\mathcal{M} \vDash_w \varphi$ iff $\mathcal{M} \vDash_w \chi$ and $\mathcal{M} \vDash_w \psi$;

4. when φ has the form $(\chi \rightarrow \psi)$, then $\mathcal{M} \vDash_w \varphi$ iff $\mathcal{M} \nvDash_w \chi$ or $\mathcal{M} \vDash_w \psi$;

5. when φ has the form $K_a\chi$, then $\mathcal{M} \vDash_w \varphi$ iff for every possible world v such that wR_av, $\mathcal{M} \vDash_v \chi$.

The definition of truth in a world allows us to define two further important notions. We say that a formula φ is *true in a model* $\mathcal{M}$ if and only if it is true in every world $w \in W$ of the Kripke-structure $\mathcal{M}$. We say that a formula φ is a *valid formula* if and only if it is true in every world $w \in W$ of every Kripke-structure $\mathcal{M}$.

What we have described so far is the minimal system **K** of modal logic, with the only peculiarity that the informal reading that we have assumed for the modal operator is "the agent a knows that...". Nevertheless, it is clear that to obtain a logic of knowledge this is not enough. What one needs, rather, is to specify the formal properties that are typical of knowledge and to represent them in the logic. Putting specific restrictions over the accessibility relation R_a, it is possible

to obtain many modal logics stronger than **K**, where more principles are valid formula. The problem is that it is not sufficiently clear which modal system photographs in the correct way the formal properties of knowledge. Since the purpose of this article is to examine which logic can be accepted by epistemologists with different philosophical orientations, we will extend **K** only with those principles which are generally taken for granted in the epistemological debate. Therefore, the only restriction that we want impose to our logical system is that it has to satisfy the following principle:

(T) $K_a\varphi \rightarrow \varphi$

What (T) says is that, if one knows a proposition, then this very same proposition must be true. This does not only follow from any analysis of knowledge as true belief plus something, but it also seems to be a valid minimal description of the meaning of knowledge. Indeed, if one says that he/she knows that p but it is not the case that p, it seems reasonable to conclude that he/she *does not* know that p, but rather only *believes* that p[2].

If we want that the principle (T) holds in the logical framework that we are considering, we have to put a restriction on the accessibility relation R_a. More precisely, as we prove with the following theorem, we have to restrict our attention to those Kripke-structures where the accessibility relation is reflexive. The modal logic that we obtain when we work only with reflexive accessibility relations is called **T**.

Theorem 2.1. *Given the language of propositional modal logic and its Kripke-structure $\mathcal{M} = \langle W, R_a, V\rangle$, the formula (T) $K_a\varphi \rightarrow \varphi$ is a valid formula if and only if the accessibility relation R_a is reflexive.*

Proof: Assuming that the accessibility relations R_a in $\mathcal{M}$ is reflexive, then given any possible world $w \in W$ we have that wR_aw. Therefore, since $\mathcal{M} \vDash_w K_a\varphi$ holds, then $\mathcal{M} \vDash_v \varphi$ holds in every world v such that v is accessible from w. But for reflexivity we have that w is accessible from itself and, therefore, that $\mathcal{M} \vDash_w \varphi$. *Vice versa*, assuming that $K_a\varphi \rightarrow \varphi$ is a valid formula then, for every Kripke-structure $\mathcal{M}$ and every world w in it $\mathcal{M} \vDash_w K_a\varphi \rightarrow \varphi$. Given the semantics of the conditional, this amounts to say that it is not the case that $\mathcal{M} \vDash_w K_a\varphi$ and $\mathcal{M} \nvDash_w \varphi$. But, if R_a was not reflexive, we could construct a Kripke-structure such as $\mathcal{N} = \langle W, R_a, V\rangle$, with $W = \{v, w\}$ and $R_a = \{\langle w, v\rangle\}$. In $\mathcal{N}$ we have that, if $v \in V(\varphi)$ but $w \notin V(\varphi)$, then $\mathcal{N} \vDash_w K_a\varphi$ but $\mathcal{N} \nvDash_w \varphi$, contradicting our claim that $K_a\varphi \rightarrow \varphi$ is a valid formula. Therefore, R_a must be reflexive. ■

[2]This aspect is famously stressed by ´ family=Wittgenstein, familyi=W., given=Ludwig, giveni=L., (1969).

3 The principle of epistemic closure and its problems

In the previous part of this article I have introduced the modal logic **T**, in order to represent some minimal formal properties of knowledge. Moving a step further, it is now possible to prove an interesting result, which says that the principle of epistemic closure is a valid formula in **T**. Firstly, let us clarify what we mean with the name of "principle of epistemic closure".

(CP) If an agent knows that φ and he/she knows that if φ then χ, then he/she also knows that χ.

It is straightforward to translate this thesis into the language of the logic of knowledge. We thus obtain the following formal version of the closure principle:

(FCP) $(K_a\varphi \wedge K_a(\varphi \rightarrow \chi)) \rightarrow K_a\chi$

We can now prove the following theorem:

Theorem 3.1. *Given the logic of knowledge **T**, the formal closure principle (FCP) is a valid formula.*

Proof: We reason for absurd. If (FCP) was not a valid formula, there would be a world w of a Kripke-structure $\mathcal{M} = \langle W, R_a, V\rangle$, where (FCP) does not hold. Given the semantics of the conditional, this means that $\mathcal{M} \vDash_w K_a\varphi \wedge K_a(\varphi \rightarrow \chi)$ but $\mathcal{M} \nvDash_w K_a\chi$. Given $\mathcal{M} \vDash_w K_a\varphi$, we have that in every world accessible from w, φ holds. Given $\mathcal{M} \vDash_w K_a(\varphi \rightarrow \chi)$, we have that in every world accessible from w, $\varphi \rightarrow \chi$ holds. Moreover, since $\mathcal{M} \nvDash_w K_a\chi$, there is at least one world v such that wR_av where $\mathcal{M} \nvDash_v \chi$. But in this same world v we have that $\mathcal{M} \vDash_v \varphi$ and $\mathcal{M} \vDash_v \varphi \rightarrow \chi$ hold too, from which it follows that $\mathcal{M} \vDash_v \chi$. Therefore, we obtain the contradiction that $\mathcal{M} \vDash_v \chi$ and $\mathcal{M} \nvDash_v \chi$. ■

If our concerns are mainly epistemological this result has a particular relevance. In fact, what we have proved is that even if we work with a weak modal system, the principle of epistemic closure will hold in it[3]. Therefore, if we have some reason to refuse the principle of epistemic closure, then we can not adopt the formal logic **T** anymore, for it describes knowledge in a way which is inconsistent with our theory. In particular Dretske (1970) offers at least two possible reasons to refuse the closure principle[4]. In the rest of this part I will present both

[3] Notice, moreover, that in the proof of the theorem 3.1. we did not make any use of the fact that the accessibility relation between worlds is reflexive. Therefore, our proof is valid also for the basic modal logic **K**.

[4] ´ family=Luper, familyi=L., given=Steven, giveni=S., (2016) synthesizes a wide range of arguments against the closure principle, often originally raised by Dretske and Nozick. However, even if Luper's reconstruction is clear, I do not agree with his presentation of the arguments from the "analysis of knowledge". In fact, the theories of knowledge supported by Dretske and Nozick are *explanations* of why the closure principle fails and not *reasons* to refuse it. Luper commits, therefore, a sort of inversion of the right order of explanation.

of them, but I will not try to set the question about their validity. Indeed, I only want to show that it might be reasonable for an epistemologist to reject the closure principle. In fact, given the possibility that (FCP) is not acceptable, we have to look for a modal logic for knowledge weaker than the standard one described by the Kripke-structures. Our purpose, in fact, is not to take part in the epistemological debate and to identify the modal logic which better describes knowledge but, rather, it is to find a minimal logical framework which can be accepted by epistemologists of different currents.

A first critique to the principle of epistemic closure is linked to skepticism. In fact, one general way to reconstruct the argument presented by the skeptic is with the following argument:

(1) I do not know that I am not a brain in a vat
(2) If I do not know that I am not a brain in a vat, then I do not know that I have hands.

---

(3) I do not know that I have hands ∴

The premiss (2) of this argument is a consequence of an instance of (CP). If I know that I have hands and I know that if I have hands I am not a brain in a vat, then I know that I am not a brain in a vat. Therefore, if I do not know that I am not a brain in a vat, then either I do not know that I have hands, or I do not know that if I have hands I am not a brain in a vat. However, since I know that if I have hands I am not a brain in a vat, we can exclude the second disjunct and obtain (2): if I do not know that I am not a brain in a vat, then I do not know that I have hands[5].

If skepticism is expressed in the form of the syllogism presented above, there are two main strategies to criticize it. Either one denies the premiss (1), either one denies the premiss (2), namely the closure principle. The first horn was chosen by Moore (1939), who reversed the skeptic's argument in its contraposed version[6].

[5]It is worth underlining that, in order to obtain (2) from (CP), we have to take for granted that we know that if we have hands we are not a brain in a vat. Although this might seem trivial, there are two problematic aspects which deserve some further reflections. On the one hand, one may think that it is much more reasonable to deny the premiss of the argument from (CP) to (2), namely to assert that we do not know that if we have hands then we are not a brain in a vat, rather than to accept the conclusion it leads to, i.e. that we do not know that we have hands. On the other hand, there might be a skeptical scenario that we do not know, or a person who never thought about brains in a vat. But if one has never thought about a skeptical scenario, it does not seem plausible to say that he/she knows that if he/she has hands, then he/she is not in the considered skeptical scenario.

[6]For historical's sake, let me remark that Moore did not deal with the brain in a vat hypothesis in his original article of 1939, but he rather considered more traditional skeptical scenarios.

(1) I do know that I have hands
(2) If I do not know that I am not a brain in a vat, then I do not know that I have hands.

---

(3) I do know that I am not a brain a vat ∴

However, this solution implies that we do actually know that we are not brains in a vat, which is a conclusion that many might find excessively strong. Therefore, if we want to remain faithful both to the intuition that we do know that we have hands, both to the intuition that we do not know that we are not brains in a vat, we have to abandon the closure principle. Notice that this is not an argument against skepticism. If we want to criticize skepticism *because* the closure principle does not hold we need independent arguments against (CP). On the contrary, this is an argument against the closure principle, *because* skepticism does not hold. So, what this argument needs are independent reasons to refuse skepticism.

However, Dretske criticizes the principle of epistemic closure also in a second more explicit way, bringing some counterexamples to it. Perhaps the most famous one is the so-called "zebra case". Imagine that you are in a zoo with your nephew. While you are walking around, he asks you if you know what is the animal you are looking at. You observe it, you notice that it looks exactly how you expect a zebra should look like, and you also find a sign with "zebra" written on it. Without any further doubt you would reply to your nephew's question something like: "Sure! It is a zebra". Thus, you do know that the animal you are observing is a zebra. But do you know that it is not a disguised mule? Indeed, it might be a mule so well depicted by the zoo-officers to look exactly like a zebra, maybe in order to attract more visitors.

Examples like this present a sort of strange situation. On the one hand, we have a plenty of reasons to believe that the animal we are observing is a zebra. On the other hand, we do not know that it is not a disguised mule. Moreover, we are also completely aware that mules and zebras are different animals. Therefore:

(i) we know that the animal we are looking at is a zebra;

(ii) we know that if the animal we are looking at is a zebra, then it is not a disguised mule;

(iii) we do not know that the animal we are looking at is not a disguised mule.

Clearly, (i), (ii) and (iii) taken together are an instance of failure of the closure principle.

Together, these two arguments show that the principle of epistemic closure is not so obvious and trivial as one might believe at first sight. A closer examination of it shows both that it has skeptical consequences and that it does not

always fit our intuitions in concrete examples. Therefore, if we want to find a propositional modal logic which describes some minimal properties of knowledge generally accepted by epistemologists we have to weaken in some way the logic of knowledge that we have previously presented.

4 The impossible worlds and the Rantala-semantics

In the context of the logical literature, an alternative to the standard Kripke-semantics has been provided in order to account for the problem of logical omniscience. In fact, one further consequence of adopting a modal logic like **K** or stronger is that any agent knows every classical tautology. In fact, since classical tautologies are valid in every possible world, the agent always knows them, for they are trivially true in all the worlds which the agent has access to. Although it is important to keep distinct the problem of the epistemological closure principle from the one of the logical omniscience, we can try to apply the logical system used to answer to the latter of these problems also to respond to the former one[7].

Given the syntax of modal logic that we have already defined, we can introduce a slightly different semantics, namely a Rantala–semantics[8].

Definition 4.1 (*Rantala-structures*)**.** Given a propositional modal logic of knowledge, a *Rantala-structure* $\mathcal{R}$ is a quadruple $\langle W, W', R_a, V \rangle$, where:

1. W is a non-empty set. Intuitively, W is a set of "possible worlds" or "possible scenarios".

2. W' is a set. Intuitively, W' is a set of "impossible worlds" or "impossible scenarios".

3. R_a is a binary relation over $W \cup W'$, i.e. a subset of $(W \cup W') \times (W \cup W')$. Intuitively, we read vR_aw as "the possible or impossible world w is epistemically accessible from the possible or impossible world v by the agent a".

4. V is a function that assigns to every atomic propositional formula a subset of $W \cup W'$ and to every formula a subset of W'. Intuitively, V specifies in which possible or impossible worlds each atomic formula is true, and in which impossible worlds each formula is true.

As one can immediately notice, the difference between the Kripke and the Rantala structures relies on the introduction of a set of impossible worlds. To see how

[7] On the difference between the problem of logical omniscience and the one of epistemic closure see ´ family=Holliday, familyi=H., given=Wesley H., giveni=W. H., (forthcoming, pp. 8-10).

[8] The name of Rantala-semantics comes from the Finnish logician Veikko Rantala. Here I follow the presentation of its semantics given by ´ family=Wansing, familyi=W., given=Heinrich, giveni=H., (1990), who also provides an interesting comparison between the Rantala-semantics and other methods to solve the paradox of logical omniscience.

they affect the interpretation of every formula, we shall reformulate also the notion of truth in a model.

Definition 4.2 (*Truth in a world*)**.** Given a propositional modal logic for knowledge, a Rantala-Structure $\mathcal{R}$ and a world w, the notion $\mathcal{R} \vDash_w \varphi$ of being true in a world is defined as follows:

1. If $w \in W'$, namely if w is an impossible world, then $\mathcal{R} \vDash_w \varphi$ iff $w \in V(\varphi)$;
2. If $w \in W$, namely if w is a possible world, then:
   (a) when φ is atomic, then $\mathcal{R} \vDash_w \varphi$ iff $w \in V(\varphi)$;
   (b) when φ has the form $\neg\chi$, then $\mathcal{R} \vDash_w \varphi$ iff $\mathcal{R} \nvDash_w \chi$;
   (c) when φ has the form $(\chi \wedge \psi)$, then $\mathcal{R} \vDash_w \varphi$ iff $\mathcal{R} \vDash_w \chi$ and $\mathcal{R} \vDash_w \psi$;
   (d) when φ has the form $(\chi \rightarrow \psi)$, then $\mathcal{R} \vDash_w \varphi$ iff $\mathcal{R} \nvDash_w \chi$ or $\mathcal{R} \vDash_w \psi$;
   (e) when φ has the form $K_a\chi$, then $\mathcal{R} \vDash_w \varphi$ iff for every possible or impossible world v such that wR_av, $\mathcal{R} \vDash_v \chi$.

It is now possible to clarify which is the role that the impossible worlds play in the new structure now defined. A first notable aspect is that, while in regards of the possible worlds the notion of truth in a world is defined inductively, the truth-value of every formula in an impossible world is directly specified by the assignment V. In an impossible world we might have that a disjunction is true even if its two disjuncts are both false, or that even if two formulas are true their conjunction is false, and so on. The distinguished aspect of this structure is that the anomalous behaviour of impossible worlds has some consequences on the evaluation of formulas in "normal" possible worlds. In fact, in order for a modal formula like K_ap to be true in a possible world w, the formula p has to be true in every world v, both possible and impossible, such that wR_av.

The notion of valid formula has now to be defined for the new Rantala-semantics: we say that a formula φ is a *valid formula* if and only if it is true in every possible world of every Rantala-structure. Given this new definition and thanks to the introduction of the impossible worlds, we can show that the principle of epistemic closure (FCP) is not a valid formula anymore. In fact, even if $\mathcal{R} \vDash_w K_a\varphi$ and $\mathcal{R} \vDash_w K_a(\varphi \rightarrow \chi)$, it is still possible that $\mathcal{R} \nvDash_w K_a\chi$, since there might be an impossible world i such that wR_ai where $i \in V(\varphi)$ and $i \in V(\varphi \rightarrow \chi)$ but $i \notin V(\chi)$.

Moreover, notice that the introduction of impossible worlds does not imply that "everything goes". We can, as we have already done for **K**, propose a strengthening of this logical framework in order to meet at least the essential properties of the knowledge operator. Exactly as we have argued in the first part of this article, the minimal requirement for a logic of knowledge seems to be that if we know a proposition, then this very proposition is true. Again, if we impose

that the accessibility relation is reflexive, then we obtain a logic where the formula (T) $K_a\varphi \rightarrow \varphi$ is a valid formula. In this way we can define the new logic **T'**, obtained by considering only those Rantala-structures where the accessibility relation between worlds is reflexive.

5 An evaluation of the Rantala-semantics strategy

In this last part I shall draw some consequences from the previous analysis and try to evaluate if the Rantala-semantics that we have defined provides a minimal logical framework to describe the formal properties of knowledge. Firstly, I argue that it is possible to identify two reasons to believe that the Rantala-semantics actually describes a valid minimal logic of knowledge. Then I will consider two objections. While one will result to be only an apparent critique to the Rantala-semantics strategy, the second one will identify a true limit of it.

(i) A first observation is that the logic **T'** that we have defined actually provides the minimal logical framework for knowledge which we were looking for. On the one hand, the principle (T) $K_a\varphi \rightarrow \varphi$ results to be a valid formula in this system: working in **T'** we can represent the fact that if an agent knows a proposition, then that proposition is true. On the other hand, the logic **T'** does not force us to accept the closure principle, since (FCP) is not a valid formula in it. Therefore, epistemologists with different theories about knowledge can all accept the modal system **T'** as a minimal framework, which reflects only those properties of knowledge which are unanimously recognized.

(ii) Moreover, the Rantala-semantics is sufficiently flexible to provide not only a minimal common framework, but also a basis suitable for further developments. Given the minimal logic **T'**, it is possible to obtain systems with new axioms or inference rules imposing new conditions on the accessibility relation R_a or on the evaluation function V[9]. In this way, the Rantala-semantics can be used also to represent more complex theories of knowledge, in which more principles hold and should be treated as valid formulas. Epistemologists of different philosophical orientations will thus share the common framework given by **T'**, and they will also be able to describe more complex and rich systems without the need of describing a new and different semantics. Even if **T'** is a quite general and minimal system, we can start from it and obtain step by step new and stronger logics, which will formalize richer and more complex accounts of knowledge.

(iii) However, one aspect of the Rantala-semantics that some philosophers may find problematic is the fact that it makes use of impossible worlds. In fact, even if we accept to work with the framework of possible worlds of the Kripke-

[9]Compare with ´ family=Wansing, familyi=W., given=Heinrich, giveni=H., (1990), who also presents some examples of restriction.

structures, the introduction of impossible worlds poses some new problems. Indeed, although possible worlds represent sets and combinations of facts and events that are not actual, they are still consistent with the laws of classical logic. Differently, it is not straightforward to account for worlds where the most evident logical contradictions may hold. In an impossible world both a proposition and its negation might be true, two disjuncts can be true and the entire disjunction false, and so on. Nevertheless, even if impossible worlds surely present paradoxical features, I think that this problem is only apparent.

Firstly, as Nolan (2013, p. 367–370) underlines, almost every metaphysical theory about the possible worlds can be extended in order to account also for the impossible ones. The only theory which has some problems while explaining the nature of impossible worlds is modal realism, which regards possible worlds as entities really existing. However, there are also some attempts to extend the modal realist perspective in order to describe impossible worlds[10]. Moreover, one may also decide to follow an alternative direction and to consider the useful theoretical role of the impossible worlds a valid reason to reject modal realism and to defend another metaphysical perspective also in regards of the "normal" possible worlds.

Furthermore, it is not obvious at all that the introduction of impossible worlds in epistemic logic forces us to take an explicit position about their metaphysical nature[11]. In fact, the specific philosophical problems that a modal logic raises are linked to the informal interpretation that we decide to give of its operators. For instance, if we read the box symbol as representing necessity, then we have to clarify what does it mean that a proposition is necessary in a world w if and only if it is true in every possible world which is accessible from w. An analysis of the nature of possible world is essential, in this case, in order to make sense of the metaphysical interpretation of the system of modal logic that we are considering. However, if the reading that we are adopting is epistemic, we do not need to take such a metaphysical attitude. As we have already said defining the Kripke-structures, the label of possible world can be substituted without any problem with the one of "scenario". Indeed, the possible and impossible worlds are only the combinations of facts and events that an agent may find plausible descriptions of the reality or not. The informal epistemological reading of the knowledge operator does not call for any metaphysical interpretation. The fact that an agent knows a proposition if and only if that proposition is true in every world to which he/she has access only means that that proposition is part of all the descriptions that he/she considers as possibly valid representations of the reality.

[10] Compare with ´ family=Nolan, familyi=N., given=Daniel P., giveni=D. P., (2013, p. 369).

[11] ´ family=Wansing, familyi=W., given=Heinrich, giveni=H., (1990, p. 536) takes an even stronger position, saying that the question itself about the nature of the impossible worlds is "unsatisfactory".

(iv) Ultimately, despite its many virtues, I think that it is possible to identify a proper limit of the Rantala-semantics strategy. Let us distinguish two different aspects: the failure of the closure principle itself and the explanation of the fact that it does not hold. Depending on what we ask to an epistemic logic, we might then give different evaluations to the Rantala-semantics strategy. On the one hand, as I have already pointed out, the modal logic **T'** offers a formal system where the closure principle of knowledge is not a valid formula. If we adopt **T'**, indeed, we are able to represent many formal properties of knowledge and to potentially adjust the system – working on the accessibility relation and the evaluation function – to meet the characteristics of different epistemological theories. On the other hand, the Rantala-semantics does not provide an explanation of why the closure principle fails. Or, even worse, one may argue that it actually gives a *wrong* explanation of this fact. Indeed, the "cause" that determines the failure of (FCP) in the Rantala-semantics is the introduction of the impossible worlds. If we try to interpret this formal aspect from an epistemological perspective, the result is that the epistemic closure principle does not hold because the agent consider as plausible descriptions of the reality also scenarios where the laws of logic do not hold. However, the problem is that this is not the explanation that the epistemologists who refuse closure – notably Dretske and Nozick – have provided. Therefore, even if it offers a framework that can be accepted also by the epistemologists who do not accept the closure principle, the Rantala-semantics do not reflect in any way their intuitions about why this principle does not hold[12].

Finally, trying to sum up the considerations developed in this last part, it is possible to sketch an evaluation of the Rantala–semantics strategy. The result that we obtained can be regarded as twofold and it depends on what we ask to an epistemic logic. If we want a strong characterisation of a formal system, such that it reflects all the theoretical features of an epistemological theory, then the Rantala–semantics strategy does not seem to be the right way to account for the problems presented by the closure principle. Still, a more modest attitude is also possible. In fact, we can demand to a formal system only to verify as valid those principles – and only those – which an epistemological theory regards as the formal properties of knowledge. In this light, even if it does not provide any heuristic insight about the failure of (FCP), the Rantala-semantics is an interesting common framework for different epistemological perspectives, which can also be refined and strengthened in further ways.

[12] An interesting contribution on this topic is ´ family=Holliday, familyi=H., given=Wesley H., giveni=W. H., (2015), who directly formalizes the epistemological theories proposed by Dretske and Nozick. Notice, however, that although in this way a formal system gains in heuristic power, it also loses the generality that makes it acceptable by epistemologists with different ideas.

References

´ family=Dretske, familyi=D., given=Fred I., giveni=F. I., *= * (1970). “Epistemic Operators”. In: *The Journal of Philosophy* 67.24, pp. 1007–1023.

´ family=Hendricks, familyi=H., given=Vincent, giveni=V., *= * (2015). “Epistemic Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by ´ family=Zalta, familyi=Z., given=Edward N., giveni=E. N., Metaphysics Research Lab, Stanford University.

´ family=Holliday, familyi=H., given=Wesley H., giveni=W. H., *= * (2015). “Epistemic Closure and Epistemic Logic I: Relevant Alternatives and Subjunctivism”. In: *Journal of Philosophical Logic* 44.1, pp. 1–62.

— (forthcoming). “Epistemic Logic and Epistemology”. In: *Handbook of Formal Philosophy*. Ed. by ´ family=Hansson, familyi=H., given=Sven Ove, giveni=S. O., Springer.

´ family=Luper, familyi=L., given=Steven, giveni=S., *= * (2016). “Epistemic Closure”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by ´ family=Zalta, familyi=Z., given=Edward N., giveni=E. N., Metaphysics Research Lab, Stanford University.

´ family=Moore, familyi=M., given=George E., giveni=G. E., *= * (1939). “Proof of an External World”. In: *Proceedings of the British Academy* 25, pp. 273–300.

´ family=Nolan, familyi=N., given=Daniel P., giveni=D. P., *= * (2013). “Impossible Worlds”. In: *Philosophy Compass* 8.4, pp. 360–372.

´ family=Wansing, familyi=W., given=Heinrich, giveni=H., *= * (1990). “A general possible worlds framework for reasoning about knowledge and belief”. In: *Studia Logica* 49.4, pp. 523–539.

´ family=Wittgenstein, familyi=W., given=Ludwig, giveni=L., *= * (1969). *Über Gewißheit*. Ed. by ´ family=Anscombe, familyi=A., given=Elizabeth, giveni=E., Frankfurt am Main (DEU): Suhrkamp.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Hilbert, Completeness and Geometry

Giorgio Venturi

Abstract. This paper aims to show how the mathematical content of Hilbert's Axiom of Completeness consists in an attempt to solve the more general problem of the relationship between intuition and formalization. Hilbert found the accordance between these two sides of mathematical knowledge at a logical level, clarifying the necessary and sufficient conditions for a good formalization of geometry. We will tackle the problem of what is, for Hilbert, the definition of geometry. The solution of this problem will bring out how Hilbert's conception of mathematics is not as innovative as his conception of the axiomatic method. The role that the demonstrative tools play in Hilbert's foundational reflections will also drive us to deal with the problem of the purity of methods, explicitly addressed by Hilbert. In this respect Hilbert's position is very innovative and deeply linked to his modern conception of the axiomatic method. In the end we will show that the role played by the Axiom of Completeness for geometry is the same as the Axiom of Induction for arithmetic and of Church-Turing thesis for computability theory. We end this paper arguing that set theory is the right context in which applying the axiomatic method to mathematics and we postpone to a sequel of this work the attempt to offer a solution similar to Hilbert's for the completeness of set theory.[1]

Keywords. Hilbert, Axiom of Completeness, Geometry, Axiomatic Method.

[1]This article is the result of a talk given for the workshop "Due giornate di studio sulla filosofia mathematica", held in Milano, June 22nd-23rd, 2011. We would like to thanks the SEminario di Logica Permanente (SELP) for the organization and the Associazione Italiana di Logica e sue Applicazioni (AILA) for the support.

Copyright. 2018 Giorgio Venturi. Published in Italy. Some rights reserved. First published on *Rivista Italiana di Filosofia Analitica Junior* 2:2 (2011), *Philosophy of Mathematics*, pp. 80-102.
Author. Giorgio Venturi, gio.venturi@gmail.com.

1 The Axiom of Completeness

In 1899, after a series of lectures on geometry held at the University of Göttingen[2], Hilbert published a book not only fundamental for the subsequent development of geometry, but also for the way of thinking about and doing mathematics of the century that would shortly thereafter start: the "Foundations of Geometry" (*Grundlagen der Geometrie*)[3].

One of the most innovative aspects of this work is the way of thinking about and using the axiomatic method, which is no longer treated as a hypothetical-deductive method capable of proving theorems from true, self-evident, axioms. On the contrary Hilbert transformed it into a versatile method, useful to investigate the foundations of a science and building independence proofs among its axioms.

The system of axioms that Hilbert sets up in the *Grundlagen der Geometrie* is divided into five groups. In order: connection, order, parallels, congruence and continuity. We have two axioms of continuity: Archimedes's axiom and the Axiom of Completeness.

We now want to analyze the latter and try to understand what led Hilbert to formulate this axiom and why it occupies such an important role in the whole construction of the foundation of geometry.

> To the preceeding five groups of axioms, we may add the following one, which, although not of a purely geometrical nature, merits particular attention from a theoretical point of view[4].

Moreover Hilbert argues that the Axiom of Completeness "*forms the cornerstone of the entire system of axioms*"[5].

In the first German edition of 1899 there is no trace of the Axiom of Completeness. It appears from the second, in 1903, to the sixth, in 1923, in the following form:

> V.2 (Axiom der Vollständigkeit) Die Elemente (Punkte, Geraden, Ebenen) der Geometrie bilden ein System von Dingen, welches bei Aufrechterhaltung sämtlicher genannten Axiome keiner Erweiterung mehr fähig

[2] See (Toepell 1986a) and (Hallet and Majer 2004), for a precise exposition of the origins of the *Grundlagen der Geometrie* and of the development of Hilbert's reflections on geometry in this early period.

[3] When referring and quoting it we will use (Hilbert 1899) to indicate the first German edition and (Hilbert 1900F) for the first French edition. Otherwise (Hilbert 1950) refers to the first English edition, translated from the second German edition (Hilbert 1903G), while (Hilbert 1971) indicates the second English edition, translated from the tenth German edition (Hilbert 1968). However, when quoting from (Hilbert 1971), we will point the German edition where the quote first appeared. Moreover, when quoting (Hilbert 1950), we will indicate if the quote can be found also in (Hilbert 1899).

[4] (Hilbert 1950, p. 15).

[5] (Hilbert 1971, p. 28); original emphasis. From the seventh German edition onward.

> ist, d.h.: zu dem System der Punkte, Geraden, Ebenen ist es nicht möglich, ein anderes System von Dingen hinzuzufügen, so dass in dem durch Zusammensetzung entstehenden System sämliche aufgeführten Axiome I-IV, V 1 erfüllt sind[6].

The axiom, however, appeared in print for the first time in the French edition, in 1900, in the following form.

> Au système de points, droites et plans, il est impossible d'adjoindre d'autres êtres de maniére que le système ainsi généralisé forme une nouvelle géométrie oé les axiomes des cinq groupes I-V soient tous vérifiés; en d'autres termes: les elèments de la Géométrie forment un système d'êtres qui, si l'on conserve tous les axiomes, n'est susceptible d'aucune extension[7].

There is also an axiom of completeness for the axiomatization of real numbers in *Über den Zahlbegriff*, published in 1900.

> IV.2 (Axiom of Completeness) It is not possible to add to the system of numbers another system of things so that the axioms I, II, III, and IV 1 are also all satisfied in the combined system; in short, the numbers form a system of things which is incapable of being extended while continuing to satisfy all the axioms[8].

Furthermore, from the seventh edition onward the Completeness Axiom is replaced by a Linear Completeness Axiom, which in the context of the other axioms implies the Axiom of Completeness in the apparently more general form.

> V.2 (Axiom of Line Completeness) It is not possible to extend the system of points on a line with its order and congruence relations in such a way that the relations holding among the original elements as well as the fundamental properties of the line order and congruence following from Axioms I-III and from V.1 are preserved[9].

The literal translation of the Axiom of Completeness is the following.

> V.2 (Axiom of Completeness) The elements (points, straight lines, planes) of geometry form a system of things that, compatibly with the other

[6] (Hilbert 1903G, p. 16).
[7] (Hilbert 1900F, p. 25).
[8] (Hilbert 1900a, p. 1094) in (Ewald 1996). In German: *IV.2 (Axiom der Vollständigkeit) Es ist nicht möglich dem Systeme der Zahlen ein anderes System von Dingen hinzuzufügen, so dass auch in dem durch Zusammensetzung entstehenden Systeme bei Erhaltung der Beziehungen zwischen den Zahelen die Axiome I, II, III, IV.1 sämtlich erfüllt sind; oder kurz: die Zahlen bilden ein System von Dingen, wleches bei Aufrechterhaltung sämtlicher Beziehungen und sämtlicher aufgeführten Axiome keiner Erweiterung mehr fähig ist.*
[9] In (Hilbert 1971, p. 26).

> axioms, can not be extended; i.e. it is not possible to add to the system of points, straight lines, planes another system of things in such a way that in the resulting system all the axioms I-IV, V.1 are satisfied.

In order to set about analyzing the content of this axiom, there are some terms that need to be clarified: Axiome, Dingen, Geometrie. The clarification of the concepts related to these terms will be used to explain Hilbert's axiomatic approach to the foundations of geometry and the central role that the Axiom of Completeness plays in this respect. We do not want to trace the history of these terms, but to investigate the role that these concept played in that historical context.

1.1 Axiome

It is easy to imagine that the concept of axiom in Hilbert's thought mirrors his use of the axiomatic method.

> The procedure of the axiomatic method, as it is expressed here, amounts to a *deepening of the foundations* of the individual do- mains of knowledge[10].

This *deepening of the foundations* amounts in an analysis of the basic principles of a theory that are formalized by means of axioms. The goal of the axiomatic method is to answers questions about why certain theorems can be proved with those principles and others can not[11]. But how it is possible to link axiomatic analysis and mathematical practice? In other words, where does the meaning of the axioms come from? In this first period of reflections on foundational issues[12], Hilbert seems to offer a point of view so far from a formalist conception of mathematics that it may be almost seen at odds with a modern approach.

> These axioms may be arranged in five groups. Each of these groups expresses, by itself, certain related fundamental facts of our intuition[13].

This quote is partly the result of an immature reflection on the sources of knowledge in geometry[14], but it also springs from a notion of intuition that is

[10] (Hilbert 1918, p. 1109) in (Ewald 1996).

[11] In a letter to Frege, dated December 29th, 1899 (in (Frege 1980, pp. 38-39)) Hilbert wrote: *I wanted to make possible to understand and answer such questions as why the sum of the angles in a triangle is equal to two right angles ad how this fact is connected with the parallel axiom.*

[12] In the second period of Hilbert's foundational studies, whose beginning can be placed in the early twenties, with the beginning of the proof theory, we can see an evolution of the concept of axiom. Yet even in this second period, the label "formalist" does not match with his concept of mathematical practice. See (Venturi) in this respect.

[13] (Hilbert 1950, p. 2). Also in (Hilbert 1899).

[14] For a detailed study of the origins and the early influences on Hilbert's conception of geometry see (Toepell 1986a), (Toepell 1986b) and (Toepell 2000).

not the empirical intuition of space, as in the Euclidean formulation. Although recognizing the intuition of space as the starting point of any geometrical reflection, Hilbert maintains that it is not the ultimate source of meaning and truth of geometrical propositions. A different notion of intuition leads Hilbert to argue that the analysis of the foundations of geometry consists of "a rigorous axiomatic investigation of their [of the geometrical signs] conceptual content"[15]. As a matter of fact Hilbert is explicit in recognizing that the axioms of geometry have different degrees of intuitiveness.

> *A general remark on the character of our axioms I-V* might be pertinent here. The axioms I-III [incidence, order, congruence] state very simple, one could even say, original facts; their validity in nature can easily be demonstrated through experiment. Against this, however, the validity of IV and V [parallels and continuity in the form of the Archimedean Axiom] is not so immediately clear. The experimental confirmation of these demands a greater number of experiments.[16].

Accordingly, Hilbert's notion of axiom, even if it is deeply linked with intuition, does not have the evident character that it had classically. We cannot find in Hilbert the substantial coincident between intuition and evidence, that in Euclid's conception of geometry was based on the notion of spatial intuition. In this modern formulation, axioms draw their meaning from a kind of intuition that we can define *contextual*. It is an intuition encoding the *modus operandi* that is obtained working in a field of research, in this case geometry.

We can find an antecedent of this kind of intuition in Klein's words:

> Mechanical experiences, such as we have in the manipulation of solid bodies, contribute to forming our ordinary metric intuition, while optical experiences with light-rays and shadows are responsible for the developement of a 'projective' intuition[17].

However a different conception of the axiomatic method and of a formalistic treatment of mathematics[18] will lead Klein to a different approach to geometry. Indeed Klein's geometrical enquires and the Erlangen's Programme will always presuppose an uncritical treatment of the intuitive data on the nature of space, contrary to the basic principle that aims Hilbert's axiomatic method. Indeed, while Klein will try to analyze and classify the different kind of spaces, Hilbert will deal with intuitions prior to the concept of space. We will come back later to this point, while explaining the different stages that Hilbert saw in the development of a science.

[15] (Hilbert 1900, p. 1101) in (Ewald 1996).
[16] (Hilbert *1898-1899, p. 380) in (Hallet and Majer 2004).
[17] In (Klein 1897, p. 593).
[18] On this subject see (Torretti 1984).

In the preface to the *Grundlagen der Geometrie*, Hilbert is explicit in pointing out the requirements that a system of axioms must meet to be considered a good presentation of a theory.

> The following investigation is a new attempt to choose for geometry a *simple* and *complete* [*vollständiges*] set of *independent* axioms and to deduce from these the most important geometrical theorems in such a manner as to bring out as clearly as possible the significance [*Bedeutung*] of the different groups of axioms and the scope of the conclusions to be derived from the individual axioms. [19].

Here we see clearly that the meaning of the axioms is related to the technical tools they provide, as they are used in proving geometric theorems. This meaning is therefore intrinsic to the context of the theory.

Hilbert thus requires that a formal system be simple, complete and independent. We will consider later the meaning of completeness; however it is now useful to note that a certain idea of completeness is related to the requirement that the system of axioms should be able to prove *all* important geometrical theorems. Moreover, as shown by mathematical practice, the more the ideas are simple, the more they are deep and fundamental[20]. Finally, the demand for independence is for Hilbert a necessary condition for a good application of the axiomatic method. Indeed, for Hilbert the independence of a system of axioms is an index of the depth of the principles expressed by the axioms[21].

We still have to explain what accounts for the truth of the axioms. The answer to this question is clearly shown in a letter to Frege[22] in the form of the well-known equation that Hilbert saw between coherence, truth and existence.

Once shown that the criterion of existence is identified with that of consistency, we still need to clarify what in Hilbert's view exists and how.

[19] (Hilbert 1950, p. 1). Also in (Hilbert 1899).

[20] We will not discuss here the problem of simplicity, although it is partially linked to that of purity of the methods we will address later. In the *Mathematische Notizbücher* (Hilbert *1891) Hilbert writes: *The 24th problem in my Paris lecture was to be: Criteria of simplicity, or proof of the greatest simplicity of certain proofs. Develop a theory of the method of proof in mathematics in general. Under a given set of conditions there can be but one simplest proof. Quite generally, if there are two proofs for a theorem, you must keep going until you have derived each from the other, or until it becomes quite evident what variant conditions (and aids) have been used in the two proofs. Given two routes, it is not right to take either of these two or to look for a third; it is necessary to investigate the area lying between the two routes.* As can be seen from this quote, the problem of simplicity is linked to what would be the development of Hilbert's proof theory; but this would lead us too far from the historical period we are examining. On this subject see (Thiele 2003).

[21] Notice however that the system of axioms proposed by Hilbert was not entirely independent. A truly independent system of axioms for geometry, but not categorical, will be proposed in 1904 by Oscar Veblen in (Veblen 1904).

[22] Letter from Hilbert to Frege December 29th, 1899; in (Frege 1980).

1.2 Dingen

For Hilbert, the existence of mathematical entities is intimately linked to the axioms of a specific formal system. Hilbert considers the axioms as *implicit definitions* of mathematical objects.

> The axioms so set up are at the same time the definitions of those elementary ideas[23].

The idea behind this position is a clear distinction between formal theory and intuitive theory. The latter refers to any mathematical field of research that features only one subject of enquiry and homogeneous methods.

Hilbert is explicit in saying that the axiomatic method leads to a more general conceptual level.

> According to this point of view, the method of the axiomatic construction of a theory presents itself as the procedure of the mapping [*Abbildung*] of a domain of knowledge onto a framework of concepts, which is carried out in such a way that to the objects of the domain of knowledge there now correspond the concepts, and to statements about the objects there correspond the logical relations between the concepts[24].

It is important here to stress that for Hilbert the mathematical objects defined by the axioms of the *Grundlagen der Geometrie* are not strictly speaking geometrical objects but conceptual entities that can be interpreted as geometrical objects. The intended interpretation is of course that of geometry, but this does not narrow the range of possible interpretations that can be give to formulas that constitute the formal system.

We then can see three distinct levels of things: 1) empirical entities 2) formal objects 3) elementary ideas. This distinction mirrors the evolutive steps of a theory that we will see in the next paragraph.

This distinction explains the Kantian exergue that Hilbert places at the beginning of the *Grundlagen der Geometrie*: *All human knowledge begins with intuitions, thence passes to concepts and ends with ideas*[25].

One of the main problem of a formal treatment of a theory is to explain why the axiomatic system so constructed should be a good formalization of the intended intuitive theory. This is the content of an objection raised by Frege.

> Your system of definitions is like a system of equations with several unknowns, where there remains a doubt whether the equations are

[23] (Hilbert 1900, p. 1104) in (Ewald 1996).
[24] (Hilbert *1921-1922, p. 3). Translation in (Hallet 2008).
[25] (Hilbert 1950). Also in (Hilbert 1899).

> soluble and, especially, whether the unknown quantities are uniquely determined. If they were uniquely determined, it would be better to give the solutions, i.e. to explain each of the expressions 'point', 'line', 'between' individually through something that was already known. Given your definitions, I do not know how to decide the question whether my pocket watch is a point. The very first axiom deals with two points; thus if I wanted to know whether it held for my watch, I should first have to know of some other object that is was a point. But even if I knew this, e.g. of my penholder, I still could not decide whether my watch and my penholder determined a line, because I would not know what a line was[26].

The objection is justified on the basis of Frege's studies on the foundations of geometry. Indeed, he acknowledged that the axioms were self-evident propositions and that geometrical objects were abstractions of empirical objects. Frege's critic, however, is easily rebutted by Hilbert[27]. In fact he argues that that was exactly the strength of his method: to establish a formal system able to define an abstract concept, which would respond only to the requirements imposed by the axioms.

> This is apparently where the cardinal point of the misunderstanding lies. I do not want to assume anything as known in advance; I regard my explanation in sec. 1 as the definition of the concepts point, line, plane - if one ads again all the axioms of groups I to V as characteristic marks. If one is looking for another definitions of a 'point', e.g. through paraphrase in terms of extensionless, etc., then I must indeed oppose such attempts in the most decisive way; one is looking for something one can never find because there is nothing there[28].

The problem with Hilbert's reply is that it just points out a distinction of levels but does not give an explanation to the problem implicit in Frege's objection. We will call it Frege's problem and we formulate it as follows: why is the axiomatic system presented by Hilbert in the *Grundlagen der Geometrie* to be considered an axiomatization of geometry? In other words, if the axioms formalize the fundamental ideas of a theory and they are what allow the most important geometrical facts to be proved, what are the criteria that allow to identify the class of theorems we are interested in axiomatizing as theorems of geometry? And finally: in Hilbert's view, what is the definition of geometry once the axiomatic method has cut off the link between formalization and spatial intuition?

[26] Letter from Frege to Hilbert January 6th, 1900; in (Frege 1980, p. 45).

[27] Or at least this is what Hilbert would have answered, because he chose not to replay. Anyway next quote is from the letter just before the one just quoted; and we can assume that if Hilbert did not wrote Frege back is because he had already made his point.

[28] Letter from Hilbert to Frege December 29th, 1899; in (Frege 1980, p. 39).

2 Completeness

In order to understand the meaning of the Axiom of Completeness we promised to explain the meaning of the terms involved. However, the main thesis of this paper is that it is not possible to understand what Hilbert's conception of geometry was without explaining the role that the Axiom of Completeness has in the process of its axiomatization.

If we undertake the difficult task of clarifying the ideas of an author far from us in time, and in the progress of the discipline he contributed to, some methodological precautions are necessary. First of all we must avoid the use of contemporary conceptual results in anachronistic contexts. As a matter of fact, understanding the genesis of concepts means going back to the time when those ideas were not clear, not completely understood. For this reason an historical analysis of this kind, even when it is precise and competent, risks obscuring not only the intentions of those who went through that experience, but the scope and extent of the ideas that are investigated. So, the analysis we would like to pursue here aims to contextualize the choices made by Hilbert as regards foundations of geometry, without altering the originality of those ideas. We therefore propose to go to the root of the problems that Hilbert addressed, trying to understand the mathematical choices and also to unravel the philosophical ideas that moved them.

We assume as our methodological stance that concepts do not proceed in a straight line of reasoning, but they get more and more clear once they are used in solving problems. In this way, ideas and conceptions at first vague are modeled on solutions given to problems. These concepts then become indispensable tools for the discipline that uses them, so that they cannot be disregarded if we want to understand a certain matter completely. In the exact sciences the historical process is easily mystified in two forms: firstly, a retrospective look tends to discover a linear progression of knowledge, and secondly the narrative of a discipline often proceeds in the opposite direction to the one that led to its formation.

In the case under discussion here it is interesting to see how this idea of completeness, which is still vague in Hilbert's discussions, and for this reason so fruitful, contained both the synthesis and the difficulty of concepts that a few decades after played a crucial role in studies of logic and beyond.

2.1 Completeness of the axioms

Coming back to the concept of geometry, in the lectures on projective geometry in 1891, Hilbert divides geometry in three parts:

> The divisions of geometry.

1. Intuitive geometry.
2. Axioms of geometry.
   (investigates which axioms are used in the established facts in intuitive geometry and confronts these systematically with geometries in which some of these axioms are dropped)
3. Analytical geometry.
   (in which from the outset a number is ascribed to the points in a line and thus reduces geometry to analysis)[29].

There is here an important distinction: the one between geometry and geometries. It is also possible to find this distinction in the *Grundlagen der Geometrie*, but for orthographic reasons it can be found only in the French version of 1900, where in the statement of the Axiom of Completeness we can find the distinction between *Géométrie* and *géométrie* The presence of new additions and comments indicates that Hilbert followed closely the editing of this translation[30]. From now on, with Geometry we mean the intuitive theory that is the object of formalization in the *Grundlagen der Geometrie*.

Hilbert's emphasis on analytic geometry stems from its importance in geometrical investigations at that time, as, for example, in Klein's representations of geometries as groups of transformations over manifolds. However, Hilbert's goal is not analyze the nature of space, as Klein did, but to make an axiomatic inquire of our geometrical intuitions. These intuitions are prior to the concept of space and hence they cannot presuppose anything about it. Indeed few years later Hilbert sharpens his reflections on the general concept of a mathematical theory and he says that

> Usually, in the story of a mathematical theory we can easily and clearly distinguish three stages of development: naïve, formal and critical[31].

Then, for geometry, Hilbert's task is to analyze critically the continuity assumption hidden in the intuition of space.

In (Hilbert 1903), Hilbert too contributed to the clarification of the nature of the space, assuming continuity since the beginning. However, since a foundation and not just a classification was sought in the *Grundlagen der Geometrie*, Hilbert sees his work as a contribution to the *kritische* stage of the development of geometry. Thus, following the basic principle of the axiomatic method of deepening the foundations, Hilbert tries to elucidates the more fundamental principles of Geometry.

[29] (Hilbert *1891, p. 3).

[30] In the volume (Hallet and Majer 2004) there is a careful account of the editorial vicisitudes of the French translation.

[31] In (Hilbert 1903a, p. 383) in (Hilbert 1970b). In German: *In der Geschichte einer mathematischen Theorie lassen sich meist 3 Entwicklungsperioden leicht und deutlich unterscheiden: Die naive, die formale und die kritische.* My translation.

Here is outlined one of the most difficult tasks of Hilbert's axiomatization of Geometry: to find a system of axioms able to formalize all the means, also analytical, used in geometrical proofs. Linked to these problems, there are considerations on the purity of method, but we will face them later. Here it is sufficient to say that Hilbert is not concerned with problems of uniformity of methods of proofs[32].

In the same lectures on projective geometry we can find the following sentence, which still suffers from a conception that shortly thereafter would be radically changed.

> Geometry is the theory about the properties of space[33].

However, in Hilbert's lectures for the summer semester, in 1984, entitled *Die Grundlagen der Geometrie* there is no longer an explicit definition of geometry, but rather of geometrical facts. It is also worth noting that in the 1899 *Grundlagen der Geometrie* we do not find a definition of space.

> Among the phenomena, or facts of experience that we take into account observing nature, there is a particular group, namely the group of those facts which determine the external form of things. Geometry concerns itself with these facts[34].

Here there is a subtle, but basic, shift in addressing the problem of a foundation for geometry. Hilbert is not trying to define what Geometry is by means of the axioms, on the contrary he just tries to find a simple, independent and consistent system of axioms that allows a formalization of all geometrical facts. The completeness of the axioms to which Hilbert refers at the beginning of the *Grundlagen der Geometrie* has therefore to be understood in the sense of maximizing the class of geometrical facts that can be proved thanks to the proposed system of axioms.

In 1894, Hilbert was explicit in describing the goals he wanted to achieve by means of his foundational studies.

> Our colleague's problem is this: what are the *necessary* and *sufficient*[35] conditions, independent of each other, which one must posit for a system of things, so that every property of these things corresponds to a geometrical fact and vice versa, so that by means of such a sys-

[32] This is a concern typical of a classical conception of the axiomatic method that dates back to Aristotele: "[…] we cannot in demonstrating pass from one genus to another. We cannot, for instance, prove geometrical truths by arithmetic" (Posterior Analytics: 75a29-75b12). For an historical survey of this subject see (Detlefsen 2008).

[33] (Hilbert *1891, p. 5).

[34] (Hilbert *1894, p. 7).

[35] My emphasis.

> tem of things a complete description and ordering of all geometrical facts is possible[36].

Hilbert's statement of intent is clear: find necessary and sufficient conditions to describe every geometrical fact. Then the problem of defining geometry disappears, since it is implicitly and extensionally defined by geometrical facts. This is precisely the purpose of an analysis conducted with the axiomatic method. As a matter of fact, in 1902, Hilbert says:

> I understand under the axiomatical exploration of a mathematical truth [or theorem] an investigation which does not aim at finding new or more general theorems being connected with this truth, but to determine the position of this theorem within the system of known truths in such a way that it can be clearly said which conditions are necessary and suffcient for giving a foundation of this truth[37].

Thanks to this precise statement, we can make some general consideration on the axiomatic method. First of all, this method is primarily designed to formalize an already developed field of knowledge. Therefore it is a method that can be applied when a science has already reached a sufficient level of maturity, such that it can be divided from other branches of knowledge. Then it is possible to develop an intuition internal to the theory capable of identifying the class of facts that have to be axiomatized, together with the basic principles that allow their proofs. Moreover, it should be noted that Hilbert says explicitly that the goal of the axiomatic method is a clear understanding of geometrical proofs, thanks to the analysis of the *meaning* of the axioms[38], and not the discovery of new theorems.

Besides, Hilbert does not consider the axiomatic method primarily as a source of mathematical rigor, capable of giving an epistemological foundation for mathematical knowledge[39], but rather as a tool which allows us to answer why some proofs are possible and some others are not.

One of Hilbert's greatest achievements in the field of the foundational studies has been to recognize not only the distinction of levels between theory and metatheory, but also to understand that the metatheory was analyzable with mathematical tools. However, Hilbert considered meta-mathematical investigation as a deepening of knowledge about mathematics, and not as a genuine source of new results; contrary to his subsequent work and what the development of twentieth-century logic would show[40].

[36] (Hilbert *1894, p. 8).

[37] (Hilbert 1902-1903, p. 50).

[38] Recall the quote from the introduction of the *Grundlagen der Geometrie* (p. 1), where Hilbert declares that the aim of the book is "*to bring out as clearly as possible the significance [*Bedeutung*] of the different groups of axioms*".

[39] See (Ogawa 2004) in this respect.

[40] Following this line of reasoning it is perhaps reasonable to find an explanation for Hilbert's mild

In 1908, Hilbert still express opinions similar to those of 1902.

> In the case of modern mathematical investigations, ...I remember the investiga tions into the foundations of geometry, of arithmetic, and of set theory-they are concerned not so much with proving a particular fact or establishing the correctness of a particular proposition, but rather much more with carrying through the proof of a proposition with restriction to particular means or with demonstrating the impossibility of such a proof[41].

If the main point in axiomatizing Geometry is the axiomatization of all geometrical facts, what distinguishes them from other facts, whether empirical or mathematical? Hilbert answers this question clearly, but he is not clear on what motivates his choice; and it is on this terrain that Frege's problem regains strength.

2.2 Axiom der Vollständigkeit

Hilbert's aim is to find necessary and sufficient conditions to prove all relevant geometrical fact. So it is possible to define Geometry as the field of knowledge whose true propositions are the theorems that can be proved by means of the axioms presented in the *Grundlagen der Geometrie*.

As we saw in the last paragraph Hilbert's critical investigation of our geometrical intuitions should also take care of the continuity principles that are deeply linked with our intuition of space. This partially explains Hilbert's attention to analytical geometry. Judson Webb, in (Webb 1980), suggests that Hilbert's goal was to free Geometry from analytical considerations, in order to restore its dignity and autonomy. However, more than historical or methodological observations, there is also another reason that led Hilbert to deal with analytic geometry and in particular with analysis.

Hilbert talks explicitly of the "introduction of the number [*Einführung der Zahl*]", within Geometry, and its goal seems to be the arithmetization[42] of Geometry in the axiomatic context. Moreover, following his concept of axioms, as revealing their meaning in the demonstrative use, Hilbert's aim was to formalize analytical tools by means of geometrical axioms.

As a matter of fact, logic and analysis always play an important role in Hilbert's foundational work. In 1922, Hilbert expresses this view in these terms:

reaction to Gödel's incompleteness theorems. However, the quotes above are from the first period of Hilbert's interest on foundational issues i.e. before the twenties; while Gödel's theorems where proved in 1930.

[41] (Hilbert 1909, p. 72). Translation in (Ogawa 2004, p. 100).

[42] By arithmetic Hilbert means analysis and in this sense we use the expression "arithmetization of geometry".

> This circumstance corresponds to a conviction I have long maintained, namely, that a simultaneous construction of arithmetic and formal logic is necessary because of the close connection and inseparability of arithmetical and logical truth[43].

The foundational view proposed here by Hilbert is radically different from the standard one that tries to ground all mathematical knowledge on a single concept. This is what Frege and Russell tried to do with logic; or how a set theoretical, functional or categorical foundation of mathematics is interpreted in modern times. Rather Hilbert was convinced that the tools offered by logic and arithmetic were essential for a proper development of any branch of mathematics. In other words, Hilbert does not seem to have any ontological or epistemological commitments in using numbers and logic; rather it is a methodological concern[44].

> In all exact sciences we gain accurate results only if we introduce the concept of number[45].

However, according to Hilbert these tools must be investigated in a critical manner.

> But if science is not to fall into a bare formalism, in a later stage of its development it has to come back and reflect on itself, and at least verify the basis upon which it has come to introduce the concept of number[46].

In order to introduce the concept of number in Geometry, Hilbert defines a calculus of segments and then he uses the axiomatic method to show which algebraic properties of the calculus follow from the validity of geometrical propositions.

Here the axiomatic method is used with the aim of understanding the demonstrative role of the axioms of Geometry. The idea is to generate a coordinate system internal to Geometry, showing that some fundamental theorems implie

[43] (Hilbert 1922, pp. 1131-1132) in (Ewald 1996).

[44] This is why it is not easy to attribute any philosophical position to Hilbert, although the problems he addresses have obvious philosophical implications.

[45] (Hilbert *1894). In German: *In allen exakten Wissenschaften gewinnt man erst dann präzise Resultate wenn die Zahl eingefuhrt ist.* , in (Hallet and Majer 2004, p. 194).

[46] (Hilbert *1898). In German: *Aber wenn die Wissenschaft nicht einem unfruchtbaren Formalismus anheimfallen soll, so wird sie auf einem spateren Stadium der Entwickelung sich wieder auf sich selbst besinnen mussen und mindestens die Grundlagen prufen, auf denen sie zur Einfuhrung der Zahl gekommen ist,* in (Hallet and Majer 2004, p. 194). However, even in this mixture of geometry and analysis we need to be guided by intuition. In (Hilbert *1905, pp. 87-88), Hilbert says: *[O]ne should always be guided by intuition when laying things down axiomatically, and one always has intuition before oneself as a goal [*Zielpunkt*].* Translation in (Hallet 2008).

certain properties of numbers that are used as coordinates. In this way, the system of real numbers is not imposed from outside, as in analytic geometry, but arises from geometrical argumentation.

For example, the validity of Pappus's theorem (called Pascal's theorem by Hilbert) is used to show that the multiplication that it is possible to define on the coordinate system must necessarily be commutative. Thanks to axioms I-V1 Hilbert shows that the coordinate system thus defined forms an Archimedean field. However, since this Archimedean field can be countable, it is clear to Hilbert that the geometry that satisfies all axioms I-V1 can not be immediately identified with analytic geometry.

Indeed, the domain of the latter is uncountable, because it makes use of all real numbers. So, Hilbert's major concern is to define axiomatically a bijection between the points of a straight line and the real numbers. The solution of this problem is precisely the mathematical content of the Axiom of Completeness

> If in a geometry only the validity of the Archimedean Axiom is assumed, then it is possible to extend the set of points, lines, and planes by "*irrational*" elements so that in the resulting geometry on every line a point corresponds, without exception, to every set of three real numbers that satisfy the equation. By suitable interpretations it is possible to infer at the same time that *all* Axioms I-V are valid in the extended geometry. Thus extended geometry (by the adjunction if irrational elements) is none other than the ordinary space Cartesian geometry in which the completeness axiom V.2 also holds[47]

In this quotation it is possible to see how the Axiom of Completeness is used to fill that gap between Hilbertian plane geometry and analytic geometry. The way to achieve this is by adding irrational elements to the coordinate system presented in the *Grundlagen der Geometrie*. As a matter of fact, the axiomatization of the real numbers is simultaneous with the introduction of the Axiom of Completeness for geometry[48].

The irrational elements are also called ideal elements, by Hilbert. However, he immediately makes it clear that the ideal character of these elements is only relative the specific presentation of the system[49].

> That to every real number there corresponds a point of the straight line does not follow from our axioms. We can achieve this, however, by the introduction of ideal (irrational) points (Cantor's Axiom). It

[47] (Hilbert 1950, pp. 35-36).

[48] Remember that the Axiom of Completeness first appears in (Hilbert 1900a) and then in the first French edition of *Grundlagen der Geometrie*

[49] In (Hilbert *1919, p. 149), Hilbert says, *The terminology of ideal elements thus properly speaking only has its justification from the point of view of the system we start out from. In the new system we do not at all distinguish between actual and ideal elements.*

> can be shown that these ideal points satisfy all the axioms I-V [...]. Their use is purely a matter of method: *first with their help is it possible to develop analytic geometry to its fullest extent* [50].

The reference to irrational elements echoes the problem of the purity of methods, which is explicitly mentioned by Hilbert. However Hilbert's solution is not to restrict the demonstrative tools, allowing just those conforming to the essential properties of the object of the theory. Indeed, the same idea of an extralogical property of mathematical objects is contrary to the conception of axiomatic method, as Hilbert made clear also in correspondence with Frege.

> In fact, the geometric investigation carried out here seeks in general to cast light on the question of which axioms, assumptions or auxiliary means are necessary in the proof of a given elementary geometrical truth, and it is left up to discretionary judgement [*Ermessen*] in each individual case which method of proof is to be preferred, depending on the standpoint adopted[51].

Since its aim is to show the possibility or the impossibility of a proof, the axiomatic method is the highest expression of the search for the purity of methods. In an interlineated addition to the 1898/1899 lessons Hilbert writes: "Thus, solution of a problem impossible or impossible with certain means. With this is connected the demand for the purity of methods[52]". Hilbert considers the application of the axiomatic method as a precondition for any consideration on the purity of methods. Indeed, thanks to that it is possible to clear necessary conditions for the proof of a mathematical theorem. So, the choice of the demonstrative methods becomes a subjective question, since it does not depend on the nature of the problem.

> This basic principle, according to which one ought to elucidate the possibility of proofs, is very closely connected with the demand for the 'purity of method' of proof methods stressed by many modern mathematicians[53]. At root, this demand is nothing other than a subjective interpretation of the basic principle followed here.[54].

[50] (Hilbert *1899, pp. 166-167).

[51] (Hilbert 1950, pp. 82).

[52] See (Hilbert *1898-1899, p. 284) in (Hallet and Majer 2004).

[53] Remember that Hilbert's proofs were not easily accepted by the mathematical community of the late nineteenth century. Therefore, instead of restricting the methods of proof, Hilbert put forward an analysis of proofs that does not rest on external considerations on the nature of mathematical entities, but that aim to show if a demonstrative tool is necessary in a particular proof. Moreover, given the link between methods of proof and axioms, the justification of the means used in a proof is brought back to the justification of the axioms and to the inference rules. In 1925, in (Hilbert 1925), Hilbert writes: *If, apart from proving consistency, the question of the justification of a measure is to have any meaning, it can consist only in ascertaining whether the measure is accompanied by commensurate success.*

[54] (Hilbert 1950, p. 82).

As a matter of fact Hilbert used analytic geometry to the full in the application of the axiomatic method to Geometry; for example in the proof that it is possible to develop a non-Deserguean geometry. This choice shows also that Hilbert's goal was not a foundation of analytic geometry in the contemporary sense, short of running into an obvious circularity in his reasoning.

In this context we can also explain how the axiomatic method can be used to improve our mathematical knowledge. Remember that Hilbert says: "I understand under the axiomatical exploration of a mathematical truth [or theorem] an investigation which does not aim at finding new or more general theorems"[55].

> This basic principle [to enquire the main possibility of a proof] seems to me to contain a general and natural prescription. In fact, whenever in our mathematical work we encounter a problem or conjecture a theorem, our drive for knowledge [*Erkenntnistrieb*] is only then satisfied when we have succeeded in giving the complete solution of the problem and the rigorous proof of the theorem, or when we recognise clearly the grounds for the impossibility of success and thereby the necessity of the failure[56].

Therefore, we can clearly see in Hilbert's thought a dichotomy between the subjective side of the demonstrative tools and the objective side of the the logical relations between concepts. However, the objectivity of mathematics is not needed to ground the mathematical discourse; indeed, this is done by means of a consistency proof. The emphasis given to the objectivity of mathematics is just a matter of justification of the methods of proof, hence of the axioms. We need to stress here the difference between giving a foundation or a justification. As a matter of fact, if we try to interpret Hilbert's foundational efforts as a modern foundation for a mathematical theory, we see that we ran into an apparent circularity of the argumentation, because analytic geometry is used in order to show the necessity of the axioms that should give a foundation for analytic geometry. Then, this seems to support the autonomy of Hilbert's foundation of mathematics[57]. But, this point of view is incorrect, since a foundation is sought where there is no foundation in the traditional sense. Hilbert does not try to find an epistemological explanation for mathematical arguments, or an ontological classification of mathematical entities, on a mathematical ground. On the contrary he tries to justify the possibility to give a formal treatment of an intuitive theory. Even if Hilbert avoids any extra-logical commitments about objects and methods of proof, however the way he constructs the formal theory for Geometry is not autonomous from extra-mathematical considerations; we can say philosophical. Indeed Hilbert justifies the formalization of a theory appealing to

[55] (Hilbert 1902-1903, p. 50).
[56] (Hilbert 1950, p. 82).
[57] See for example (Franks 2009).

intuition, logical reasoning and the concept of number. These concepts seem to be for Hilbert the starting points for any mathematical knowledge and construction. Appealing to these notions he is able to say that the axioms presented in the *Grundlagen der Geometrie* formalize precisely analytic geometry, in its intuitive character. This choice is indeed philosophical, because it implies a precise definition of mathematics: the science of calculation, carried out by logical means. This conception is quite astonishing if we think of the development of mathematics in the last century. However it explains the role of arithmetic in Hilbert's conception of mathematics, throughout all his work; where arithmetic is here to be understood in the widest sense, including also transfinite cardinal arithmetic.

Recalling that Hilbert's goal was to find necessary and sufficient conditions for proving the more relevant geometrical facts, we can affirm that the axioms of groups I-IV, together with Axiom of Archimedes, are necessary conditions for the development of analytic geometry, and the Axiom of Completeness plays the role of a sufficient condition to adapt the formal presentation given in the *Grundlagen der Geometrie* to the intuitive idea of a geometrical theory that makes use of the whole class of real numbers. Already in 1872 Cantor felt the need for an axiom to make compatible these two sides of geometry.

> In order to complete the connection [...] with the geometry of the straight line, one must only add an axiom which simply says that conversely every numerical quantity also has a determined point on the straight line, whose coordinate is equal to that quantity [...] I call this proposition an *axiom* because by its nature it cannot be universally proved. A certain objectivity is then subsequently gained thereby for the quantities although they are quite independent of this[58].

So we can distinguish two different kinds of axioms: the ones that are *necessary* for the development of a theory and the *sufficient* ones used to match intuition and formalization.

In the lectures that precede the first edition of the *Grundlagen der Geometrie* Hilbert proposed that continuity be formalized, in ways similar to Cantor's[59] and Dedekind's[60], which were able, together with the other axioms, to guarantee the existence of a bijection between the point lying on a straight line and the real numbers. However, Hilbert soon realized that there was need of less continuity for developing Geometry. Thus, following the general principle of the axiomatic method of pointing out the necessary conditions, Hilbert chose the Axiom of Archimedes. Indeed Hilbert's aim was to explain how and why geo-

[58] In (Cantor 1872, p. 128).

[59] (Cantor continuity axiom): every descending (with respect to the relation of inclusion) sequence of non empty real intervals has no-empty intersection.

[60] (Dedekind continuity axiom): given any partition of the real line in two classes $A \leq B$ (i.e. $\forall a \in A$ and $\forall b \in B$, we have $a \leq b$) there is a real number c such that $a \leq c \leq b$, for every $a \in A$ and $b \in B$.

metrical proofs were possible, considering knowledge as knowledge of causes. In a letter to Frege, on December 29th 1899 (in (Frege 1980, pp. 38-39)), Hilbert wrote: "*It was of necessity that I had to set up my axiomatic system:* I wanted to make it possible to understand *those geometrical proposition that I regard as the most important results of geometrical inquiries*" (my emphasis).

> By the above treatment the requirement of continuity has been decomposed into two essentially different parts, namely into Archimedes' Axiom, whose role is to prepare the requirement of continuity, and the Completeness Axiom which *forms the cornerstone of the entire system of axioms*. The subsequent investigations rest essentially only on Archimedes' Axiom and the completeness axiom is in general not assumed[61].

Following this line of reasoning, the Axioms of Completeness can be seen as the first, historically documented, instance of Skolem's paradox; of course Hilbert was not driven by considerations on the nature of logic, but the Axiom of Completeness can be seen as a way of solving the problem of the existence of a theory for analytic geometry that cannot prove that real numbers are uncountable. As a matter of fact Hilbert seems to argue in favor of an intuitive connection with real numbers. Writing against the genetic method that tries to define real numbers, starting with natural numbers, Hilbert says:

> The totality of real numbers, i.e. the continuum [...] is not the totality of all possible series of decimal fractions, or of all possible laws according to which elements of a fundamental sequence may proceed. It is rather a system of things whose mutual relations are governed by the axioms set up and for which all propositions, and only those, are true which can be derived from the axioms by a finite number of logical processes[62].

In other words, this matching of intuition and formalization, which tries to harmonize the intuitions behind the system of real numbers and the real line, is the intuitive content of the fifth group of axioms of the *Grundlagen der Geometrie*.

In conclusion, Hilbert's analysis of the notion of continuity led him to formalize the Axiom of Completeness as a sufficient condition for analytic geometry, in the form of a *maximality* principle.

There are some presuppositions that need to be made explicit in Hilbert's ideas. First of all, the scope of the axiomatization needs to be known right from

[61] (Hilbert 1971, p. 28). From the seventh German edition onward
[62] In (Hilbert 1900, p. 1105) in (Ewald 1996).

the beginning. Moreover, Hilbert chose to include analytical tools in the formalization of Geometry. This choice seems surprising if we consider that at that time the development of Geometry led to the introduction of very remote concepts, not only from classical geometry, but also from considering calculation as the most important tool in Geometry[63]. The answer to this problem may be Hilbert's conviction that "In all exact sciences we gain accurate results only if we introduce the concept of number[64]".

All this shows how important logic and arithmetic are for Hilbert. So, together with the fact that formalization needs to take care of demonstrative methods used in a certain field of knowledge, it explains how Hilbert's ideas developed to the proof theory.

3 Idea of completeness and contemporary axiomatics

Having explained what Hilbert means by completeness and what he was aiming for in placing it at the center of his axiomatic presentation of Geometry, we would like here to study how this idea developed after Hilbert.

We would like to say here that we do not want to explain how the notion of completeness became what we now call semantic completeness, syntactic completeness and categoricity[65]. On the contrary, we would like to see if the idea of a maximal axiom that tries to match intuition and formalization has been used in other contexts.

In the analysis of the foundations of Geometry, Hilbert faced the problem of finding a link between the subjective perception of mathematical reality and the objective character of mathematical truth. However, this link was not fully justified, because he never even try to address the problem of explaining the concept of Geometry. Hilbert's solution is satisfactory as far as the Axiom of Completeness, translated into a modern terminology -with second order logic-, implies the categoricity of the model. However, since it is possible to develop arithmetic in the system of the *Grundlagen der Geometrie*, by the first Gödel's incompleteness theorem, this system is deductively incomplete, with respect to first order logic. But for what concern the sense of completeness we used to explain the Axiom of Completeness, we can say that Hilbert did managed to build a complete system of axioms, i.e. capable to prove all relevant theorems of Euclidean geometry and to formalize all methods of proof used in it. Anyway at that time not only Gödel's results were lacking, but also a good formalization of logic, able

[63] See (Hintikka 1997) for a detailed analysis of the importance of combinatorial aspects in Hilbert's thought.

[64] (Hilbert *1894).

[65] See (Awoday and Reck 2002) in this respect.

to represet the logical tools used in formalizing Geometry.

There is a substantial link between the problem of matching intuition and formalization and the problem of a mathematical treatment of logic. Indeed whenever there is a need to formalize concepts that have intuitive roots, we have to reflect on whether reasoning on these concepts is really possibile; and at the border between subjectivity of judgements and objectivity of truths there is logic.

In this respect the Axiom of Completeness is used to delimit the scope of axiomatization and it witnesses an extra-logical relation with the subject matter of Geometry.

Hilbert's work can be seen as an instance of a more general procedure aiming to establish some necessary conditions for the development of a theory and to find a maximal principle as a sufficient condition for the formalization.

Another example, besides the case of geometry, is the formalization of the concept of computability. In this case the need for a principle capable for completing the theory is really important, since what is formalized is a meta-mathe matical concept. In this context, the analogue of the Axiom of Completeness is Church-Turing thesis. It says that the class of functions defined by the λ-caluculus (equivalently of general recursive functions and of functions computa ble by a Turing machine[66]) is the class of all the functions that are intuitively computable. Then, since all these functions are intuitively computable, Church-Turing thesis is a sufficient condition that characterizes the class of computable functions. There seems to be an important difference between the Axiom of Completeness and Church-Turing thesis, since one is an axiom, but the other a thesis. However the difference is only apparent, because as far as their use in proofs in concerned both serve as a justification of the use of the other axioms. Indeed Hilbert says explicitly that the Axiom of Completeness is not used in his geometrical investigations; exactly as the Church-Turing thesis is not used in proving theorem of recursion theory, but just invoked to justify that all and only those functions are intuitively computable. Again we can see that Church-Turing thesis bridges the gap between the formalization of a concept and the our intuitive idea.

Another example of this kind is the formalization of the concept of natural number by means of the Peano-Dedekind axioms[67]. In this case the presentation is completed by the axiom of induction as a second order principle: given a

[66] Indeed all these classes are provably the same.

[67] Besides the scheme for induction we have:

1. $\forall x(x \neq 0 \rightarrow \exists y(S(y) = x))$,
2. $\neg\exists x(S(x) = 0)$,
3. $\forall x, y(x \neq y \rightarrow S(x) \neq S(y))$.

non empty set M, an element $0 \in M$ and an injective function $S : M \to M$

$$\forall P \subseteq M\big(0 \in P \wedge \forall x(x \in P \to S(x) \in P) \to P = M\big).$$

This axiom says that every subset of M satisfying the axioms and closed under the successor function, must necessarily be the structure of natural numbers. In other words is not possible to extend the system of natural numbers with new objects and to get a new system of things that satisfies the Dedekind-Peano axioms, minus induction.

As in the case of the Axiom of Completeness we are here dealing with a method which, by using second order principles, fixes the structure intended to formalize an intuitive concept uniquely. As for Geometry, by means of these axioms we give a definition of natural number. It is interesting to note that in both situations the result is achieved through the identification of a property which formalizes the demonstrative power of a concept: continuity in the first case, induction in the second.

Therefore, it is interesting to ask whether this axiomatic notion is still relevant and how the progress of logic served to clarify this relationship between intuition and formalism.

4 The case of set theory

In the axiomatic context set theory plays a prominent role. This theory, in fact, was given a satisfactory axiomatization capable of formalizing almost all mathematics, and maintained and improved the ability to analyze up to a minimum the demonstrative tools used in mathematical practice, thanks to versatile methods for building independence proofs.

In the last century the development of mathematics and the invention of category theory have undermine the widespread idea that set theory could be the foundation of mathematics[68]. However if we confine to Hilbert's idea of foundation of a science as outlined in these pages -to apply the axiomatic method in order to find necessary and sufficient conditions- we can say that set theory provide fine tools to analyze the main possibility of proof of a theorem[69] and a unifying language where it is possible to pose any mathematical problem. Hence it is a good framework for applying Hilbert's axiomatic method to mathematics.

Indeed, set theory deals mainly with problems independent of ZFC, the classical first order formalization of the theory. However, the analysis of these prob-

[68] On this topic see MacLane's and Mathias's articles in (Judah, Just and Woodin 1992).

[69] This is also the aim of what is now called Reverse Mathematics, although its main focus are systems that leis in between RCA_0 and second order arithmetic. For this reason in Reverse Mathematics the axiomatic method is applied to theorems about countable structures. So, even if its analysis is finer, its scope is much smaller then that of set theory. See (Marcone 2009), and (Simpson 2009) for a presentation of aims and methods of Reverse Mathematics.

lems does not end with their independence proofs, but seeks to identify which principles are needed for their proofs; as Hilbert did in the case of Geometry. As a consequence these principles often cannot hide their combinatorial origin.

Secondly, as the analysis of the concept of axiom has shown, for Hilbert the idea of completeness is related to the idea of exhaustiveness of the methods of proof. However, the incompleteness phenomenon arising from Gödel's theorems makes it always possible to extend these methods, although in such a way that it is possible to compare them by means of their consistency strength[70].

A further source of difficulty is the fact that set theory uses arguments that, even if formalized in first order logic, are substantially of higher order. For example, the axioms expressing the existence of large cardinals, while affirming the existence of sets with certain first order properties, imply the existence of a model for set theory, or of class-size objects. For this reason a reflection about the methods used in set theory should also take into account a meta-theoretical discussion of logic, not necessarily first order logic[71].

Moreover, if we try to formulate an axiom that makes set theory complete with respect to the intuitive idea of set, one collides with some conceptual difficulties. These are due to the fact that the very concept of set is a mental operation of reducing to unity a plurality of things. Therefore the "set of" operation cannot be limited to a fixed domain, without asking if this latter is itself a set. The history of the axiomatization of the concept of set is in fact a continuing attempt to impose the least restrictive limitations, in order to avoid an inconsistent system; as Russell's paradox showed for Frege's system.

However, even facing these inherent difficulties, the need for an axiom similar to Hilbert's Axiom of Completeness was historically felt quite early in the development of set theory.

In 1921 Fraenkel expressed this idea as follows:

> Zermelo's axiom system do not ensure any character of "categorical" uniqueness. For this reason there should be an "Axiom of Narrowness" similar, but opposite, to Hilbert's Axiom of Completeness, in order to impose the domain to be the smallest possible, compatibly with the other axioms. In this way we can eliminate those classes, existing in Zermelo's system, that are unnecessary for a mathematical purpose[72].

[70] We say that a theory T has consistency strength stronger than a theory S if in first order Peano arithmetic (i.e. the induction axiom is a scheme) it is possible to prove $con(T) \rightarrow con(S)$, where $con(T)$ is the sentence expressing the consistency of T. Surprisingly, and luckily, the theories that are the object of study are linearly ordered, with respect to consistency strength. This order is induced by the one existing among the axioms that postulate the existence of large cardinals. For an overview of this subject see (Kanamori 1994).

[71] For an historical presentation of this problem, see (Moore 1980).

[72] (Fraenkel 1921). In German: *Das Zermelosche Axiomensystem sichert dem Bereich keinen "kat-*

Once noticing that set theory is a good framework for applying the axiomatic method, as Hilbert conceived it, to mathematics, it would be interesting to inquire about the possibility of defining a notion of completeness able to capture some intrinsic aspect of set theory, within an axiomatic framework. In other words, in which way it makes sense to try to reconcile the idea of a complete theory with the phenomenon of incompleteness?

We defer the attempt to answer these questions to another work.

egoriscehn" Eindeutigkeitscharakter. Dazu ist ein weiteres, dem Hilbertschen Vollständigkeitsaxiom umgekehrt analoges "Beschränktheitsaxiom" erforderlich das dem Bereich den kleinsten mit den Axiomen verträglichen Umfang auferlegt. Hierdurch werden verschiedene, für mathematische Zwecke unnötige Klassen von Mengen ansageschieden , die im Zermeloschen System Platz haben. My translation.

Bibliography

Awodey, S. and E. H. Reck (2002). Completeness and categoricity. Part I: nineteenth-century axiomatics to twentieth-century metalogic, *History and Philosophy of Logic* 23, 2002.

Cantor, G. (1872). Über die Ausdehnung eines Satzes aus der Theorie der trignometrischen Reihen, in *Mathematische Annalen*, 5, pp. 123-132.

Corry. L. (2006). "The origin of Hilbert's axiomatic method", in Renn (2006).

Detlefsen, M. (2008). The purity of methods, in (Mancosu 2008).

Dreben, B. and Kanamori, A. (1997). "Hilbert and set theory", *Synthese* 110, 1997.

Ehrlich, P (1997). "From completeness to archimedean completeness: an essay in the foundations of euclidean geometry", *Synthese* 110, 1997.

Ewald, W. ed. (1996). *From Kant to Hilbert: a source book in the foundations of mathematics* Volume II, Clarendon Press, Oxford.

Ewald, W. and W. Sieg, (eds.) (2008). *David Hilbert's lectures on the foundations of logic and arithmetic, 1917-1933*, Springer, Berlin.

Ewald, W. Hallet, M. and W. Sieg, (eds.) *David Hilbert's lectures on the foundations of logic and arithmetic, 1894-1917*, Springer, Berlin. Forthcoming.

Fraenkel, A. (1921). Über die Zermelosche Begründung der Mengenlehre, abstract, in *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 30(2).

Franks, C. (2009). *The autonomy of mathematical knowledge: Hilbert's program revisited*, Cambridge University Press, Cambridge.

Frege, G. (1980). *Philosophical and mathematical correspondence*, Basil Blackwell, Oxford.

Hallet, M. (2008). Reflections on the purity of method in Hilbert's *Grundlagen der Geometrie*", in (Mancosu 2008).

Hallet, M. and U. Majer, (eds.) (2004). *David Hilbert's lectures on the foundations of geometry, 1891-1902*, Springer, Berlin.

Hendricks, V. F. and al. (eds.) (2007). *Interactions. Mathematics, physics and philosophy, 1860-1930*, Springer, New York.

Hilbert, D. *Mathematische Notizbücher*, 3 notebooks, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Handschriftenabteilung, Cod. Ms. D.

Hilbert 600:1-3.

Hilbert, D. (*1891). Projektive Geometrie, lectures notes for a course held during the year 1891, in Könisberg. In (Hallet and Majer 2004) pp. 21-55.

Hilbert, D. (*1894). Die Grundlagen der Geometrie, lectures notes for a course held during the year 1894, in Könisberg. In (Hallet and Majer 2004), pp. 72-123.

Hilbert, D. (*1898). Feriencursus: Über den Begriff des Unendlichen, lectures notes for a course held during the years 1898, in Göttingen. In (Hallet and Majer 2004), pp. 160-178.

Hilbert, D. (*1898-1899). Grundlagen der Euklidischen Geometrie, lectures notes for a course held during the years 1898-1899, in Göttingen. In (Hallet and Majer 2004), pp. 221-286.

Hilbert, D. (*1899). Elemente der Euklidischen Geometrie, *Ausarbeitung* form (Hilbert *1898-1899). In (Hallet and Majer 2004), pp. 302-395.

Hilbert, D. (1899). *Grundlagen der Geometrie* (*Festschrift*). In (Hallet and Majer 2004), pp. 436-525.

Hilbert, D. (1900F). *Les principes fondamentaux de la géométrie*, Gauthier-Villars, Paris, 1900.

Hilbert, D. (1900). Mathematishen Probleme, 1900. Translated as Mathematical problems in (Ewald 1996), pp. 1096-1105.

Hilbert, D. (1900a). Über den Zahlbegriff. Translated as On the concept of number in (Ewald 1996), pp. 1089-1095.

Hilbert, D. (1902-190303). Über den Satz von der Gleichheit der Basiswinkel im gleichschenkligen Dreieck, in *Proceedings of the London Mathematical Society*, 35, pp. 50-68.

Hilbert, D. (1903G). *Grundlagen der Geometrie*, Zweite, durch Zusätze vermehrte und mit fünf Anhängen versehene Auflage. Teubner, Leipzig.

Hilbert, D. (1903). Über die Grundlagen der Geometrie", *Mathematiche Annalen*, 56, pp. 381-422. Translated in English in (Hilbert 1971), p. 150-190.

Hilbert, D. (1903a). Über die Theorie der algebraischen Invarianten", Mathematical paper read at the international Mathematical Congress Chicago. Printed in (Hilbert 1970b), pp. 376-383.

Hilbert, D. (1905). "Über die Grundlagen der Logik und der Arithmetik", 1905.

Translated as "On the foundations of logic and arithmetic", in (Van Heijenoort 1967).

Hilbert, D. (*1905). Logische Principien des mathematischen Denkens", lecture notes for a course held during the year 1905, in Göttingen. In (Ewald, Hallet and Sieg).

Hilbert, D. (1909). Wesen und Ziele einer Analysis der unendlichvielen unabhängigen Variablen. In (Hilbert 1970c).

Hilbert, D. (1918). Axiomatishes Denken. Translated as "Axiomatic thought", in (Ewald 1996), pp. 1105-1115.

Hilbert, D. (*1919). Natur und mathematisches Erkennen, lecture notes for a course held during the year 1919, in Göttingen. Published for the first time by David Rowe, in 1992.

Hilbert, D. (*1921-1922). Grundlagen der Mathematik, lecture notes for a course held during the academic year 1921-1922, in Göttingen. In (Ewald and Sieg 2008).

Hilbert, D. (1922). Neubegründung der Mathematik. Erste Mittelung. Translated as The new grounding of mathematics. First report, in (Ewald 1996), pp. 1117-1148.

Hilbert, D. (1925). Über das Unendliche. Translated as On the infinite, in (Ewald 1996).

Hilbert, D. (1950). *The Foundations of Geometry*, The Open Court Publishing Company, La Salle. Translated from the 2^{nd} German edition.

Hilbert, D. (1968). *Grundlagen der Geometrie*. Mit Supplementen von Paul Bernays. 10. Auflage. Teubner, Stuttgart.

Hilbert, D. (1970b). *Gesammelte Abhandlungen II*, Springer.

Hilbert, D. (1970c). *Gesammelte Abhandlungen III*, Springer.

Hilbert, D. (1971). *Foundations of geometry*, (Second English ed.), La Salle: Open Court. Translated by the 10^{th} German edition.

Hintikka, J. (ed.) (1995). *From Dedekind to Gödel*, Synthese Library vol. 251.

Hintikka, J. (1997). Hilbert vindicated, *Synthese* 110.

Judah, H. Just, W. and W. H. Woodin (eds.) (1992). *Set theory and the continuum*,

Springer, New York.

Kanamori, A. (1994). *The higher Infinite. Large cardinals in set theory from their beginnings*, Springer, Berlin.

Klein, F. (1897). Gutachen, betreffend den dritten Band der Theorie der Transformationsgruppen von S. Lie anlässlich der ersten Verteilung des Lobatschewsky Preises, *Mathematichen Annalen*, 50, pp. 583-600.

Gabriele Lolli: "Hilbert e la logica", in *Le Matematiche*, LV(1), 2000.

Mancosu, P. (ed.) (2008). *The philosophy of mathematical practice*, Oxford University Press, Oxford.

Majer, U. (1997). "Husserl and Hilbert on completeness. A neglected chapter in early twentieth century foundations of mathematics", *Synthese* 110, 1997.

Majer, U. (2007). "Hilbert's axiomatic approach to the foundations of science - a failed research program?", in (Hendricks 2007).

Marcone, A. (2009). Equivalenze tra teoremi: il programma di ricerca della reverse mathematics, *La Matematica nella Società e nella Cultura, Rivista dell'Unione Matematica Italiana* 2, pp. 101-126.

Moore, G. H. (1980). Beyond first-order logic: the historical interplay between mathematical logic and axiomatic set theory, *History and Philosophy of Logic* 1.

Ogawa, Y. (2004). The Pursuit of Rigor: Hilbert's axiomatic method and the objectivity of mathematics, *Annals of the Japan Association for Philosophy of Science* 12(2).

Ortiz-Hill, C. (1995). "Husserl and Hilbert on completeness", in (Hintikka 1995).

Peckhaus, V. (2003). "The pragmatism of Hilbert's program", *Synthese* 137, 2003.

Peckhaus, V. (2005). "Pro and contra Hilbert : Zermelo's set theories", *Philosophia Scientiae* 9(2), 2005.

Renn, J. and al. (eds.) (2006). *The genesis of general relativity, Vol. 4 Theories of gravitation in the twilight of classical physics: the promise of mathematics and the dream of a unified theory*, Springer, New York, 2002.

Shore, R. (2010). Reverse mathematics: the playground of logic, *The Bulletin of Symbolic Logic*, 16, pp. 378-402.

Simpson, S. C. (2009). *Subsystems of second order arithmetic*, Cambridge Uni-

versity Press, Cambridge.

Thiele, R. (2003). Hilbert's Twenty-fourth problem, in *America Mathematical Monthly*, January.

Toepell, M. (1986a). *Über die Entstehung von David Hilberts "Grundlagen der Geometrie"*, Dissertation, Göttingen. Vandenhoeck & Ruprech.

Toepell, M. (1986b). On the origins of David Hilbert's *Grundlagen der Geometrie*, in *Archive for History of Exact Sciences*, 35(4).

Toepell, M. (2000). The origin and the further development of Hilbert's *Grundlagen der Geometrie*, in *Le Matematiche*, LV(1).

Torretti, R. (1984). *Philosophy of geometry from Riemann to Poincaré*, Springer, New York.

Van Heijenoort, J: *From Frege to Gödel: a source book in mathematical logic, 1879–1931*, Harvard University Press, Cambridge, 1967.

Veblen, O. (1904). A system of axioms for geometry, in *Transaction of the American mathematical society*, 5.

Venturi, G. The concept of axiom in Hilbert's thought, preprint.

Webb, J. C. (1980). *Mechanicism, mentalism and metamathematics*, Kluwer, Boston.

Judson C. Webb: "Hilbert's formalism and arithmetization of mathematics", *Synthese* 110, 1997.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

The Link between Misinterpretation, Intentionality, and Mental Agency in the Natural Language Interpretation of "Fake"

Janek Guerrini

Abstract. In formal semantics of natural language, an intersective interpretation works for many adjectives: x is a French lawyer iff x ∈ {x: x is French} ⋂ {x: x is a lawyer}. For those adjectives for which this does not work, like "excellent", we still have, at worst, a subsective modification ({x: x is an excellent violinist} ⊂ {x:x is a violinist}). Neither of these applies to "fake", whose formal interpretation is a traditional challenge. In this paper, I propose an analysis of the semantics of "fake" in which the speaker's attribution of intentionality (derived or original) to the object or person of which she predicates fakeness is central. In fact, the boundaries between the properties that 'fake' modifies and those it leaves unchanged are moved in function of this attribution of intentionality. In a famous 1994 paper, Dretske argues that for something to be specifically mental it does not merely need to exhibit original intentionality. It also has to be capable of misrepresentation, i.e. be a structure having a content independent of its causes. I argue that this intuition is implicitly contained in the natural language use of "fake".

Keywords. Intentionality, Misrepresentation, Fake, Privative Adjectives, Formal Semantics.

Copyright. 2018 Janek Guerrini. Published in Italy. Some rights reserved.
Author. Janek Guerrini, janek.guerrini@ens.fr.
Received. 25 February 2018. **Accepted**. 17 December 2018.

1 The Problem

Nouns and adjectives are commonly described in the semantic type-theoretical approach as functions of type <e,t> that take an element or an individual e, and return a truth-value t. For example, "lawyer" is a function that says "give me an individual, and I'll tell you whether or not it's a lawyer." These functions are the characteristic functions of sets, so that properties in the traditional sense can be seen interchangeably as functions and as sets of individuals. What happens when I combine two nouns, or two adjectives, or a noun and an adjective?

Consider sentence (1)

(1) x is a French lawyer

which says $x \in \{French\} \cap \{lawyer\}$. In fact, "French" is an intersective adjective. in that (A) holds for any N.

(A) $||French\ N|| = ||French|| \cap ||N||$

This approach seems a very elegant and obvious account of the compositionality of nouns and adjectives. Unfortunately, it works only for a portion of adjectives. Some cases for which it does not work are adjectives like "fake" or nouns like "toy". A fake violinist is not someone who is both fake *and* a violinist.

The problem can be stated as follows. A formal account of "fake" needs to be able to tell us why and how a "fake gun" is neither a gun nor merely a non-gun. Notice, in fact, that (2) is well-formed.

(2) That gun is a fake gun.

2 A first sketch of the account

We introduce an interpretation of nouns and adjectives as structured sets of properties.

$||G|| = \langle R,P \rangle$

Where:

- G is a noun or adjective.
- R is the set of all *relevant* properties of G that all instances of G must have.
- P is the set of all *prototypical* properties of G that all prototypical instances of G have. We assume $R \subseteq P$.

It is central to point out that R and P are not structured sets of properties themselves, but plain sets of properties. In this approach, properties are still the standard mathematical object: sets of individuals that have a given property.

Only the meanings of natural language words like adjectives and nouns are not merely properties, but structured sets of properties.

A basic example for the noun "gun":

1. Gun

   R= {shoots, kills}

   P= {has the physical form, has a barrel, shoots, kills}

Now we can give an analysis of 'fake':

f(⟨R,P⟩) := ⟨{ seems to *r* but cannot r:r∈R},{$P - R$}⟩

That is, f is a function that takes as an input a *structured set of properties*, both the relevant and the prototypical properties, and returns another structured set of properties, where

- every prototypical property is left unchanged
- every relevant property r is modified from "r" to "seems to r and cannot r".

A basic example:

1. fake gun = f(gun)

   R = seems to {shoot, kill}and cannot {shoot, kill}

   P = { has the physical form, has a barrel}

3 Discussion

3.1 3.1 Instability across contexts

I am not arguing that anytime we pronounce "fake gun" we intend the exact interpretation given above. Rather, if the context makes shooting and killing relevant to being a gun, 'fake' composes by taking only those properties that are relevant as input. But these R-properties change over different contexts, as showed by the example below.

Imagine a world in which guns, instead of keys, are used to open doors. Every gun has unique bullets, you shoot on your door, the door reads the bullet, and if the door recognizes it as its specific bullet it opens. These guns can still shoot and kill. Now I tell you: that's a fake gun.

All I am saying with this utterance is that it opens no door, and not that it is not able to kill. I am using a different R set than I would be using to say, in our actual world, something like "The gun he used to rob the bank was fake". In the gun-key world, the property of being able to kill might just be a prototypical property and not a relevant one. *How* the properties which are relevant to

a predication change over different contexts is not the subject of this paper. By contrast, I have shown that those properties change, and this is sufficient to propose an account of how 'fake' composes with heads by assuming it only applies to those relevant properties.

The fact not all adjectives have an intersective or subsective meaning, like "French" in a phrase such a French lawyer, was observed by Reichenbach as early as the '40s. In the sentence «John is a slow rider», "what is said is not that John is slow in general but only that John is slow in his driving: thus the word 'slow' [...] operates as a modifier of 'drive'".

Attempts to describe fake not only as a non-privative predication, but also as an operation on some internal structure of NPs started as soon as the '80s. Lakoff and Johnson's account of privative NPs is based on the idea that privative Adjs operate ranging over the complex internal semantic structure of terms, especially in the case of artifacts (Lakoff and Johnson 1980). In 'fake gun', for instance, they put forward an analysis which includes three representational dimensions: perceptual, functional, and genealogical.

Franks' account shows that despite their apparent vagueness in classification ("what is a gun?"), NPs work in a quasi-classical way: their internal attribute-value structure determines uncertain predications, but the single attributes ('features') work in a perfectly binary fashion (Franks 1995). Franks' account is the first that posits a complex internal structure that distinguishes between central and diagnostic features. 'Fake' negates only the central features, keeping the diagnostic ones. This makes him predict that the central features of a gun do not obtain of a gun, but the gun does not seem to possess these, making that account of privatives significantly different from this account of 'fake'.

Partee (1987), by contrast, did not posit any rich internal structure, providing a non-intersective extensional account in which adjectives like 'fake' coerce their argument to a broadened extension. Thus in (2) «the first occurrence of gun, modified by fake, is coerced, whereas the second, unmodified, occurrence is not. Normally, in the absence of a modifier like fake or real, all guns are understood to be real guns, as is evident when one asks how many guns the law permits each person to own, for instance. Without the coerced expansion of the denotation of the noun, not only would fake be privative, but the adjective real would always be redundant. »

More recently, Del Pinal (2015) proposed another kind of internal complex structure on which 'fake' acts. NPs are constructed as having one principal extension-determining, competence-linked sub-structure and another, sub-structure containing all the core facts about that noun, secondary and normally not involved in compositionality. 'Fake' works differently than other modifiers because it takes as an input the Cstructure. Hence a fake gun appears to fulfill all the core-facts we know about a gun. Such an extensional semantics can be

employed in a HeimKratzer-like, traditional formal semantics.

This need of positing a rich structure, but still directly extension-linked (without intensional mediation), comes as a result of considering Putnam's influential argument against definitional theories (1970). The argument goes as follows: for any property that supposedly defines an artifact, we can always find a counterfactual situation in which an object falls under the extension linked to that artifact in spite of not having that property. Therefore, words are linked directly to the extension they denote, and it is only through word-use that we understand exactly what individuals fall under an extension.

This helps us address the issue of intensionality. Clearly this is not the topic of the paper: the proposed account presupposes an intensional structure, but it also works by partitioning an extensional space, hence embracing a view sympathetic with Putnam (and seeing the properties we posited above as sets of individuals). However, I want to state my sympathy for intensional internal structures. I want to limit myself to observe that if reference and compositionality occur on a part of the features of an internal structure (as is the case in R vs P), and if the bundle of relevant features is not stable across contexts, then Putnam's argument is less effective. Of course I can think of a counterfactual situation in which an individual not having the property of 'being a hunting big cat with a mane" falls under the extension of lions. But this is because in that situation the properties relevant to this categorization are different.

3.2 Improving the analysis

This is why I did not call R.properties essential. An object's essential property, if it exists at all, is possible world-invariant and would have to hold, in virtue of this, across conversational contexts. For instance, if being a mammal is an essential property of humans, then it is necessarily an essential property of humans. Fake objects, by contrast, enjoy bigger flexibility in that what the gun is made for changes depending on the possible worlds that are being referred to. It looks like, at least in the case of tools, relevant properties amount to the purpose the object was made for. And there are good reasons to make an even stronger claim, namely that some intentionality in the meaning of fake or in the object to which we apply the function f plays a big role in the interpretation of the whole. The following thought experiment should clarify this point:

> Bob and Carl have built a super-powerful telescope: not only can it show you the furthest galaxies, but you can also look through the objects you are pointing it to. By playing around with the telescope and looking at an ice-cold, lifeless and extremely distant galaxy, Bob happens to point at an atom-conglomerate which randomly appears exactly like a gun. Bob calls Carl and they observe it together. How-

> ever, by looking inside the gun with the distance-rays-function of their powerful telescope, Bob and Carl, who are also weapon-freaks and know a lot about how weapons work, notice that, given how the internal mechanism works, the gun could never shoot and kill. Everything else resembles exactly a gun.

In this situation, it wouldn't be appropriate to say: "look, a fake gun!", for the only way in which this can become felicitous is by cooking up a context that posits a builder and a purpose assigned to the gun.

If this is the case, then "fake" implies at least some kind of derived intentionality. By intentionality here I intend the standard Brentano definition (Brentano 1874). Brentano referred to intentionality as the power of minds to be about, to represent, or to stand for, things, properties, and states of affairs. In philosophy of mind debates this is nowadays referred to as "original" or "primitive" intentionality. "Derived" intentionality is borrowed from original intentionality: a gun borrows its intentionality from its conceiver. A gun is about shooting and killing, a hammer is about beating nails.

One could with good reason contest that it is perfectly fine and reasonable to talk about a fake lawyer and that this lawyer was not created with the purpose of seeming a lawyer. But this does not change the fact that when stating a sentence of the form

(3) x is a fake G

you are implying G has some kind of intentionality. This intentionality may be derived, as for the gun, or original, as for the lawyer.

But the example with Bob and Carl leaves us with a question about the account of "fake" that we put forward above. The conglomerate of atoms seems to shoot, cannot shoot and has all of the prototypical properties of a gun, including the physical form , and nevertheless cannot be said to be a fake gun.. Yet the account I proposed above predicts it to be a fake gun. What is missing?

There are two ways to restrict the account as a means of making the right predictions:

> I. Not only does "fake" change relevant properties to "seems to R and cannot R", but it crucially **adds** that it was built by its builder with the purpose of seeming R.
>
> II. "fake" as we defined it is fine. We just have to specify that it applies only to things that already have a purpose, a derived intentionality. That is why it is semantically awkward to say something like "look, a fake atom!", modulo the possibility of pragmatic adjustments that posit a builder of that atom. It is really not clear what this should mean and it does not at all look like something a competent speaker

> would say. The sentence above is therefore unclear unless we try to think out a story in whichthat atom was built by someone who deliberately put into it the purpose of resembling through and through an atom. In this case we would no longer be talking about semantics as it would be a pragmatic adjustment.

Basically, our decision must be whether to add some operations to what the function "fake" already does (I.) or to restrict its domain (II.).

Before making the decision, it may come in handy to unpack what exactly "seem" means when we are saying that a fake G seems to R and cannot R. A reasonable account for sentences like

(4) j seems to be playing

might be that when you utter such a statement you are saying that you think that an observer will think that j is playing. Now we can unfold a more complete interpretation of "fake":

f(⟨R,P⟩) := ⟨{ is capable of fooling j into thinking that r and cannot r:r∈R},{$P - R$}⟩

Where j is a particular observer, a 'judge' or, in its absence, an average observer. This intuition is drawn from the Moltmann 2010 account of taste predicates. Consider the two following examples:

(a) Ron's eyes are shaped in a particular manner, which cannot see the difference between a gun and a drill. Dan usually threatens him with a drill. When he tells the story to his brother, he says "And then I was pointing my fake gun towards him, and…"

(b) "Monopoly bills are fake".

In case (a), j is Ron. In case (b), j is clearly the average observer modulo the assumption of a lack of knowledge about the nature of monopoly.

To translate this into an example, postulating that in a given context, i.e. in a given possible-worlds set

3 lawyer

R= {has a law degree, is member of the bar}

P= {wears a robe, has good rhetoric}

4 fake lawyer=f(lawyer)

R= is capable of fooling j into thinking that{has a law degree, is member of the bar} and it is not the case that {has a law degree, is member of the bar}

P= {wears a robe, has good rhetoric}

The judge parameter is, alongside with the intensional structure, what differentiates this analysis from Del Pinal's, which states that a fake gun was made to have the perceptual features that make it look like a gun, but doesn't make any predictions regarding to whom the fake gun should look like a gun.

We still have to decide whether "fake" not only modifies but also adds some operation to the function it applies to (I.), or it has just a domain restricted to what the speaker attributes some derived or original intentionality (II.). It is very unclear which of the two options delivers a better natural language meaning. The different predictions made by these two hypotheses concern whether the mental-agent-requiring content is at-issue or presuppositional. At-issue content is plainly relevant to truth conditions. So if the at-issue content of a sentence is false, then its negation is true. Presuppositional content, by contrast, works differently. Take for instance the famous Russellian example "The king of France is bald". Because the presupposition of the sentence is not fulfilled, it is neither true nor false, and the same hold for its negation. A characteristic of presuppositions is that, unlike at-issue content, they project out of embedded clauses. Take following sentences:

(i) If that thing over there is a fake atom, I'll be really surprised.

(ii) That thing over there was built by someone, that's for sure, but if it's a fake atom I'll be really surprised.

Under hypothesis I sentence (i) presupposes that "that thing over there was built by someone," while (ii) triggers no presuppositional content. In fact, the content that is supposedly presupposed in (i) is asserted in (ii), and thereby made at-issue. Under hypothesis II neither sentence presupposes anything. Another example that might help clarify is following:

Under hypothesis I sentence (i) presupposes that "that thing over there was built by someone," while (ii) triggers no presuppositional content. In fact, the content that is supposedly presupposed in (i) is asserted in (ii), and thereby made at-issue. Under hypothesis II neither sentence presupposes anything. Another example that might help clarify is following:

A: Oh my gosh, there's a fake atom over there!

B: What you just said is false, since no one built that thing.

B: What you just said is false, since no one built that thing.

Again, if B' seems more of a plausible reaction than B, then it makes a (all but uncontroversial) case for II. So all really depends on which intuitions we have: is there (more of) a presupposition failure in (i) there than in (ii)? Is B' more plausible than B? An informal survey among my colleagues found contrastive intuitions. Experimental work might be needed on this, as judgements are very

subtle here. Del Pinal, on the other hand, seems to be more empathetic with I, as he makes the way in which a thing comes into being at-issue. If we were to make the same choice, we would have following meaning assigned to 'fake':

> f(⟨R,P⟩) := ⟨{ was built with /has the precise purpose of fooling j into thinking that r and cannot r:r∈R},{$P - R$}⟩

3.3 Intentionality and misrepresentation

In his 1994 paper, Dretske argues that intentionality per se is not mental. For take the example of a compass:

> (a) the compass indicates the North-pole
>
> (b) the North-pole is coextensional to the habitat of polar bears

The conclusion of

> the compass indicates the habitat of polar-bears

is not justified. We have generated an intensional and referentially opaque context. This is enough in order for something to have original intentionality. Dretske defines intentionality in the following terms:

> if ascribing a property to x generates an intensional context, then x exhibits original intentionality.

We already defined derivative intentionality as pertinent to those objects (like a gun) whose representing such-and-such can be explained in terms of the intentionality of something else. Then how is a compass different from a gun?

> The intentionality of the device is not, like the intentionality of words and maps, borrowed or derived from the intentionality (purposes, attitudes, knowledge) of its users. The power of this instrument to indicate north to or for us may depend on our taking it to be a reliable indicator (and, thus, on what we believe or know about it). but its being a reliable indicator does not itself depend on us.

As the compass de facto exhibits original intentionality, original intentionality cannot be the distinguishing mark of thought. So mere intentionality involved in the semantics of fake wouldn't be enough to prove that we represent any mental agent behind 'fake'. But what artefacts, however sophisticated, cannot do, argues Dretske, is misrepresenting something without our help:

> Although clocks, compasses, thermometers, and fire alarms-all readily available at the comer hardware store-can misrepresent the conditions they are designed to deliver information about, they need our

> help to do it. Their representational successes and failures are underwritten by and, therefore, depend on our purposes and attitudes, the purposes and attitudes of their designers and users. As representational devices, as devices exhibiting a causally detached meaning, such instruments are not therefore eligible ingredients in a recipe for making thought. (Dretske 1994)

Now: "fake" always hides mental agents with misrepresentational states behind its meaning, for note that:

- we showed that we assume of a hypothetic average observer that he would be fooled by the object / person. To be fooled, the observer must form himself a misrepresentation. In other words, there must be some semantic content independent of its causes.
- we attribute to the builder from which the object borrows intentionality or to the person of which we predicate fakeness the active misrepresentation of the relevant properties of the thing of which they are a fake version.

Thus

(a) Misrepresentation needs a specifically mental

(b) The adjective "fake" needs misrepresentation

(c) we always need a mental agent or an intentionality derived from a specifically mental agent when we use "fake".

Note indeed that mere derived intentionality is not enough for fake objects. For instance, if I want to build a paperweight but by pure accident end up making something that resembles a gun, is that a "fake gun"? No, and yet it definitely has derived intentionality. I built it with a purpose (being a paperweight) that it can accomplish. What is needed behind the fake object is an *active misrepresenter who has the goal of that exact misrepresentation.*

In conclusion, I turn to explaining why (2) is a perfectly fine sentence.

(2) That gun is fake.

That object can be described as a "gun" only from the point of view of the person who is fooled, and as "fake" only from the point of view of a fully informed person. But we know that there are observers that represent that thing as a gun. It seems that there is a perspective shift between "gun" and "fake" taking place in (2), from the fooled person to the speaker (or from the average person to the well-informed speaker in case of a the standard filling the void argument). The speaker says "gun" by putting himself in the average observer's frame, and then switches to his better-informed frame in order to predicate "fakeness" of the object.

By shifting frame of reference, we are shifting between different boundaries between R and P . The domain of conversation modifies these boundaries only insofar as it modifies what type of intentionality is behind of the objects of the conversation. We delimit R and P in function of the mental agents we have to posit in order to be able to predicate 'fakeness': the average observer and the lawyer or the builder of the gun. We understand each other and agree in the vast majority of occurrences in what respects an object is "fake". This shows how cognitively convenient it is for us to posit mental agents. How good we, as humans, are at understanding others' representational and misrepresentational states.

For this reason, accepting the analysis proposed in this paper and the linked view does not necessarily mean endorsing an intentional view of the mind rather than a computational one. Dennett himself Dennett 1987 concedes that intentionality is for humans a convenient way to think about others' mental states. Ascribing to our chess opponent psychological states with intentionality makes it easier for us to imagine what she is thinking, planning, avoiding, fearing etc. All one is committed to when accepting this analysis is that we do indeed ascribe a mental intentionality mainly characterized by capability of misrepresentation to others, not that the intrinsic nature of the mind itself is intentional.

Acknowledgements

I am particularly thankful to Salvador Mascarenhas for his valuable and constructive suggestions, as well as for all the fun chats, during the planning and development of this work. I am also grateful for useful comments and lively conversations on this topic to Achille Varzi, Jeremy Kuhn, and Giacomo Mazzucchelli, as well to the audience of the 2018 University of Alberta Philosophy Graduate Conference.

References

Brentano, Franz (1874). *Psychology From an Empirical Standpoint*. Routledge 1995.

Coquand, Thierry (2015). *Type Theory*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/archives/sum2015/entries/type-theory/.

Del Pinal, Guillermo (2015). "Dual Content Semantics, privative adjectives, and dynamic compositionality". In: *Semantics and Pragmatics* 8.7, pp. 1–53.

Dennett, D. and J. Haugeland (1987). "Intentionality". In: *The Oxford Companion to the Mind*. Ed. by R.L. Gregory. Oxford: Oxford University Press, pp. 139–143.

Dennett, Daniel (1987). *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Dretske, Fred (1994). "If You Can't Make One, You Don't Know How It Works". In: *Midwest Studies in Philosophy* 19.1, pp. 468–482.

— (1995). *Naturalizing the Mind*. Cambridge, Mass.: MIT Press.

Fodor, Jerry (1975). *The Language of Thought*. New York: Crowell.

Franks, Bradley (1995). "Sense generation: A quasi-classical approach to concepts and concept combination". In: *Cognitive Science* 19.4, pp. 441–505. URL: http://dx.doi.org/10.1207/s15516709cog1904_2.

Jacob, Pierre (2014). *Intentionality*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/archives/win2014/entries/intentionality/.

Klein, Ewan (1980). "A semantics for positive and comparative adjectives". In: *Linguistics and Philosophy* 4.1, pp. 1–45.

Lakoff, George and Mark Johnson (1980). *Metaphors we live by*. Chicago, IL: The University of Chicago Press.

Parsons, Terence (1970). "Some problems concerning the logic of grammatical modifiers". In: *Synthese* 21.1, pp. 320–334.

Partee, Barbara (1987). "Lexical Semantics and Compositionality". In: *An Invitation to Cognitive Science (Second Edition). Volume 1: Language*. Ed. by Lila Gleitman. Cambridge, Mass.: MIT Press.

— (2010). "Privative: Subsectives plus coercion?" In: *Presupposition and discourse: Essays offered to Hans Kamp*. Ed. by R. Baule, U. Reyle, and T.E. Zimmerman. Amsterdam: Elsevier.

Putnam, Hillary (1970). "Is semantics posible?" In: *Metaphilosophy* 1.3, pp. 187–201. URL: http://dx.doi.org/10.1111/j.1467-9973.1970.tb00602.x.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

Temporal Subitizing and Temporal Counting: a Proposal between Vision and Action

Andrea Roselli

Abstract. How is our temporal experience possible? When we hear a song, we are aware that every note is before and after another note (and that's how we remember it), but we also 'experience-as-present' more than one note at a time. To answer this question, I suggest an analogy with the difference drawn, in the spatial case, between the two different mechanisms of counting and 'subitizing' (the immediate visual capture of a certain number of items as a single object). My proposal is to identify two different mechanisms even in the temporal case: a temporal counting, a coconscious experiential 'single look' of a temporal interval; and a temporal subitizing, an atomic storing operation which organizes every event in a mathematical, point-like sequence. These two mechanisms are taken to be operative always and together; we never cease to store the events encountered in a temporal line, but we also experience a subgroup of them as present.

Keywords. Models of Temporal Understanding, Phenomenal Temporality, Specious Present.

Copyright. © 2018 Andrea Roselli. Published in Italy. Some rights reserved.
Author. Andrea Roselli, roselli.uniroma3@gmail.com.
Received. 30 March 2018. **Accepted**. 26 October 2018.

1 The models of our temporal phenomenology and a shared problem

The three main accounts[1] of our experience of time and presentness are the Cinematic Model, the Retentional Model and the Extentional Model. Cinematists reject the idea of a Specious Present. They maintain that our temporal phenomenology is a succession of momentary states of consciousness. In our phenomenology we always know what comes first and what second; then, the best model to describe our temporal awareness is one in which there are momentary states of consciousness (physiologically momentary: about 30ms, time under which we can't distinguish the order of two stimuli[2]). But how can they account for perceptions of motions? While, in fact, there is a distinct frame to point at when we want to know where does the experience "I see the green apple on the desk" come from (there is a frame containing the green apple), we can't do the same with the also very familiar experience "I see the green apple falling from the desk" (in every frame the apple is in one position: it is never falling). If our perceptual consciousness consists of a succession of momentary experiences, we never really perceive the apple falling in the same way we perceive it 'being green'. Where, then, does this dynamical feature of our experience come from? It seems that a story needs to be told about how, from this succession of experiences, we have an experience of succession. One thing is to have in mind the different positions that an object occupied in time and have the cognitive understanding that it moved, and another thing is to directly perceive it moving[3].

It is to save this last intuition that the two other models of temporal experience were born – Retentionalism and Extentionalism. These models are realists about phenomenal temporality: change, succession and persistence can be directly perceived or apprehended[4]. Both Extensional and Retentional theorists agree that a temporal spread of contents can be apprehended as a unity. Not only, then, simultaneous contents can be experienced together, but even contents that are successive; contents which are apprehended as unified in this way belong to a single specious present. How is it possible, however, to perceive an extended present? When we hear three close auditory tones, we seem to hear the musical phrase as present, and yet we also hear the notes as successive, and

[1] See Rashbrook (2013), Prosser (2013, 2016), and Hoerl (2014b,a, 2015) for an extensive discussion on the matter.

[2] Stimuli of around 1ms need to be separated from one another by in interval of around 30 msec if they are to be perceivedasa succession – a result which holds across sensory modalities. Stimuli which are separated by shorter intervals are not perceived as distinct.

[3] Obviously enough, many refined arguments could be put forward by the Cinematist to defend her position: all I'm trying to do here, however, is to present the main models of our temporal phenomenology to show how the Specious Present is present in them.

[4] There is the possibility to build a 'Cinematist Realist' model, but virtually every philosopher of time who defends Cinematism is an Anti-realist about phenomenal temporality.

therefore as extending over an interval. How could a succession of elements – elements which are experienced as *before* and *after* – also be experienced as present *in toto*? Retentionalist and Extentionalist, while accepting both the idea of an extended Specious Present, give different accounts of this apparent paradox.

Retentionalists agree that our experiences occur within episodes of consciousness which lack an objective, clock-time extension: but these episodes, they maintain, are composed by an immediate experience *and* a representation (or retention) of the recent past; the result is that the contents of these experiences represent temporally extended intervals. The stream of consciousness, then, is composed of succession of momentary states – just as the Cinematists claim: the difference, however, is that the experience of these momentary states is one of duration. The confinement to a momentary present is seen by Retentionalists as a condition for contents to be experienced together: phenomenal unity needs the simultaneous presentation of contents to a single momentary awareness. Retentionalists, however, are typically accused to have invented "nothing but a new word" (Dainton 2000, p. 155): what is a retention, and in what differs from a memory? Until we explain how does it work, it is just an *ad hoc* solution. How is that possible that a portion of what the Retentionalist herself calls recent 'past' is added to our present, point-like experience, creating a new whole? Shouldn't there be some sort of difference between the present and the retained past? Isn't it, then, just another version of the Cinematic model, in which we simply call the awareness of the recent past with a different name? If we choose the other horn of the dilemma, however – clearly differentiating memories and retentions – we risk to multiply the experiences: shouldn't we hear-as-present a sound in all the different point-like Specious Presents that contain it? This is why Extentionalists claim that the Specious Present is not merely experiential, but extends over clock-time; they hold that the atomic unit of our perception is an extended period of time: we have an experience of succession because we directly experience the succession. The Retentionalist doctrine that diachronic phenomenal unity can only exist in strictly momentary states of consciousness is rejected, in favour of a more 'natural' model of temporal awareness: change and persistence are incorporated in our experience in a quite straightforward way, since our stream of consciousness is composed of a succession of an extended chunk of experience; the main Extentionalist claim, then, is that experience itself is extendend, and not just its content (vehicle and content share their temporal properties). The Extentionalists' Specious Present is itself temporally extended, and its parts succeed one another in time in just the way they seem to: our experiences extend over a period of real time, in a way which (almost infallibly) matches the phenomenal period it presents.

Realists about phenomenal temporality, such as Extentionalists and Reten-

tionalists, explain the immediacy associated with experiences of change, persistence, succession, in a quite direct way; their problem however, one that Cinematists don't seem to face, is to explain how is it possible that the succession experienced in the extended present doesn't collapse in a temporal *unicum*: how is it possible, for contents that are all experienced as present, to be presented to our conscious life as in succession rather than simultaneously? How come that not only *objectively*, but even *phenomenologically*, there is a before and an after in a Specious Present? Shouldn't the extended present be experienced as a totul-simul (we directly experience the succession of notes without confusing their order)?

How, moreover, should we divide one extended present from another? While it was obvious in the Cinematist case (every single perception, such as a note, is one present experience), it is not so obvious in the Retentionalist or Extentionalist case: how long are these extended present experiences, and how they succeed one another without giving the feeling of a continuous hiccup (which is a stream, of course, but a very unappealing one)? There is a double dilemma, then, for the realist about phenomenal temporality: how could it be that within these wholes there is a succession, a before and an after? And how could it be that each experienced whole seamlessly gives way to the next?

In this paper I sketch a possible way out from this double dilemma; what the three different models of our temporal phenomenology have in common is that they all try to reduce one side of our temporal phenomenology to the other; Cinematists give priority to the phenomenology of succession, and try to minimize the experience of a Specious Present; Extentionalists and Retentionalists give priority to the Specious Present, but they have problems when it comes to explain why our extended experiences of a temporal 'present' don't merge all the perceptions in one simultaneous datum. The novelty of my proposal consists in the acceptance of the paradox – cognitive neuroscience may indicate us the way to a better model, and our possibility to act and react will be crucial in this phenomenological model of our temporal perception.

2 Synchronic and diachronic unity

There are two macro-areas of concern regarding the phenomenology of our temporal experience: questions about synchronic unity at a time, and questions about diachronic unity over time. Not only, in fact, do we experience many successive movements of an object in front of us as fluidly reunited in a temporal extended now, our present moment; we also experience an endless stream of these 'nows', without being capable of pinpointing, locating or even remotely feeling any kind of definite boundary between them. There have been attempts to argue in favor of a unified account, providing one answer to both questions: how-

ever, it seems that there are some structural differences that make it impossible. Oliver Rashbrook (2013) argues very convincingly that similar solutions hide two very different notions of 'togetherness'. While in fact, on the one hand, 'being experienced together' is a transitive relation in our experience of synchronic unity at a time, it is a non-transitive relation in our experience of diachronic unity over time (the continuity of consciousness tells a very different story from that of a single, prolonged experience during our waking hours). But the relation can't be both, at least not in a *unified* account of our temporal phenomenology.

There seems to be a genuine problem here. Consider the auditory experience of a fast piano song; our phenomenological experience of 'the present' is a single look, so to speak, to a brief succession of notes. We simply *can't* experience-as-present only one note at a time (remember: it is a *fast* song). Still, after one minute not only are we *aware* that we are not experiencing the beginning of the song: we also don't experience-as-present the first notes of the song. There must be in play here two very different ways to have a temporal experience: on the one hand, there is a brief but extended present, that even if distinguishes the succession of (say) three notes, comprises them all in a single temporal present experience – as the single vision of three dots on a screen: you can tell that there is one on the left, one in the center, and one on the right: still, you don't need to look singularly in turn at every one of them to tell. In this case, 'togetherness' is a transitive relation. On the other hand, there is a completely different way to temporally experience the song: instead of a single look, it resembles much more the operation of storing the notes in succession; in this case, 'togetherness' has a whole different meaning.

It seems that there are two different phenomenological processes going on: if we had absolute pitch and a prodigious, Mozart-like memory, at the end of the song we would remember perfectly the stream of the notes, being capable of saying which were played before, and which were played after; if we chose a random note, we would be able to tell which notes were in its past, and which notes were in its future; our total temporal experience of the song, then, is that of a continuous stream of temporally ordered single notes; a mathematical succession of points, so to speak. A totally different process, however, is responsible for our direct temporal experience while the song is being played. Think of what you would answer if someone asked you, during the song, "what are you hearing now?": instead of an ordered succession of single notes being present and successively, in turn, being stored in the past, your present experience would much more likely be that of a brief succession of multiple notes, which – even if they are in succession – are all felt as part of the same present; there are more-than-one notes in our experiential now.

Let me make another example. Suppose you live in a poor and dangerous neighborhood; one night, you got frightened by the sudden sound of two close

gun shots (say, 100ms from one another). Try to imagine your temporal experience: even if you heard two separate shots, you do not experience the first 'as past' when you hear the second. Nonetheless, when the police officers interrogate you, you have no problem telling that one shot was *before* the other; you are absolutely aware that, technically, when one was present the other was in its past. Indeed, at the end of that ugly night, you remember a stream, a sequence of temporal ordered gun shots; the single experience that you had when you heard the two close shots – when you could actually act – is lost, replaced by an ordinate succession available for your memory.

Maybe the simplest option is even the right one: if we experience two such different things, it could be because there are two different phenomenological processes going on, and our temporal experience is twofold: to make my proposal clearer, I am going to propose an analogy with a spatial debate, which has significantly been tackled with the recourse of such a dualism between two distinct ways of operating of our intellect: counting and 'subitizing'. It is important to stress at this point, however, that the analogy serves merely to indicate the direction I am taking: I do not intend to claim that there is a straightforward relation between my temporal model and the spatial models that make use of the notion of 'subitizing'; it is obviously possible to explore the conceptual link between them, but it is beyond the purposes of the present paper.

'Subitizing' is a latinism coined[5] in the mid-fifties to describe the immediate visual capture of a certain number of items, to be distinguished from the usual action of counting. The idea behind it was to see if there were a cognitive description of our everyday-life different performances in front of streams of not-grouped and grouped numbers (4939724 and 4,939,724; car plates; bank accounts; etc.). Experimental results[6] showed a significant difference between judgments made for displays composed of one to four items, and for displays of more items; of course, response times always rise with the increase of the number of the items showed, but it is often claimed that there is a dramatic difference between the two groups[7] (see Figure 1).

[5] See Kaufman et al. (1949).

[6] See for example Trick and Pylyshyn (1994), or Camos and Tillmann (2008).

[7] Inside the range 1–4 objects, there is an increase of the time necessary for an accurate response of about 50ms every added element; in the range +4 objects, however, the increase in response time becomes of about 300ms.

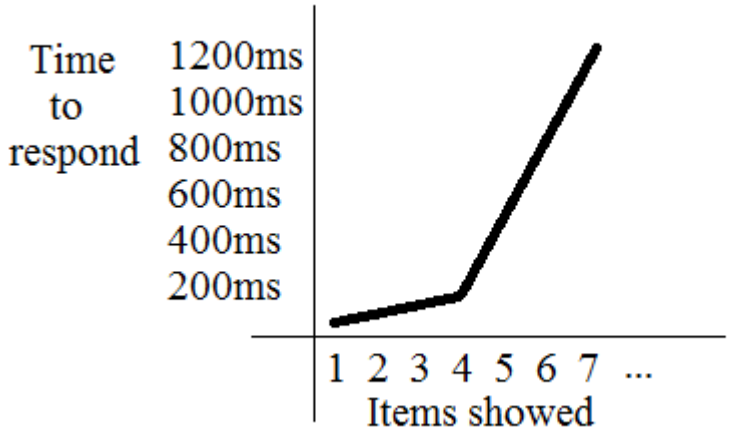


Figure 1[8]

In current scientific literature we find a lot of different models to explain these results. Sometimes (rarely) the limit between subitizing and counting is set after the third object, instead of the fourth; given that there is never an indisputable discontinuity in the curves of response, moreover, there are even those who deny that there are two different mechanisms to determine visual numerosity. Gallistel and Gelman (1991), for example, famously claimed that even small sets of items are quantified by serial counting, albeit with faster speed than for larger sets: subitizing, then, would just be a fancy word to say 'fast counting'. Others see in our ability to subitize small groups of numbers a similarity to object recognition: Mandler and Shebo (1982), for example, argued that subjects recognize the characteristic geometric configuration of sets of objects (for example: 1, point; 2, line; 3, triangle). This pattern recognition would fail for sets of more items, at which point the subject would then start to (slowly) count. Trick and Pylyshyn (1994) attributed subitizing to the parallel assignation of pointers called 'fingers of instantiation' to each object in a visual display; these 'fingers', it is assumed, are available in a limited number (four), as it is suggested by multiple object tracking experiments. Subitizing, then, would be based primarily on preattentive processing, and be dissociated from serial counting. To similar conclusion came Dehaene and Cohen (1994), Simon and Vaishnavi (1996), Robertson et al. (1997), Piazza et al. (2002), Maloney et al. (2010).

A disquisition on the single models' merit exceeds the purposes of this thesis; it is sufficient to say that, even if many possible explanations have been put for-

[8]I created this figure on the basis of the data presented in Akin and Chase (1978), Klahr and Wallace (1976) and Mandler and Shebo (1982); in their result, it is shown the non-trivial fact that the subjects of the experiments needed, in order to press a button and tell the exact number of elements on a screen, 25 to 100 more milliseconds every added element in the range 1-4, but after the fourth element the difference for every added element suddenly raised to 250-350 milliseconds per added element. The 'elbow' shown in the figure is also confirmed by Trick and Pylyshyn (1994), in which it is considered the percentage of errors made by the subjects.

ward to explain such a dramatic difference between our abilities to enumerate objects, there seems to be a convincing amount of proofs pointing in the direction of the existence of two different mechanisms at the basis of our different performances in front of a visual display of objects. What I'm proposing, without suggesting a straightforward relation, is a temporal analogy. It seems, in fact, that even in the temporal case there are at work two different processes: while a *temporal subitizing* has an 'action guidance task', which is responsible for our directly experienced present – a single 'temporal look' at an extended period of time that comprises a succession of more notes (for example) in an immediate co-conscious present temporal experience – a *temporal counting* has the cognitive task to store in succession the events perceived. Before turning to the main argument of this paper, however, it is necessary to draw a distinction between long-term, short-term and working memory.

3 Long-term, Short-term and Working memory

Three types of memory are distinguished in scientific literature: long-term memory, short-term memory, and working memory. It is crucial to introduce this distinction at this point of the paper because, as I will argue in the next session, it is possible that temporal subitizing has to do with short-term or working memory mechanisms, while temporal counting has to do with long-term memory processes. Let me proceed in order and clarify the distinctions, first.

Long-term memory is a vast store of knowledge and a record of prior events, and it exists according to all theoretical views. Short-term memory reflect faculties of the human mind that can hold a limited amount of information in a very accessible state temporarily. One might relate short-term memory to a pattern of neural firing that represents a particular idea and one might consider the idea to be in short-term memory only when the firing pattern, or cell assembly, is active. The individual might or might not be aware of the idea during that period of activation. As Nelson Cowan (2001, 2008) showed, short-term memory differs from long-term memory in respect to temporal decay and (crucial for the present argument) chunk capacity limits. Working memory, then, has been conceived and defined in three different, slightly discrepant ways: as short-term memory applied to cognitive tasks, as a multi-component system that holds and manipulates information in short-term memory, and as the use of attention to manage short-term memory. Regardless of the definition, there are some measures of memory in the short term that seem routine and do not correlate well with cognitive aptitudes and other measures (those usually identified with the term "working memory") that seem more attention demanding and do correlate well with these aptitudes. What is clear, however, is that working memory is not completely distinct from short-term memory. It is a term that was originally

used to refer to memory as it is used to plan and carry out behavior. One relies on working memory to retain the partial results while solving an arithmetic problem without paper or to combine the premises in a lengthy rhetorical argument. Measures of working memory have been found to correlate with intellectual aptitudes (and especially fluid intelligence) better than measures of short-term memory and, in fact, possibly better than measures of any other particular psychological process (see for example Conway et al. 2005). This reflects the use of measures that incorporate not only storage but also processing, the notion being that both storage and processing have to be engaged concurrently to assess working memory capacity in a way that is related to cognitive aptitude.

But what are the relations between long-term, short-term, and working memory mechanisms? Short-term memory is derived from a temporarily activated subset of information in long-term memory. This activated subset may decay as a function of time unless it is refreshed, although the evidence for decay is still tentative at best. A subset of the activated information is the focus of attention, which appears to be limited in chunk capacity (how many separate items can be included at once). New associations between activated elements can form the focus of attention. The distinction between short-term memory and working memory is clouded in a bit of confusion but that is largely the result of different investigators using different definitions. Cowan et al. (2006) proposed, on the basis of some developmental and correlational evidence, that multiple functions of attention are relevant to individual differences in aptitudes. The control of attention is relevant, but there is an independent contribution from the number of items that can be held in attention, or its scope. According to this view, what may be necessary for a working memory procedure to correlate well with cognitive aptitudes is that the task must prevent covert verbal rehearsal so that the participant must rely on more attention-demanding processing and/or storage to carry out the task. The idea is that a working memory test will correlate well with cognitive aptitudes to the extent that it requires that attention be used for storage and/or processing. In sum, the question of whether short-term memory and working memory are different may be a matter of semantics. There are clearly differences between simple serial recall tasks that do not correlate very well with aptitude tests in adults, and other tasks requiring memory and processing, or memory without the possibility of rehearsal, that correlate much better with aptitudes. Whether to use the term working memory for the latter set of tasks, or whether to reserve that term for the entire system of short-term memory preservation and manipulation, is a matter of taste. The more important, substantive question may be why some tasks correlate with aptitude much better than others.

The distinction between long-term and short-term memory depends on whether it can be demonstrated that there are properties specific to short-term mem-

Rivista Italiana di Filosofia Analitica Junior 9:2 (2018)

ory; the main candidates include temporal decay and a chunk capacity limit. The question of decay is still pretty much open to debate, whereas there is growing support for a chunk capacity limit. The distinction between short-term memory and working memory is one that depends on the definition that one accepts. Nevertheless, the substantive question is why some tests of memory over the short term serve as some of the best correlates of cognitive aptitudes, whereas others do not. The answer seems to point to the importance of an attentional system used both for processing and for storage. The efficiency of this system and its use in working memory seem to differ substantially across individuals, as well as improving with development in childhood and declining in old age.

4 Temporal subitizing and temporal counting

What we are looking for is an account of our phenomenology capable of explaining why we see temporal extended phenomena, such as motions, as clear and directly as colors. The account should also indicate the extension of this extended present; it should explain why we are simultaneously capable of decomposing our auditory experience of a song in a succession of notes (point-like presents) and still composing more-than-one notes in a single present experience; the different models of our temporal phenomenology sketched above (Cinematism, Retentionalism, Extensionalism) try to give a unified account of these two phenomenological aspects, but it is always one of them reduced – and thus, in a certain sense, sacrificed – to the other. A good way to characterize our phenomenology, then, could be represented by the distinction between two different ways to temporally experience the events, inspired by the debate regarding the mental processes that allow us to transform a given perceptual input into a proper motor output.

When we think back, not only we know that event A preceded event B: we also lose the sensation of a unique temporal experience of them – we only feel them as part of an ordered stream. In our present, however, the situation is different: we can't help but subitize the contents of our perception; we can't look at a ball as being in different positions at different times: we see the motion. We can even force ourselves to consider only a point-like instant, but we can't perceive it as being so: our temporal phenomenology of the present is always extended. As in a single vision of an image there is a left and a right, in our extended temporal experience of the present we recognize an after and a before. When we see an image containing three points we subitize: we are almost immediately conscious of the fact that there are three points; if we wish, we can also focus on every one of them singularly, 'counting' them, but we can't help to simultaneously have a general vision of the figure as containing one object on the left, one in the center, one on the right. If more objects are added, however, we lose the ability to

subitize: we start to focus on little areas of the image, subitizing on those, and moving our focus (that's what we do when, for example, we group numbers as in 345,678,912).

Think of the present experience of hearing two close sounds, one much longer than the other (say, 200ms of the note DO and 100ms of the note LA). Our temporal phenomenology tells a story of one present: we had one experience, we didn't 'have time' to have an experience and then another one (when we experience the note LA, the note DO isn't in our phenomenological past: it doesn't 'feel past'); still, there is a sense in which we have experienced the different duration; it is as if our experience were simultaneously made of parts and still integral and undivided. The proposal, then, is to think of two different ways of experiencing the continuous encounter of a succession of numbers (events): on the one hand we 'temporally count' them, storing them singularly and attributing them a particular, point-like present, as in the series 1 2 3 4 5 6 7 8 9: every number is alone in its present; on the other hand we 'temporally subitize' them, directly experiencing a series of them as already being together, animated, and making the cinematic metaphor disappear, as in the series 123, 456, 789 – where 123 is a single experience of motion from 1 to 3: the total experience does not consist of three stationary image, but of a motion. When the subject subitizes, then, she 'knows' that the first note (DO x 200ms) was played before and longer than the LA, but what she experienced was a co-conscious present experience of 'DOOO-LA'; the first note wasn't in her experiential past when she heard the second one. The model should thus translate not only the two different temporal experiences famously discerned by Broad, but also the fact that we are contemporaneously aware that something is moving *and* feel it moving.

Even in the spatial case, when presented with a great number of objects, we simultaneously subitize and count: we shift our viewpoint around the display and keep track of our count, but we also tend to see subgroups of objects, subitizing them. Our temporal experience is continuously presenting us with events, and even if we are able to *count* them, storing them in order as if they were disposed in a uni-dimensional mathematical line (knowing which note is before and what is after), we also *subitize* subgroups of events, experiencing them in a co-conscious present, seeing them in a single look. When we hear three notes of a song, then, there is an immediate awareness of the auditory elements, we hear them in a single, co-conscious experience; if the notes become ten, on the other hand, we lose the overall sensation of a single experience, and at the tenth note we already feel that the first is 'past': we can't 'see' the ten notes as a single object.

The difference drawn between subitizing and counting in a temporal context could be correlated with the differences described in current literature between long-term memory and short-term and working-memory and, in partic-

ular, with attentive and pre-attentive estimation mechanisms (as suggested, for example, in Burr, Turi, and Anobile 2010). Even when subjects do not have the time or opportunity to count the number of objects in the field of view, they can *estimate* numerosity rapidly (approximate *estimation* of number has been demonstrated in humans, see for example Whalen, Gallistel, and Gelman 1999). The ability to estimate number correlates strongly with mathematics achievement (Halberda, Mazzocco, and Feigenson 2008), suggesting it is strongly linked to other number-based capacities. Estimation of numerosity is rapid and effortless but not errorless. Error increases in direct proportion to the number of items to be estimated, a property known as *Weber's law*. The *Weber fraction*, defined as the just noticeable difference or precision threshold divided by the mean, is usually found to be quite constant over a large range of base numerosities. Thus, subitizing may be nothing special, merely a consequence of the resolution of estimation mechanisms and the quantal separation at low numbers. However, this idea has not received experimental support; as Burr, Turi, and Anobile (2010, p. 20) comment, "subitizing tends to be resistant to attempts to disrupt it". In particular, it seems that subitizing depends strongly on attentional resources, while estimation of larger quantities depends far less on attentional load. Subitizing is often considered to be a pre-attentive process, while enumeration of larger numbers is considered to require attention (although this is more controversial). There has been some debate as to whether subitizing uses the same or different mechanisms than those of higher numerical ranges and whether it requires attentional resources. Recent results[9] seem to show that the mechanisms operating over the subitizing and estimation ranges are not identical, and that pre-attentive estimation mechanisms works at all ranges, but in the subitizing range attentive mechanisms also come into play. The question is thorny, but there is a good experimental base to claim that in the temporal cases discussed above there may be two different mechanisms at work.

An easy objection, at this point, would come from the request of a precise indication of the boundaries of our temporal subitizing. I don't have an answer to that, but it isn't necessarily a flaw of the model here exposed. Experiences such as 'hearing a song' strongly suggest the existence of a present temporal window – we experience-as-present a non-point-like extension of the song, but much shorter than the song itself. When we temporally subitize, we try to keep under one, general look a duration of time (for example, many notes of a song); the operation becomes harder and harder with the passage of time and the accumulation of notes, and the first notes of the song start to slide away. But not only are the boundaries between the two way to experience the events not manifest: they could also depend, for example, on how much we are inclined to focus on the single notes rather then a rhythm; on our ability to anticipate the future; on

[9] See for example Burr, Turi, and Anobile (2010).

how well we know the song, etc. My hypothesis is that there isn't an unambiguous and unique window in which we temporally subitize, then, but I think there is, however, a clear phenomenological distinction between temporal subitizing and temporal counting. The difficulty, as we have seen, could be also attributed to the limits of the working memory. Progress in the field of the distinction between the different types of memories could be relevant to better explain the phenomenological difference in the temporal case that I have sketched here. It is also crucial to underline, finally, that it would be misleading to think that the subitizing is only related to visual experiences[10]; though more immediate to understand and test, it is probably better to refer to a 'sensorial subitizing', instead of a mere visual one.

Conclusions

The question is: how is our temporal experience possible? Many conflicting elements must coexist: our present is extended, but not a *totul simul*; it has boundaries, but they are shifting and not manifest; it is part of a seamless stream, but distinct from the past and the future. My answer to the question is the identification of two different mechanisms: a temporal subitizing, a co-conscious experiential 'single look' of a temporal interval; and a temporal counting, an atomic storing operation which organizes every event in a mathematical, point-like sequence.

Given the great amount of changes and events experienced, the two mechanisms are taken to be operative always and together: we never cease to store the events encountered in a temporal line, but we also experience a subgroup of them as present. Even if we are aware that technically, from a physical-mathema tical perspective, 'the present' is point-like, our phenomenological present gathers recent events in a co-conscious experience. The two mechanisms described are at the basis of our twofold temporal experience: the awareness that every note is before and after another note, and the 'experiencing as present' of more than one note.

[10]Thanks to an anonymous referee for this useful remark.

References

Akin, O. and W. Chase (1978). "Quantification of three-dimensional structures". In: *Journal of Experimental Psychology: Human Perception and Performance* 4.3, pp. 397–410.

Bradley, F. H. (1922). *The Principles of Logic*. Oxford: Oxford University Press.

Burr, D. C., M. Turi, and G. Anobile (2010). "Subitizing but not estimation of numerosity requires attentional resources". In: *Journal of Vision* 10.6, p. 20.

Camos, V. and B. Tillmann (2008). "Discontinuity in the enumeration of sequentially presented auditory and visual stimuli". In: *Cognition* 107, pp. 1135–1143.

Chi, M. T. H. and D. Klahr (1975). "Span and rate of apprehension in children and adults". In: *Journal of Experimental Child Psychology* 19.3, pp. 434–439.

Conway, A.R.A. et al. (2005). "Working memory span tasks: a methodological review and user's guide". In: *Psychonomic Bulletin & Review* 12, pp. 769–786.

Cowan, Nelson (2001). "The magical number 4 in short-term memory: a reconsideration of mental storage capacity". In: *Behavioral Brain Sciences* 24, pp. 87–185.

— (2008). "What are the differences between long-term, short-term, and working memory?" In: *Progress in brain research* 169, pp. 323–338.

Cowan, Nelson et al. (2006). "Scope of attention, control of attention, and intelligence in children and adults". In: *Memory and Cognition* 34, pp. 1754–1768.

Dainton, B. (2000). *Stream of Consciousness. Unity and continuity in conscious experience*. London: Routledge.

Dehaene, S. and L. Cohen (1994). "Dissociable Mechanisms of Subitizing and Counting: Neuropsychological Evidence From Simultanagnosic Patients". In: *Journal of Experimental Psychology Human Perception & Performance* 20.5, pp. 958–975.

Elbert, T. et al. (1991). "The processing of temporal intervals reflected by CNV-like brain potentials". In: *Psychophysiology* 28, pp. 648–655.

Euler, M. (1997). "Sensations of Temporality: Models and Metaphors from Acoustic Perception". In: *Time, Temporality, Now. Experiencing Time and Concepts of Time in an Interdisciplinary Perspective*. Ed. by H. Atmanspacher and E. Ruhnau. Berlin, Heidelberg: Springer-Verlag.

Flanagan, O. (1998). "The Robust Phenomenology of the Stream of Consciousness". In: *The Nature of Consciousness: Philosophical Debates*. Ed. by N. Block, O. Flanagan, and G. Guzeldere. Cambridge, MA: MIT Press.

Franz, V. H. et al. (2000). "Grasping visual illusions: no evidence for a dissociation between perception and action". In: *Psychological Science* 11.1, pp. 20–25.

Gallistel, C. R. and R. Gelman (1991). "Preverbal and verbal counting and computation". In: *Cognition* 44, pp. 43–74.

Goodale, M. A. and A. D. Milner (1992). "Separate visual pathways for perception and action". In: *Trends in Neurosciences* 15.1, pp. 20–25.

Halberda, J., M. M. Mazzocco, and L. Feigenson (2008). "Individual differences in non-verbal number acuity correlate with maths achievement". In: *Nature* 15.1, pp. 665–668.

Hoerl, C. (2014a). "Do we (seem to) perceive passage?" In: *Philosophical Explorations* 17, pp. 188–202.

— (2014b). "Time and the domain of consciousness". In: *Annals of the New York Academy of Sciences* 1326, pp. 90–96.

— (2015). "Seeing motion and apparent motion". In: *European Journal of Philosophy* 23.3, pp. 676–702.

Kaufman, E. L. et al. (1949). "The discrimination of visual number". In: *The American Journal of Psychology* 62.4, pp. 498–525.

Klahr, D. and J. G. Wallace (1976). *Cognitive Development: An Information-Processing View*. Mahwah (US): L. Erlbaum Associates.

Maloney, E. A. et al. (2010). "Mathematics anxiety affects counting but not subitizing during visual enumeration". In: *Cognition* 114, pp. 293–297.

Mandler, G. and B. J. Shebo (1982). "Subitizing: An analysis of its component processes". In: *Journal of Experimental Psychology: General* 111.1, pp. 1–22.

Piazza, M. et al. (2002). "Are Subitizing and Counting Implemented as Separate or Functionally Overlapping Processes?" In: *NeuroImage* 15, pp. 435–446.

Poppel, E. (1978). "Time Perception". In: *Handbook of Sensory Physiology*. Ed. by H. W. Held, M. Leibowitz, and H. L. Teuber. Vol. Perception. Berlin: Springer, pp. 713–729.

— (1997). "The Brain's Way to Create "Nowness"". In: *Time, Temporality, Now. Experiencing Time and Concepts of Time in an Interdisciplinary Perspective*. Ed. by H. Atmanspacher and E. Ruhnau. Berlin, Heidelberg: Springer-Verlag.

Prosser, S. (2013). "Passage and Perception". In: *Noûs* 47, pp. 69–84.

— (2016). *Experiencing Time*. Oxford: Oxford University Press.

Rashbrook, O. (2013). "The Continuity of Consciousness". In: *European Journal of Philosophy* 21, pp. 611–640.

Riggs, K. J. et al. (2006). "Subitizing in tactile perception". In: *Psychological Science* 17.4, pp. 271–272.

Robertson, L. et al. (1997). "The interaction of spatial and object pathways: Evidence from Balint's Syndrome". In: *Journal of Cognitive Neuroscience* 9.3, pp. 295–317.

Ross, J. (2003). "Visual discrimination of number without counting". In: *Perception* 32, pp. 867–870.

Simon, T. and S. Vaishnavi (1996). "Subitizing and counting depend on different attentional mechanisms: evidence from visual enumeration in afterimages". In: *Perception & Psychophysics* 58.6, pp. 915–926.

Stanislas, D. and C. Laurent (1994). "Dissociable mechanisms of subitizing and counting: neuropsychological evidence from simultanagnosic patients". In: *Journal of experimental psychology* 20.5, pp. 958–975.

Szelag, E. (1997). "Temporal Integration of the Brain as Studied with the Metrono me Paradigm". In: *Time, Temporality, Now. Experiencing Time and Concepts of Time in an Interdisciplinary Perspective*. Ed. by H. Atmanspacher and E. Ruhnau. Berlin, Heidelberg: Springer-Verlag.

Trick, L. M. and Z. W. Pylyshyn (1994). "Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision". In: *Psychological Review* 101.1, pp. 80–102.

Whalen, J., C. R. Gallistel, and R. Gelman (1999). "Nonverbal counting in humans: The psychophysics of number representation". In: *Psychological Science* 10, pp. 130–137.

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445
http://www.rifanalitica.it

Determinismo causale e responsabilità morale: un approccio semicompatibilista

Lorenzo Testa

Abstract. Questo articolo si propone di analizzare il recente sforzo teorico di J.M. Fischer in merito alla compatibilità del determinismo causale con la responsabilità morale. Dopo l'analisi concettuale dei termini fondamentali della discussione contemporanea sul determinismo e la libertà del volere, metterò in luce l'originalità della teoria di Fischer basata sulla difesa del semicompatibilismo. Secondo tale prospettiva teorica è possibile essere responsabili del proprio agire anche nel caso in cui la verità del determinismo causale dovesse eliminare la presenza di possibilità alternative. Per sostenere questa tesi farò riferimento alla nozione di controllo (Fischer e Ravizza 1998) unitamente alla presentazione del controesempio elaborato da Frankfurt (1969) sulla non necessità di poter fare altrimenti per essere ritenuti responsabili del proprio agire.
Infine, dopo aver affrontato alcune critiche agli esperimenti mentali elaborati a partire da quelli di Frankfurt, mi concentrerò sugli attacchi diretti alla compatibilità fra responsabilità e determinismo, vale a dire sulle critiche che pur non rifacendosi alla necessità della possibilità di fare altrimenti cercano di mettere in discussione la compatibilità fra responsabilità morale e determinismo causale.

Keywords. Determinismo, John Martin Fischer, Harry Frankfurt, Libero arbitrio, Responsabilità morale, Semicompatibilsmo.

Copyright. © 2018 Lorenzo Testa. Pubblicato in Italia. Alcuni diritti riservati.
Autore. Lorenzo Testa, quaiomes@gmail.com.
Ricevuto. 30 Marzo 2018. **Accettato**. 28 Dicembre 2018.

1 Determinismo causale, libero arbitrio e semicompatibilismo: l'analisi concettuale

La maggior parte di noi ritiene, almeno a un livello pre-filosofico, che la possibilità di agire diversamente da come si è effettivamente agito sia un requisito necessario per l'attribuzione di responsabilità. Intuitivamente, cioè, si ritiene che se il corso degli eventi possa essere uno e uno solo, allora non avrebbe senso ritenere un agente responsabile del proprio comportamento: l'assenza della possibilità di agire diversamente renderebbe assurda l'attribuzione di responsabilità. Questo argomento è convincente nei casi di manipolazione diretta: un agente minacciato, ipnotizzato, condizionato o drogato (i casi di manipolazione possono essere molti, e di diversa specie) non viene ritenuto responsabile di ciò che fa poiché non avrebbe potuto fare altrimenti. Quanto appena notato è spesso in accordo con le nostre pratiche comuni: difficilmente si ritiene responsabile delle sue azioni un agente che abbia operato sotto il controllo di qualcuno o qualcosa. Ciò ha offerto la possibilità a diversi autori di estendere l'argomento a casi che a prima vista possono sembrare analoghi: ci si concentrerà in particolar modo sul tentativo di applicare lo stesso ragionamento all'ipotesi della verità del determinismo causale.

Prima di addentrarsi nella discussione di questo argomento, è bene chiarire i termini appena utilizzati. Per determinismo si intende, in generale, la teoria secondo la quale «ogni evento è determinato dal verificarsi di condizioni sufficienti per il suo accadere» (De Caro 2004, p. 11). Più specificamente, il determinismo di tipo causale sostiene che «un evento B è un effetto di altri eventi antecedenti A che lo necessitano, essendone cause sufficienti» (Magni 2005, p. 40). Non è raro che si faccia confusione fra questa versione di determinismo e l'argomento del fatalista, secondo cui qualunque scelta venga compiuta nel presente, il futuro si svolgerebbe comunque allo stesso modo, tanto che non farebbe nessuna differenza agire in una maniera o in un'altra (Moore 1912, p. 111). Il determinismo, al contrario, ammette queste differenze: la tesi deterministica non equivale alla teoria secondo la quale tutto sarebbe già "scritto" nel destino del mondo. Le leggi di natura, nell'ipotesi deterministica, non esercitano un diretto controllo sulle nostre vite e sulle nostre azioni (Dennett 1984).

Altra condizione necessaria allo sviluppo degli argomenti è la chiarificazione del concetto di libero arbitrio. Si tratta di un compito difficile, in particolar modo se si cerca di evitare di prendere direttamente posizione in favore del compatibilismo (secondo cui libero arbitrio e determinismo possono essere contemporaneamente veri) o dell'incompatibilismo (secondo cui libero arbitrio e determinismo sono mutualmente esclusivi). Seguendo Magni (2005, p. 47) è possibile individuare due condizioni necessarie al possesso del libero arbitrio senza entrare nella questione della sua compatibilità con il determinismo. Le due condizioni

sono:

1. «L'esistenza di *possibilità alternative*: la possibilità, cioè, di volere o scegliere altrimenti rispetto a come di fatto si vuole o si sceglie. Non è libero l'agente che può volere solo in un unico modo e non ha di fronte a sé più opzioni possibili.»

2. «Il *controllo* dell'azione e della scelta da parte dell'agente. [...] La volontà deve quindi dipendere da noi, essere in nostro potere (*up to us*).»

Un'ultima nota concettuale: sebbene il termine "compatibilismo" indichi tradizionalmente la posizione filosofica secondo la quale ad essere compatibili siano il determinismo e la libertà del volere, vi è un'altra accezione dello stesso termine, la quale è emersa nella discussione recente. Nella sua seconda accezione il compatibilismo indica la difesa della compatibilità fra determinismo e responsabilità morale: secondo certi autori, infatti, è possibile intendere la responsabilità morale in modo tale da renderla possibile in congiunzione con la verità del determinismo, anche nel caso in cui quest'ultimo escluda la possibilità di fare altrimenti.

2 Essere responsabili in un mondo deterministico

2.1 Responsabilità morale senza possibilità di fare altrimenti: la proposta del semicompatibilismo

È possibile essere responsabili delle proprie azioni senza possedere la possibilità di fare altrimenti? O, in altri termini, è vero che la possibilità di fare altrimenti sia una condizione necessaria per essere responsabili? Si potrebbe pensare di individuare come requisito necessario all'attribuzione di responsabilità la seguente condizione: si è agenti responsabili se si ha accesso a corsi di azioni differenti, se cioè è possibile per l'agente scegliere quale corso di azioni innescare. Se l'agente non possiede la possibilità di compiere questa decisione, allora non è responsabile delle sue azioni. Come nota Fischer (1994, p. 99), in questo requisito sono presenti almeno due condizioni distinte. La prima è che esistano possibilità alternative accessibili all'agente, la seconda che l'agente eserciti egli stesso il controllo sulla scelta del corso di azioni da innescare. Se non fosse rispettata la prima condizione l'agente non avrebbe la possibilità di fare altrimenti, mentre se fosse la seconda condizione a mancare non ci sarebbero ragioni sufficienti per ritenere l'agente responsabile, dato che egli non ha compiuto la scelta (ma essa è frutto, ad esempio, del caso).

Il progetto di Fischer è quello di mostrare come il requisito necessario per esser ritenuti responsabili delle proprie azioni sia il possesso di un determinato tipo di controllo, e non l'accesso a possibilità alternative (Fischer e Ravizza 1998).

Il problema sollevato da Fischer è che il concetto di controllo sia più complesso di quello che possa apparire a prima vista: egli propone due versioni dello stesso esperimento mentale per esporre questa sua tesi (Fischer 1994, pp. 135-150).

Si immagini un conducente intento a guidare la propria automobile lungo una strada. Ad un certo punto il guidatore intende girare il volante a sinistra per svoltare in quella direzione: poiché la sua macchina funziona correttamente, il guidatore gira il volante a sinistra e imbocca la strada in quella direzione. Sembra plausibile ritenere che il guidatore sia responsabile della sua azione di girare il volante a sinistra. Sembra anche plausibile ritenere che il guidatore eserciti il totale controllo del suo mezzo: se avesse scelto di svoltare a destra, avrebbe girato il volante in quella direzione e avrebbe imboccato una strada diversa. Stando a questo primo semplice caso si potrebbe essere indotti a pensare che se il conducente non avesse potuto fare altrimenti, non sarebbe stato responsabile della sua azione e quindi della direzione della sua automobile. Dunque l'accesso a corsi di azioni differenti sarebbe condizione necessaria all'attribuzione di responsabilità.

Le cose però possono non essere così semplici: un caso simile a quello appena proposto, ma con una rilevante differenza, sembra mettere in discussione la nostra intuizione preliminare.

Si immagini che il protagonista di questo nuovo esperimento mentale stia guidando un'automobile guasta. In particolare, l'autovettura è difettosa nella trasmissione del comando impartito dal volante alle ruote nel modo seguente: nel caso in cui il conducente decidesse di svoltare a destra, il mezzo non risponderebbe al suo comando e svolterebbe nella direzione opposta (la situazione inversa non accadrebbe invece se il conducente decidesse di svoltare a sinistra). Considerando anche che il conducente non sia a conoscenza del fatto che la sua automobile sia guasta, si immagini la seguente situazione: il guidatore decide di svoltare a sinistra al bivio, gira il volante verso quella direzione e svolta senza alcuna differenza con il caso precedente. Avrebbe cioè esercitato il controllo del mezzo in un senso che ci sembra necessario per l'attribuzione di responsabilità. Se però il conducente avesse deciso di svoltare a destra, egli avrebbe girato il volante in quella direzione, eppure l'automobile avrebbe preso la direzione opposta alla sua volontà e al suo comando impartito al mezzo. In questo caso allora ci sembrerebbe scorretto affermare che il conducente sia responsabile della direzione dell'auto, benché il risultato finale sia sempre lo svoltare a sinistra.

Ciò che l'esperimento proposto dovrebbe mettere in luce è la distinzione fra due tipi di controllo differenti: Fischer chiama «regulative control» il tipo di controllo che incorpora la possibilità di fare altrimenti, e «guidance control» quello che non include la possibilità di fare altrimenti (Fischer 1994, p. 34). Il *regulative control*, a ben guardare, può essere considerato come una versione "potenziata" del *guidance control*: esso infatti comprende la possibilità di esercitare il

guidance control in corsi di azioni alternativi a quello attuale.

Solitamente (come mostra l'esperimento mentale del primo guidatore, che rappresenta una situazione piuttosto comune) i due tipi di controllo non si presentano chiaramente distinti, e questo potrebbe essere all'origine della difficoltà dell'analisi del concetto di controllo. Il secondo caso proposto mostra però come possa darsi il caso in cui si sia in grado di esercitare *guidance control* senza il possesso del *regulative control*: nel secondo caso esaminato, infatti, se il conducente decidesse di svoltare a sinistra lo potrebbe fare senza difficoltà, e sembrerebbe plausibile ritenerlo responsabile della sua decisione e della sua azione. L'elemento importante da notare è che egli non aveva però la possibilità di fare altrimenti: se avesse deciso di svoltare a destra, l'automobile avrebbe rivelato il suo guasto producendo comunque una svolta a sinistra. In questa ipotesi – che nel caso proposto non si presenta effettivamente – non si sarebbe giustificati a ritenere il conducente responsabile. Questa differenza nell'attribuzione di responsabilità è spiegabile grazie alla distinzione che Fischer propone fra *guidance* e *regulative control*: nel primo esperimento mentale il conducente esercita contemporaneamente i due tipi di controllo, nel secondo esercita solo il *guidance control*. Il *regulative control* è inscindibilmente legato alla possibilità di fare altrimenti, mentre il *guidance control* può esercitarsi anche nel caso in cui manchi tale possibilità (Fischer 1994). Nel secondo caso, infatti, il conducente non avrebbe potuto fare altrimenti per via del guasto alla macchina, eppure parrebbe corretto ritenerlo responsabile della decisione e della successiva azione di svoltare a sinistra. L'esempio proposto sembra mettere in crisi la necessità di poter agire diversamente per esser ritenuti responsabili[1].

In altri termini, è possibile cioè per Fischer analizzare la questione della responsabilità concentrandosi sulla *actual sequence*: potrebbe non essere necessario, per esser responsabili del proprio agire, invocare scenari alternativi a quello che effettivamente si è svolto. Se concentrandosi sulla *actual sequence* si riuscisse a mostrare come sia possibile ritenere qualcuno responsabile del proprio agire, allora si arriverebbe alla conclusione della non necessità della possibilità di fare altrimenti per la responsabilità morale: questo è in fondo l'obiettivo dello sforzo teorico di Fischer e del suo semicompatibilismo (Fischer e Ravizza 1998).

Stando a quanto esposto fino ad ora, però, non sembra difficile muovere una critica a prima vista capace di destabilizzare fortemente la proposta di Fischer. Il conducente dell'esperimento mentale costruito poco sopra sarebbe in realtà dotato della possibilità di fare altrimenti: egli avrebbe potuto, infatti, provare a girare il volante a destra, girarlo effettivamente, e soltanto in seguito rendersi conto di non poter esercitare il proprio controllo sulla direzione del mezzo. Il critico di Fischer potrebbe mostrare che anche nel secondo caso proposto il conducente stia esercitando il *regulative control* associato alla possibilità di fare

[1] Un esempio molto simile è riscontrabile già in Locke (1689).

altrimenti, e che quindi tale tipo di controllo sia necessario all'attribuzione di responsabilità. Questa critica sarebbe corretta se non fosse possibile elaborare un esperimento mentale in grado di eliminare anche quel tipo di possibilità alternative lasciate aperte dal caso del conducente dell'automobile, ma come si vedrà nel paragrafo successivo, Frankfurt (1969) ha dato il via a una serie di esperimenti mentali capaci di far fronte a tale critica.

2.2 I controesempi *à la* Frankfurt al Principio delle possibilità alternative

Si consideri il seguente caso, che rappresenta un tipo di esperimento mentale *a là* Frankfurt. In questa situazione, il cittadino Marco deve esprimere la sua preferenza di voto potendo scegliere fra due candidati, A e B[2].Anche in questo caso vengono costruiti due diversi scenari.

Nel primo scenario, il giorno delle elezioni Marco entra nella cabina elettorale ancora indeciso sul candidato che deciderà di votare: al momento dell'espressione del suo parere sceglie di dare il proprio voto ad A, lo vota e restituisce la propria scheda. Sembra plausibile sostenere che Marco sia responsabile del proprio voto a favore di A.

Nel secondo scenario Marco ha subito un'operazione qualche giorno prima delle elezioni. Durante l'operazione il chirurgo, tale dottor Ferri, impianta all'insaputa di Marco un meccanismo nel suo cervello. Questo meccanismo nascosto non solo è in grado di monitorare le attività cerebrali di Marco e di comunicarle al dottore, ma è anche in grado di permettere al chirurgo di influire sui meccanismi cerebrali di Marco. Arrivato il momento di votare, Marco entra nella cabina elettorale ancora indeciso sul candidato che voterà. Quando deve esprimere il proprio parere, esattamente come nel primo scenario, decide che voterà per A, vota per A e restituisce la scheda elettorale. A differenza del primo scenario, però, se Marco avesse deciso di votare per B lo scienziato avrebbe captato la sua intenzione attraverso il meccanismo segreto. Si ponga che il chirurgo abbia inserito il meccanismo nel cervello di Marco proprio per evitare che egli voti per B: se Ferri avesse captato l'intenzione di Marco di votare per B, allora avrebbe avuto il potere di modificare l'intenzione di Marco per fare in modo votasse per A. Questo però non accade, dato che Marco decide spontaneamente di votare per A. Dato che Marco ha votato per A senza l'intervento di alcun manipolatore esterno, sembra corretto ritenerlo responsabile della sua azione, sebbene egli non potesse agire altrimenti (se avesse deciso di votare per B, infatti, lo scienziato avrebbe modificato la sua intenzione trasformandola nell'intenzione di votare per A).

[2]L'esempio è ispirato da Fischer (1994).

L'esperimento mentale appena proposto sembra sfuggire alla critica evocata alla fine del paragrafo precedente: in un caso *à la* Frankurt, infatti, l'agente non potrebbe nemmeno provare ad agire diversamente da come agisce, dato che il manipolatore esterno sarebbe in grado di modificare l'intenzione dell'agente non appena essa venisse formulata dall'agente stesso.

È parso a diversi autori che Frankfurt abbia fornito un valido argomento in favore della compatibilità fra determinismo causale e responsabilità morale. Il motivo risiede nel fatto che sembra plausibile ritenere responsabili delle proprie decisioni e delle proprie azioni i protagonisti degli esperimenti mentali *à la* Frankfurt, sebbene tali esperimenti mentali eliminino la possibilità dei soggetti stessi di fare altrimenti. Tuttavia, l'idea che Frankfurt abbia aperto la strada alla compatibilità fra determinismo e responsabilità morale è stata messa in discussione in diversi modi. Fischer riassume le critiche in forma di dilemma:

> Gli esperimenti mentali *à la* Frankfurt presuppongono o che il determinismo causale sia vero o che sia falso. Nel primo caso è tutto da dimostrare che l'agente sia realmente responsabile, mentre nel secondo caso è semplicemente falso che l'agente non abbia possibilità di agire altrimenti. (Fischer 2006, p. 126)

Vale la pena affrontare questo dilemma, poiché mette seriamente in discussione la validità degli esperimenti mentali *à la* Frankfurt e, con essi, la plausibilità di una teoria della responsabilità morale capace di fare a meno della possibilità di fare altrimenti.

3 Il dilemma dei controesempi *à la* Frankfurt

3.1 La critica indeterministica e i *flickers of freedom*

Secondo i sostenitori dell'indeterminismo causale, la preoccupazione maggiore per l'imputabilità della responsabilità morale è la presenza di possibilità alternative. Secondo la proposta di alcuni fra questi autori (Kane (1985), Ginet (1996)) sono in realtà presenti possibilità alternative anche nei controesempi costruiti alla maniera di Frankfurt: tali esperimenti mentali mancherebbero pertanto il bersaglio qualora cercassero di dimostrare che la presenza di possibilità alternative non sia un requisito necessario alla responsabilità morale, per il fatto che in realtà *ci sono* possibilità alternative anche in tali esperimenti mentali, nonostante le apparenze.

Nel caso proposto di Marco e del dottor Ferri, quest'ultimo sarebbe capace di intervenire solo sulla base dell'individuazione dell'intenzione di Marco di votare per B. Seguendo Kane (1985) e Ginet (1990) questo creerebbe il seguente problema: Marco dovrebbe mostrare al tempo T1 un qualche segno (per esempio un

certo *pattern* neuronale) per permettere al malvagio neurochirurgo di captare la sua intenzione di votare per B al tempo T3, cosicché possa intervenire al tempo T2 per modificare l'intenzione di voto di Marco. Se le cose stanno così, pare che l'indeterminista abbia trovato un buon argomento contro gli esperimenti mentali *à la* Frankfurt: al tempo T1 (cioè quello della formazione del *pattern*) Marco avrebbe potuto fare altrimenti, ad esempio avrebbe potuto mostrare un diverso segnale. Avrebbe cioè avuto un barlume di libertà [*Flicker of freedom*], e quindi anche nei casi elaborati a partire dall'esperimento mentale di Frankfurt sarebbero presenti possibilità alternative. Se tutto ciò è vero, affermare sulla base dell'argomento di Frankfurt che la responsabilità non richiede possibilità alternative è falso, dato che esse sono in realtà presenti.

La risposta di Fischer è tanto semplice quanto, a mio parere, convincente: questi *flickers of freedom* (Fischer 1994, pp. 131-159) sembrano troppo deboli perché si possa parlare di responsabilità morale. I difensori degli esperimenti mentali *à la Frankfurt* tendono a minimizzare i *flickers of freedom*, ad esempio sostituendo il diverso *pattern* neuronale esibito con un improvviso rossore sulla pelle del viso nel caso in cui l'agente avesse mostrato l'intenzione di votare per B. Questa mossa sembra mettere in difficoltà il critico di Frankfurt: saremmo disposti a ritenere qualcuno responsabile per aver esibito sulla propria pelle un rossore improvviso, per giunta in modo involontario? La discussione tuttavia può essere portata ad un livello di maggior complessità.

Si provi ad esempio a immaginare una situazione in cui l'agente non sia responsabile del proprio agire (per esempio, il caso in cui l'agente sia stato costretto, drogato o ipnotizzato). Se si aggiungesse a tale situazione un *flicker of freedom*, saremmo portati a modificare il nostro giudizio sull'agente e ritenerlo responsabile? Considerando che il barlume di libertà possa consistere in un differente *pattern* neuronale, o in un altro piccolo atto involontario, sarebbe difficile ritenere l'agente responsabile sulla base di questa sua possibilità di fare altrimenti. Si pensi a un cleptomane che commetta un furto perché spinto da un irresistibile impulso a portare a compimento quel gesto: se – come da ipotesi – l'impulso fosse genuinamente irresistibile, non saremmo giustificati a ritenerlo responsabile della sua azione. Si aggiunga ora la possibilità, da parte del cleptomane, di mostrare una configurazione neuronale leggermente diversa da quella che si configura nella *actual sequence*. Dovremmo modificare il nostro giudizio e ritenerlo responsabile in virtù di questa possibilità alternativa, la quale non è in suo potere? Riallacciandoci alla distinzione fra *guidance* e *regulative control*, è possibile notare come anche ammettendo che la presenza di *flickers of freedom* generi corsi di azioni alternativi a quello attuale la validità teorica degli esperimenti mentali *à la* Frankfurt non verrebbe compromessa. Negli scenari ispirati all'esperimento mentale di Frankfurt, infatti, l'agente non possiede alcun tipo di controllo sul "segnale" che esibisce, permettendo a Black di captare

le sue intenzioni ed eventualmente modificarle. Mancando il *guidance control*, nel quadro teorico proposto da Fischer sarebbe ingiustificato ritenere responsabile l'agente del suo *pattern* neuronale o del rossore improvviso sulla sua pelle, e quindi sarebbe un errore considerare le possibilità alternative offerte dai *flickers of freedom* come rilevanti per l'attribuzione di responsabilità.

Dunque, anche accettando che gli esperimenti mentali *à la* Frankfurt non dovessero riuscire a eliminare ogni residuo di possibilità alternative[3], Fischer tenta di mostrare come essi siano in grado di eliminare quanto meno le possibilità alternative rilevanti sul piano dell'attribuzione di responsabilità. Questo però può essere considerato un risultato importante per l'analisi della responsabilità morale, tanto importante da poter essere usato come un'efficace risposta agli argomenti dell'indeterminista che utilizzi la strategia dei *flickers of freedom*. Con le parole di Fischer: «Dunque ritengo che o si possano eliminare del tutto le possibilità alternative – anche se l'indeterminismo fosse vero – o che le rimanenti possibilità alternative non siano sufficientemente robuste» (Fischer 2006, p. 128).

3.1.1 La critica di K. Wyma e la discussione con J. Fischer

L'argomento di Fischer ha ricevuto una critica interessante che vale la pena discutere. Wyma (1997) ricorda un episodio accadutogli durante l'infanzia. Imparando ad andare in bicicletta, tentò un giorno di compiere un breve percorso senza l'ausilio delle rotelle. Suo padre, temendo per la sua incolumità ma al tempo stesso desideroso di lasciare al figlio la possibilità di dimostrare di essere ormai capace di andare in bicicletta, lo segue da vicino pronto a intervenire nel caso in cui si presentasse il rischio di una caduta. Al primo segnale di barcollamento, il padre sarebbe intervenuto aiutando il figlio a non cadere dalla bicicletta. Wyma ricorda di aver percorso il tragitto che si era prefissato senza mostrare segni di tentennamento e senza ricevere aiuti dal padre: era dunque riuscito nel suo intento per conto proprio, senza che tuttavia avesse la possibilità di fallire, dato che il padre lo avrebbe prontamente aiutato. L'analogia con gli esempi *à la Frankfurt* è facile da individuare: nel caso proposto da Wyma il padre ricopre il ruolo del dottor Ferri, pronto a intervenire nel caso in cui il corso degli eventi stesse per andare diversamente da quanto desiderato. Wyma individua un legame con il suo ricordo e la discussione a proposito dei *flickers of freedom* con l'obiettivo di confutare la proposta di Fischer di considerare come moralmente irrilevanti i *flickers* stessi. Wyma concorda infatti con Fischer nel ritenerli involontari: né il rossore improvviso sul viso né lo sbandamento sulla bicicletta possono essere riconosciuti come atti volontario. Tuttavia, secondo

[3] Elzein (2017) ritiene, forse poco cautamente, che sia ormai pressoché universalmente riconosciuta l'impossibilità di creare uno scenario in cui ogni possibilità alternativa sia stata eliminata.

Wyma noi riteniamo *prima facie* le persone moralmente responsabili del loro agire, e siamo pronti a rivedere i nostri giudizi nel caso in cui si venga a conoscenza di nuovi fattori capaci di modificare la valutazione della situazione. In questo modo i *flickers of freedom* sembrano generare un problema difficile da risolvere: per Wyma essi ci mostrano che Marco stava per decidere di votare per B, e a causa di ciò il dottor Ferri interviene. Per Fischer, come sappiamo, il flicker non rappresenta una possibilità alternativa sufficientemente robusta per fondare i nostri giudizi di responsabilità. Quindi, seguendo Fischer, il fatto che Marco *fosse sul punto di* compiere la decisione di votare per B (e che quindi esibisse involontariamente sul proprio viso una macchia di colore rosso), non è sufficiente per inficiare la bontà dell'argomentazione di Frankfurt rivolta contro il PAP [*Principle of Alternate Possibilities*]. Per Wyma, invece, Fischer cade in errore quando cerca di difendere i controesempi *à la* Frankfurt dalle critiche che si rifanno alla presenza di *flickers of freedom*: è pur vero che l'agente degli esperimenti mentali compie un'azione in fondo inevitabile (per la presenza del dottor Ferri, nel nostro esempio), ma egli è responsabile proprio perché esistono possibilità alternative moralmente rilevanti, anche se esse non sono né decisioni, né azioni.

In altre parole, la discussione fra Wyma e Fischer trova il suo fulcro nella rilevanza morale assegnata ai *flickers of freedom*: per Wyma, anche ammettendo l'involontarietà dei *flickers* e il loro essere diversi da una decisione, essi hanno rilevanza morale in quanto mostrano che l'agente era sul punto di decidere di agire diversamente. Per Fischer, invece, i *flickers of freedom* non hanno sufficiente "robustezza" per poter essere considerati moralmente rilevanti: qualcosa che non rientra né nel campo delle azioni né nel campo delle decisioni pare non possedere la rilevanza morale necessaria per venire considerata la base su cui fondare le attribuzioni di responsabilità degli individui. In definitiva: se si considerano i *flickers of freedom* moralmente rilevanti al pari delle decisioni (pur non essendolo), allora i sostenitori degli esperimenti mentali *à la* Frankfurt si trovano in seria difficoltà, dato che gli scenari da essi proposti falliscono nel presentare situazioni in cui all'agente non siano accessibili possibilità alternative capaci di modificare l'attribuzione di responsabilità.

Per rispondere a Wyma è necessario rifarsi alla nozione di *guidance control* esposta nel paragrafo precedente. Nell'esempio del giovane Wyma alle prese col suo primo percorso in bicicletta senza l'ausilio delle rotelle, scrive Fischer, come «non sia la possibilità di sbandare leggermente a rendere il successo di Wyma veramente suo. Ciò non ha nulla a che fare col fatto che egli avrebbe potuto sbandare un poco, ma con il modo in cui egli ha guidato la bicicletta». Per Fischer, cioè, rimane irrilevante il fatto che Wyma potesse sbandare leggermente (cioè avere accesso a corsi di azioni differenti): non è su questa base che riteniamo l'azione del ciclista alle prime armi un successo. È piuttosto il fatto che Wyma

abbia esercitato un certo tipo di controllo – il *guidance control* – ad aver reso l'azione dell'agente una sua azione svolta per conto proprio. Ancora una volta, cioè, non è sulla base dei *flickers of freedom* (come lo sbandare sulla bicicletta, secondo Wyma) che riteniamo un agente responsabile del proprio agire. In questo modo Fischer difende la tesi secondo la quale la possibilità di un agente di accedere a corsi di azione alternativi non rappresenti una condizione necessaria all'attribuzione di responsabilità.

3.2 La critica deterministica

A mio parere, l'accusa agli esperimenti mentali *à la* Frankfurt che veda in essi il determinismo come presupposto costituisce un problema più circoscritto. Se si supponesse fin dall'inizio la verità del determinismo, sarebbe ingiustificato affermare che Marco sia responsabile della sua intenzione di votare per A nella *actual sequence*. Dopotutto, la responsabilità dell'agente anche in mancanza di possibilità alternative è ciò che l'argomento vorrebbe provare: l'argomento di Frankfurt sarebbe pertanto *question begging*. I critici di Frankfurt possono infatti sostenere che in un mondo deterministico la responsabilità di Marco sia tutta da dimostrare, e che l'argomento di Frankfurt sia insufficiente nel fornire tale spiegazione. Assenza di possibilità alternative e determinismo non sono sinonimi: possono essere assenti possibilità alternative anche in un mondo indeterministico, così come – almeno secondo certe accezioni di possibilità – possono rimanere aperte possibilità alternative in un mondo deterministico. Se si tiene a mente questa distinzione (chiara già allo stesso Frankfurt) un difensore del PAP potrebbe sostenere che Frankfurt non abbia offerto un argomento valido nel caso in cui lo scenario dell'esperimento mentale fosse deterministico.

Esiste un modo piuttosto semplice di rispondere a questo tipo di obiezioni, giacché come scrive Fischer:

> I sostenitori di un compatibilismo basato sugli esperimenti mentali *à la* Frankfurt *non* sostengono che l'agente sia moralmente responsabile del suo comportamento, utilizzando come fondamento tali esperimenti mentali. Un suo sostenitore dovrebbe semplicemente dire 'Non sono a conoscenza del fatto che l'agente sia o meno responsabile del suo comportamento, ma se egli *non* lo è, *non* è per via dell'assenza di possibilità alternative'. (Fischer 2006, p. 128)

L'argomento di Frankfurt può venir formulato in modo da non difendere direttamente la responsabilità dell'agente: il punto messo in evidenza sarebbe piuttosto un altro. L'argomento di Frankfurt dovrebbe infatti mettere in luce il fatto che se noi consideriamo l'agente responsabile delle sue azioni, non lo facciamo sulla base del fatto che egli sia in possesso della possibilità di fare altrimenti. D'altra parte gli sforzi di Frankfurt sono concentrati sulla confutazione

del PAP, il che giustifica la replica di Fischer a chi ritenga gli esperimenti mentali *à la* Frankurt *question-begging* nel caso della verità del determinismo (Frankfurt 1969). Che si ritenga responsabile o meno l'agente, ciò che conta è che il giudizio non sia fondato sulla possibilità di fare altrimenti.

Ciò si può notare considerando una versione leggermente modificata del PAP, così formulata: «La mancanza di possibilità alternative è una condizione che da sola – senza considerazioni aggiuntive (siano elementi contingenti o necessari) – rende un agente non moralmente responsabile del suo comportamento» (Fischer 2006, p. 129). Se Frankfurt ha ragione, questo principio è falso, indipendente dalla verità del determinismo. Questo dovrebbe quindi servire come replica a chi ritiene che l'argomento di Frankfurt non sia convincente, sulla base del fatto che se il determinismo fosse vero sarebbe tutta da provare la responsabilità dell'agente nella *actual sequence*.

I problemi sollevati dall'approccio di Fischer non sono però finiti: se viene accettata la teoria semicompatibilista, bisogna comunque rendere conto della responsabilità morale nella *actual sequence*. Fino a questo punto, infatti, si è visto come la responsabilità morale non richieda la possibilità di fare altrimenti: non si è cioè ancora mostrato se davvero si possa ritenere responsabile un agente nella *actual sequence*. Il merito di Frankfurt, dopotutto, è quello di aver permesso agli autori successivi di prendere seriamente in considerazione la possibilità di concentrarsi sulla *actual sequence*, non quella di aver dato un resoconto completo della responsabilità dell'agente. Nel seguito della trattazione verranno considerati alcuni fra i più importanti argomenti contro la plausibilità della compatibilità fra responsabilità morale e determinismo che non si servano della possibilità di fare altrimenti. Questo tipo di critiche va sotto il nome di argomenti *diretti* contro la responsabilità morale nel caso della verità del determinismo, dove con questo termine si indica appunto l'approccio che non si serva della necessità di poter fare altrimenti. Un modo alternativo di individuare questa posizione è quello di considerare tale approccio come critico nei confronti di una concezione della responsabilità basata sulla *actual sequence*.

4 Gli argomenti diretti contro la responsabilità morale

In Kane (1996), Ekstrom (1998) e Pereboom (2001) si trovano diversi argomenti per attaccare direttamente – cioè senza invocare la presenza di possibilità alternative - la compatibilità fra responsabilità morale e determinismo causale. Come nei paragrafi precedenti, viene qui discussa l'argomentazione di Fischer (1994) a difesa del suo semicompatibilismo.

In Kane (1996) si trova un deciso rifiuto della possibilità di conciliare determinismo e responsabilità senza servirsi dell'esistenza di possibilità alternative: «[Gli agenti devono] avere il *potere di compiere decisioni che possano essere spiegate solamente e in ultima analisi come derivanti dal loro stesso volere* (ad esempio dal loro carattere, dai loro motivi e sforzi della volontà). Nessuno è in possesso di questo potere in un mondo deterministico» (Kane 1996, p. 54). Perché però in un mondo causalmente determinato si perderebbe la possibilità di ritenere responsabile delle proprie azioni un agente (se si esclude il problema delle possibilità alternative)? Questa sembra più una presa di posizione che una valida critica: anche Pereboom (2001) cade in un errore simile quando afferma che «se tutto il nostro comportamento si trovasse "già scritto" prima che nascessimo, nel senso che ciò che accade prima della nostra nascita - attraverso un processo causale deterministico - determina inevitabilmente il nostro comportamento, allora non possiamo essere biasimati in modo legittimo quando compiamo delle scelte sbagliate» (Pereboom 2001, p. 71). Anche qui però il compatibilista (e il semicompatibilista) potrebbe chiedersi perché il determinismo causale dovrebbe rendere nulla la possibilità di ritenere valida l'attribuzione di responsabilità (senza invocare la presenza di possibilità alternative). Sembra insomma necessaria un'argomentazione valida, piuttosto che il ricorso a quella che sembra una preoccupazione non chiarita attraverso l'analisi filosofica.

Molto spesso, gli incompatibilisti formulano un argomento valido nei casi di manipolazione diretta, estendendolo poi al determinismo. Questa estensione sembra però ingiustificata (o quantomeno problematica): se un esempio di manipolazione diretta come la coercizione elimina la possibilità di ritenere qualcuno responsabile, perché dovrebbe avvenire lo stesso nel caso della verità del determinismo causale? Dennett (1984) si trova un buon argomento basato sulla indebita antropomorfizzazione delle leggi di natura: esperimenti mentali basati su situazioni di diretta manipolazione non sono equiparabili al tipo di costrizione derivante dalla verità del determinismo causale[4]. Al determinista rimane infatti aperta la possibilità di distinguere fra diversi tipi di determinazione; la sua strategia è quella di affermare che il problema non risieda tanto *nel fatto che* l'agente sia causalmente determinato, quanto piuttosto nel *tipo* di determinazione (Dennett 1984, p. 12).

Si prendano in considerazione le seguenti parole tratte da (Kane 1985, p. 8): «Ciò che il determinismo nega è un certo senso di importanza dell'agente in quanto individuo». Si fa quindi riferimento all'importanza che l'individuo ha e alla minaccia rivolta a questa importanza dalla verità del determinismo. Secondo Mele (1999) un modo di esprimere la preoccupazione dell'incompatibilista sarebbe il seguente: un agente sarebbe indipendente, nel senso rilevan-

[4] L'uso stesso del termine "costrizione" può essere rilevante: sembra scorretto, dopo un'attenta analisi, affermare che le leggi di natura siano in grado di costringere qualcuno.

te, solamente se fosse in grado di dare un contributo alla spiegazione del suo comportamento, comportamento che non potrebbe venir interamente spiegato dall'insieme delle leggi di natura e dello stato di cose del mondo precedente alla percezione dell'agente di poter agire liberamente. Se si prende seriamente in considerazione la formulazione di Mele del concetto di indipendenza così come viene intesa dagli incompatibilisti, si noterà come la sfida al compatibilismo divenga quella di trovare un senso preciso in cui l'agente possa spiegare il suo comportamento senza che esso venga esaurito nelle leggi di natura e nello stato di cose del mondo.

Nota Fischer però che «anche se il determinismo fosse vero, il rifarsi a stati del mondo precedenti sommati alle leggi di natura non può spiegare il nostro comportamento e i suoi risultati senza *anche* spiegare che *noi diamo un certo contributo ad essi*» (Fischer 2006, p. 135). Il determinismo non è fatalismo: gli agenti fanno cioè parte della catena causale e il loro comportamento rientra nella descrizione degli stati di cose del mondo. La catena causale, nella teoria determinista, riesce comunque a mantenere la differenza fra le azioni che sono imputabili all'agente, e quelle che non lo sono (per esempio nei casi di manipolazione diretta, come la coercizione). Non c'è quindi bisogno di appoggiarsi all'indeterminismo per mantenere la differenza fondamentale fra tipi di azioni imputabili all'agente e tipi di azioni che non sono imputabili all'agente stesso.

Un ulteriore argomento degli incompatibilisti si può riscontrare in (Nozick 1981, p. 312): tale critica alla possibilità di conciliare determinismo e responsabilità morale senza servirsi della necessità delle possibilità alternative si serve di una intuizione comune apparentemente plausibile, e cioè che essere agenti liberi e responsabili significhi essere capaci di produrre una differenza nel mondo. Questo sembra un argomento sensato a prima vista, ma forse la sua plausibilità risiede in un'intuizione prefilosofica: un'analisi più dettagliata è capace di mettere in dubbio il principio per cui per essere liberi e responsabili sia necessario produrre una qualche differenza nel mondo.

Si pensi al caso seguente[5]: un pittore dipinge un quadro di grande importanza, che viene riconosciuto immediatamente come un'opera di grande successo dalla comunità dei critici d'arte ed esposto in prestigiosi musei. All'insaputa del pittore, se egli non avesse realizzato quel dipinto, un altro pittore avrebbe prodotto di lì a poco un quadro perfettamente identico al suo, che sarebbe stato ritenuto dello stesso grande valore e allo stesso modo sarebbe stato esposto negli stessi musei. L'esperimento mentale mette in luce un possibile controesempio all'affermazione di Nozick secondo la quale un segno distintivo del nostro essere agenti liberi sia quello di produrre una certa differenza nel mondo. Nel caso proposto, infatti, noi riteniamo responsabile il pittore della sua creazione artistica anche se un altro artista avrebbe potuto portare in essere lo stesso sta-

[5] L'esperimento mentale è tratto da Fischer (2012).

to di cose: il pittore non ha prodotto alcuna differenza significativa nel mondo (giacché, se non avesse dipinto il quadro, un altro lo avrebbe fatto al posto suo), eppure lo riteniamo responsabile della propria opera e lodevole per le sue doti. Sembra quindi che per essere responsabili non sia necessario "fare la differenza" nel mondo, e che Nozick non abbia analizzato a sufficienza la relazione fra responsabilità e influenza dell'agente sugli stati di cose esterni.

Ecco dunque che gli argomenti degli incompatibilisti non facenti riferimento all'esistenza delle possibilità alternative sembrano o essere *question-begging* e dettati più dalla preoccupazione della verità del determinismo che dalla sua analisi attenta, o passibili di critica da parte di un attento compatibilista.

Rivista Italiana di Filosofia Analitica Junior 9:2 (2018)

Riferimenti bibliografici

De Caro, Mario (1999). "Libertà metafisica e responsabilità morale". In: *Paradigmi* 17.51, pp. 519–546.

— (2004). *Il Libero Arbitrio. Una Introduzione*. Biblioteca di Cultura Moderna 1171. Roma: Laterza.

— (2014). "Analisi concettuale e scienza: il dibattito contemporaneo sul libero arbitrio". In: *Libero arbitrio. Storia di una controversia filosofica*. A cura di Emidio Spinelli Mario De Caro Massimo Mori. Roma: Carocci, pp. 365–382.

Dennett, Daniel (1984). *Elbow Room. The Varieties of Free Will Worth Wanting*. Cambridge, Mass: MIT press.

Ekstrom, Laura Waddell (1998). "Freedom, Causation and the Consequence Argument". In: *Synthese* 115.3, pp. 333–354.

Elzein, Nadine (2017). "Frankfurt-Style Counterexamples and the Importance of Alternative Possibilities". In: *Acta Analytica* 32.2, pp. 169–191.

Eshleman, Andrew (2016). *Moral Responsability*. A cura di Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: https : / / plato . stanford . edu / archives/win2016/entries/moral-responsibility/.

Fischer, John Martin (1985). "Frankfurt-style examples and semi-compatibilism". In: *Free Will and Values: Adaptive Mechanisms and Strategies of Prey and Predators*. A cura di Robert Kane. Albany: SUNY press, pp. 281–308.

— (1994). *The Metaphysics of Free Will*. Oxford: Blackwell.

— (2002). "Frankfurt-style compatibilism". In: *Contours of Agency: Essays on themes from Harry Frankfurt*. A cura di Sarah Buss e Lee Overton. Cambridge Mass: Mit Press, pp. 1–26.

— (2006). *My Way. Essays on Moral Responsibility*. Oxford: Oxford University Press.

— (2012). *Deep Control. Essays on Free Will and Freedom*. Oxford: Oxford University Press.

Fischer, John Martin e Mark Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Fonnesu, Luca (2014). "Libertà e responsabilità: dall'utilitarismo classico al dibattito contemporaneo". In: *Libero arbitrio. Storia di una controversia filosofica*. A cura di Emidio Spinelli Mario De Caro Massimo Mori. Roma: Carocci, pp. 337–363.

Frankfurt, Harry G. (1969). "Alternate Possibilities and Moral responsability". In: *Journal of Philosophy* 66, pp. 828–839.

Frankfurt, Harry G. (1971). "Freedom of the Will and the Concept of a Person". In: *Journal of Philosophy* 68, pp. 5–20.

Ginet, Carl (1990). *On Action*. Cambridge: Cambridge University Press.

— (1996). "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing". In: *Philosophical Perspectives* 10, pp. 403–417.

Kane, Robert (1985). *Free Will and Values*. Albany: SUNY press.

— (1996). *The Significance of Free Will*. Oxford: Oxford University Press.

Locke, John (1689). *An Essay concerning Human Understanding*. A cura di Peter H. Nidditch. Oxford: Clarendon Press, 1990.

Magni, Sergio Filippo (2005). *Teorie della libertà. La discussione contemporanea*. Roma: Carocci.

McKenna, Michael e Justin D. Coates (2016). *Compatibilism*. A cura di Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/archives/win2018/entries/compatibilism/.

Mele, Alfred R. (1999). "Kane, luck, and the significance of free will". In: *Philosophical Explorations* 2.2, pp. 96–104.

Moore, George E. (1912). *Ethics*. Oxford: Clarendon Press.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge Mass: Harvard University Press.

Pereboom, Derk (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Van Inwagen, Peter (1983). *An Essay on Free Will*. Oxford: Oxford University Press.

Wyma, Keith D. (1997). "Moral Responsability and Leeway for Action". In: *American Philosophical Quarterly* 34.1, pp. 57–70.