Can sentences self-refer? Gödel and the liar

Rupert Read
University of East Anglia
Christian Greiffenhagen
The Chinese University of Hong Kong

Abstract

In this article we discuss the issue of 'self-reference', i.e., the question whether (or in which sense) sentences may be said to refer to themselves. Following Wittgenstein, we suggest that the clearest thing to say is that sentences cannot *of themselves* 'do' or 'say' anything, but that it is *human beings* that 'do' and 'say'. Consequently, instances of self-reference have to be considered as part of specific human practices. We illustrate these general remarks through the examination of the Liar Paradox and Gödel's Incompleteness Theorem (which uses a formally undecidable sentence, which is sometimes taken to 'say' 'I am not provable.'). We emphasise that in the context of the 'foundations of mathematics' it is important to separate technical-mathematical from philosophical questions, and argue (again following Wittgenstein) that Gödel's Incompleteness Theorem was more a contribution to the former than the latter. In other words, Gödel's result runs the risk of being over-interpreted, and of falling foul of Wittgensteinian philosophy, if it is interpreted philosophically to include focally a sentence that literally self-refers.

INTRODUCTION

> Asking what the sense is. Compare:
> 'This sentence makes sense.'—'What sense?'
> 'This set of words is a sentence.'—'What sentence?'
> (Wittgenstein 1953: §502)

Methodological considerations

We offer in what follows what might be helpfully described as a 'respecification' (Garfinkel 2002) of a question that is important to the comprehension of a famous mathematical result. Drawing on Wittgenstein, we achieve that re-specification by re-examining (and ultimately questioning) that question and placing it in the context of actual human practices and what they (can) do.

The paper we have written concerns, focally, the ordinary meaning of devices of 'self-reference'. It thus connects and contrasts mathematical logic with ordinary life. Based on what we have just said, by now it is probably obvious to readers of this journal how Wes Sharrock's work can be seen to be very much in the ballpark of our enquiry. And, down the years, we both have profited from Wes's thoughts on the subject-matter of this paper.[1] In fact, this paper might well have proved impossible without Wes's intellectual inspiration. Before getting into the substance of the paper, we'd like to take a moment to suggest to the reader an intellectual background through which to view our enquiry; a background that emerges from the significant portion of each of our careers that has encompassed joint work with Wes.

In Rupert's case, that work has focussed particularly on Thomas Kuhn and Peter Winch, the great practice-oriented, Wittgensteinian 20th century philosophers of science and 'social science.' In the present context, we'd like especially to draw attention to the vital role, as we (and Wes) understand it, of Wittgensteinian considerations about the actual nature of rule-following (and of innovation) in the work of these two philosophers of the sciences. We are thinking here especially of Section V of Kuhn's (1962) *The Structure of Scientific Revolutions*, 'The priority of paradigms', and of Section III 5–6 of Winch's (1958) *The Idea of a Social Science and its Relation to Philosophy*. It will be helpful to bear these (and the kinds of things that Wes and Rupert co-wrote about them, in their *Kuhn: Philosopher of Scientific Revolution* [Sharrock and Read 2002] and *There is No Such Thing as a Social Science: In Defence of Peter Winch* [Hutchinson et al. 2008]) in mind in what follows. Reminding oneself of the mundane and normally effortless nature of rule-following by practicing scientists working within a paradigm—and by ordinary members of society—may help to prevent one from slipping into what we will claim is the error of thinking that something 'special' must successfully be happening when the Gödel sentence is created / written down; something achieved 'by the sentence itself' rather than by someone figuring out how (if at all) it can actually be used intelligibly, or explained in ordinary prose.

Rupert started working with Wes in 1995–6, when he arrived at Manchester and became a philosophical member of the ethnomethodologically-oriented intellectual community there.

For many years, Christian has worked with Wes on a variety of sociological and conceptual investigations of mathematics, which are based on the two key inspirations for Wes's work: again Wittgenstein, in this case, his philosophy of

[1] Thus our greatest debt in writing this paper is to him. Big thanks also to the editors of this special issue, and to an external referee.

mathematics (Wittgenstein, 1978; see Shanker 1987), and Garfinkel, in this case ethnomethodological studies of the natural sciences and mathematics (e.g., Garfinkel 1986; Lynch 1985; Livingston 1986). Wes has applied insights from both to a multitude of problems and topics in the social sciences. In our work on mathematics, this has meant to emphasize that any attempt to understand the nature of mathematics has to start with the *practices* of arithmetic and mathematics. Furthermore, Wes has drawn attention to the fact that when it comes to working with empirical materials, handling those materials is by no means straightforward. In particular, when one examines the theoretical or conceptual conclusions that are supposedly derived from empirical evidence, one very often notices a 'gap' between the presented evidence and the theoretical or conceptual conclusions. With respect to psychology, Wittgenstein (1953: §371) famously claimed that 'there are experimental methods and conceptual confusion', where that confusion pertains to interpreting the data arrived on the basis of experimental methods. With respect to sociology, Garfinkel (1967) demonstrated how sociologists relied upon common-sense reasoning and knowledge as an unacknowledged resource to derive their sociological theories from empirical evidence. In the Wednesday lunchtime meetings at the staff club at Manchester University, in Friday discussions at the Grafton pub, as well as other both formal and informal discussions, Wes has taught these important lessons to colleagues, students, and visitors.

Christian started to work with Wes in 2002, when they embarked on a year-long journey trying to understand Eric Livingston's (1986) *The Ethnomethodological Foundations of Mathematics*: Wes teaching Christian about both introductory and advanced Wittgenstein, as well as the intricacies of ethnomethodology, Christian using his background in mathematics to answer questions about formal systems, formal proofs, and the foundations of mathematics (see Greiffenhagen and Sharrock, forthcoming). In our first 'official' project on 'The Role of the Notion of 'Social Practice' in Philosophy and Sociology of Mathematics', funded by the Arts and Humanities and Research Board (AHRB), for which Rupert was a Co-Investigator, we examined the evidence for claims that logic, grammar, or arithmetic might be culturally relative (Greiffenhagen and Sharrock 2006a, b; 2007). By using a Wittgensteinian perspective to re-read the empirical case studies that were the basis of the claims of logical, mathematical, or linguistic relativism, Wes and Christian showed that those claims were not derived from the presented the empirical evidence, but rather were the consequence of philosophical assumptions.

The present paper originated during this period and drew on similar conceptual resources. (In subsequent projects, Wes and Christian tried to see what an empirical investigation of mathematical practice might look like, and conducted video-based ethnographic studies of the presentation of mathematics in lectures as well as the creation of mathematics in doctoral supervision [Greiffenhagen and Sharrock 2011].)

Summary of our argument

The argument that we develop in this paper is that words (and consequently sentences) do not have any effect in and of themselves (even if understood to operate 'within a system'), but that it is human actions and practices that ensure that words and sentences have meaningful consequences.[2] It is only by losing sight of human action that one becomes enchanted by words and symbols which seem to be altogether out of the control of humans, and which seem to mean whatever they mean and do whatever they do, magically, without any at-least-implicit reference to a speaker, community, or context. We feel that such enchantment is particularly likely in fields such as mathematics where the procedures of the subject are designed so as to eliminate any seemingly human elements.

We will try to illustrate this by focussing on the issue of 'self-reference', which has been the subject of much debate (and controversy) in the philosophy of language and mathematics. Sentences such as 'I am a liar.', 'I am not provable.', or 'This sentence makes sense.' seem to lead to all sorts of paradoxes that need to be explained and accounted for by philosophers. Offering a radically Wittgensteinian twist on some seemingly familiar territory, we want to show that these paradoxes themselves are based on a metaphysical picture of how words operate. We will suggest that such a picture is not compulsory and is in fact far less desirable than might at first appear.

Not sentences but humans self-refer

Following Stanley Cavell,[3] one might helpfully say that there is a tendency, especially in philosophy, to avoid responsibility for one's thoughts and words, as well as other deeds. In particular, there is a tendency to put forward the view that it is not humans that 'say' certain things but the words *themselves*. However, in our view there is no way of understanding language (or any other human practice, including mathematics) that does not take into account that it is not words that do the speaking or doing but people.[4]

---

[2] Now of course, J.L. Austin taught us that there are some words that might seem to have an effect in and of themselves: words used 'performatively', as in 'I pronounce you husband and wife', etc. But this does not in the least contradict what we are saying: because such words actually only have their effects in very tightly specified contexts. Outside of the actual institutionalised practice of marriage, saying 'I pronounce you husband and wife' achieves nothing but a joke or a bizarreness-reaction or at best a gentle irony; for example, if I say it to two sheep in a field; or to a happily-married old couple.

[3] See his writings on 'acknowledgement' throughout his career, especially Cavell (1969); see also Denis McManus (2006).

[4] What do we mean by 'people'? Our use of the term 'people' is shorthand for people's actions and conventions; everything that Wittgenstein means when he talks of language games that are part of human ways of life. When we read a road sign or an anonymous instruction, or when we hear a

We therefore would like to suggest the importance of it remaining perspicuous that words in and of themselves cannot 'do' or 'say' anything (and that any suggestion to the contrary is merely metaphorical). Words need something else in order to have any effect. Words are properly always part of human practices and thus cannot be regarded as intelligibly as doing, in and of themselves, any of the extraordinary, often magical, things that are attributed to them. *A fortiori,* they cannot refer to themselves. Cook (1978: 23) puts it like this: 'it should strike us as absurd to maintain that sentences say something. Sentences don't talk; people do. And so sentences don't say anything, either.'

It is sometimes argued—and we go along with the argument, by and large, on Wittgensteinian grounds—that words only have sense within a context or system. However, that does not solve our problem, as again the system itself cannot 'do' anything. It is not the *system* which 'does' the relevant communicative action, but human *practice(s)*.

Wittgenstein's (1953) *Philosophical Investigations* (*PI*) and his (1978) *Remarks on the Foundations of Mathematics* (*RFM*) include many discussions of the kinds of paradox that may arise by losing sight of the role that human practice plays in language. Our argument in a sense applies Wittgenstein's question about 'calculating machines' to 'self-referring sentences':

> Does a calculating machine *calculate*? (*RFM*, V, §2, p. 257)

> If calculating looks to us like the action of a machine, it is *the human being* doing the calculation that is the machine. (*RFM*, IV, §20, p. 234)

ORDINARY SELF-REFERENCE

There has been a lengthy debate in such disciplines as Literary Studies and Philosophy over the exact nature of the (supposed) 'self-reflexivity' of many things: texts, sentences, utterances, words, or occasionally even objects. For a while it became almost a commonplace in English Departments that poems can and do (even: always do) refer to themselves, and in Fine Art Departments that artworks refer to themselves. And it is surely commonplace in Philosophy Departments that certain sentences can refer to themselves, in particular when couched in logically perspicuous languages.[5] In this section, we will therefore investigate sentences such as 'I

robocall or follow the instructions given by a GPS voice, these are intelligible because the language involved is composed, written, printed etc. by people, and these communications rely upon people to understand them without necessarily knowing anything in particular about who or what is responsible for producing them.

[5] An example at the intersection of all these is Michel Foucault's intriguing and highly amusing (but problematic) short work, *This Is Not a Pipe* (1983). The problem is that Foucault at times (e.g., p.27, p.30) appears to assume that statements really can refer of themselves ('*de re*') to themselves.

am lying.' or 'This sentence makes sense.', in order to get an initial grasp on the notion of 'self-reference'.

Metaphysical lines of projection

How *can* a proposition say something of itself, outside of the external *factum* of our determination that it should be read so to say? Is there a means in the language by which a sentence can (itself) be said to be pointing to itself? Is it possible to establish a language that is not vulnerable to human decision as to its meaning and interpretation?

If it were, how are we to make sense of the view that a sentence can point to itself? Are we to imagine 'metaphysical lines of projection' (cf., *PI* §141) going from the sentence back to itself (just as we might imagine such lines going from, e.g., the word 'sun' to the Sun)? Very well; let us endeavour to countenance the imagining of such lines:

"This sentence makes sense."

Figure 1

Then we might ask: Which sentence? And the answer would be to follow a (metaphysical) line:

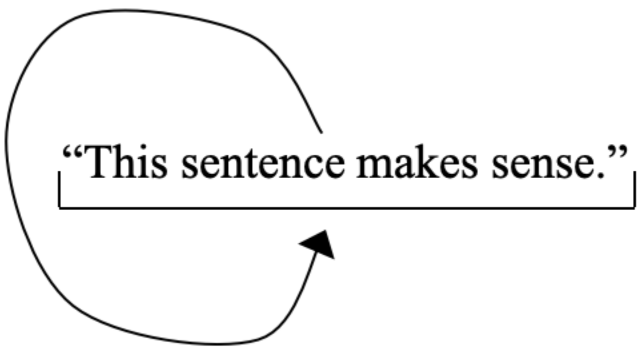

Figure 2

A happier way forward that avoids the problems in Foucault's account would be the work of ethnomethodologists such as Lena Jayyusi (1993), who makes the following point about scenes depicted by photographs: 'It is not that *the scene speaks for itself*; but *that it speaks itself* in such a way that it can provide the ground and object (the topic and resource) *of our speaking for and about it*, in the various ways that we do' (p. 45).

However, we might *still* want to ask: Which sentence? (because the question could arise as to whether or not the arrow was part of the sentence referred to):

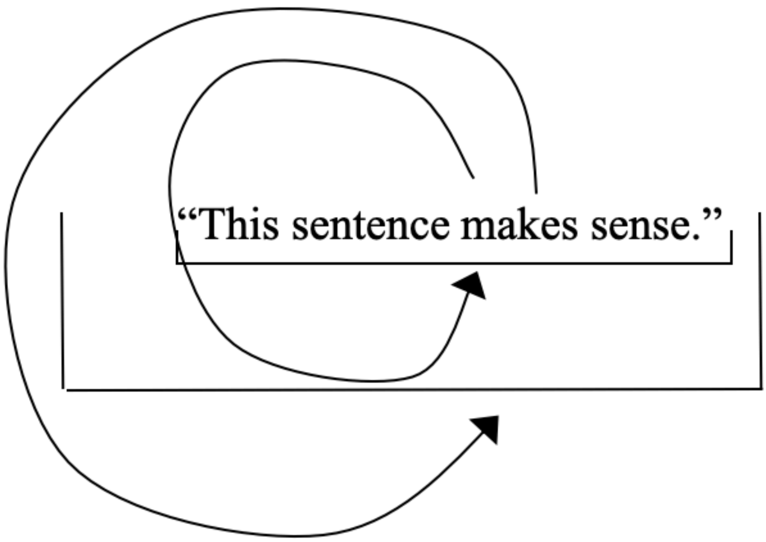

Figure 3

And so on and so forth…

We would argue that any putative instance of self-reference will involve some kind of perpetual ambiguity as shown in these figures.

It might be thought relevant (and problematic) that the example considered above involves a demonstrative ('This'). But this actually makes no difference. One could equally well substitute other methods of 'self-reference'. For example, the proposition could be written as 'The sentence on the blackboard does make sense.' (and be the only sentence on the blackboard). The same problem would apply, as one would still be able to ask whether the sentence was the same sentence after the lines of projection (ensuring that it was the sentence being referred to) had been added. And one would have to have the lines of projection, in one way or another, because without these arrows (or similar devices) there would be no *prima facie* plausibility to the claim that the sentence was 'pointing to' or 'hooking up with' itself *by itself* (i.e., without any involvement from language-*users*).

To self-refer a sentence cannot just simply stand there; it must *do* something. It must point to itself, or something similar. As we have just sought to show, the claim that a sentence is (by itself) saying something of itself is not an idea that has been often made very good sense of. It seems nearly always to be a case of (bad?) poetry and/or attempts of the kind of self-referentiality that Deconstructionist critics (arguably quite mistakenly, indeed incoherently) believe much Modern poetry accomplishes. Therefore, one surely does not *have* to read such a sentence as referring to itself.

If one reads it that way, one is *choosing* to do so, on the basis of imponderable evidence.

Wittgenstein's take on 'self-reference'

Wittgenstein provides us with the resources to show more thoroughgoingly that and how the mainstream conceptions of 'self-reflexivity' and 'self-referentiality' are in the main incoherent and based in the end on an (upon reflection) unattractive metaphysical picture of how language operates. Wittgenstein tries to show why the purported 'self-induced' or 'automatic' disambiguation of any self-referential sentence has not been coherently defined. His main argument, the reader will perhaps now be unsurprised to hear, runs roughly as follows: nothing in a sentence (or alternatively, a rule) 'itself' ensures its application (or sense); it is only because sentences (and rules) are part of human practices that they are meaningful.

Note, however, that this does not mean that it is thus human 'agreement' (in the ordinary sense of that word) that determines the sense of a sentence (no 'voting' is involved here; see Wittgenstein (*PI*, §§240–2). Normally, there simply is no wedge between the sentence and its meaning (the rule and its application)—any such wedge can only be inserted afterwards (by the analyst or theorist). As Sharrock and Button (1999) in their discussion of rule-following argue: 'understanding the rule' and 'understanding what to do' are the same thing, i.e., learning to follow a rule *is* learning what to do to act in accord with it. Furthermore, as one of us has argued extensively elsewhere, there need be no implicit metaphysics of rules present here (cf., Guetti and Read, 1996; Read and Guetti, 1999; Read and Sharrock, 2002).

Wittgenstein (*PI* §86) argues that there is no such thing as a 'self-interpreting' item of language, and that seemingly magical or metaphysical connections between, for example, objects and designations, are just inchoate reflections of our grammar. For example, a railway timetable does not literally 'tell' one how to use it, even if it contains lots of horizontal and vertical arrows.[6] Likewise, as already intimated, a sentence alleged to refer to itself does not do so by itself. There is no pointing to oneself, *simpliciter,* unless one is an agent (e.g., a human), no matter what arrows (visible or otherwise) the display contains or exhibits.

[6] It is humans that *tell* each other things, while other uses of the term are typically metaphors. Sharrock and Coleman (2000) remind us of the difference between two kinds of 'telling the time': 'it is obvious that a calendar tells the date and a clock tells the time in a different way from the way a person does' (p. 88). We might want to add that in these discussions it is easy to conflate these two senses.

Any sentence-text needs placing in a communicative context *in which the competence of the recipient is part of that context*. The intelligibility of an anonymous or decontextualised text in part depends then on the reader's understanding (or even misunderstanding) of the text.

'Self-reference' as pointing to something else

Sentences that are cited as apparent attempts of unambiguous 'pure' self-reference can at best be taken to be pointing toward some *other* string of words and declaring of it that it is a sentence (or that it makes sense). A moment's reflection will show us that outside of high theory, outside the average English or Philosophy Department (and perhaps some philosophically-misled mathematicians—see below), this is the way putatively 'self-referential' sentences *are* taken: not in isolation, but as referring to some *other* (usually following) proposition.

In common parlance, outside certain very peculiar philosophical/logical contexts (contexts in which the relevant thought-community 'enforces' upon its members the idea that there is such a thing as ['*de re*'] self-reference), the only types of 'self-reference' that may be said to exist in a relatively unproblematic way are part–whole or attribute–thing relationships (as in metonymy, ordinary cases of recursion, references to oneself, and so on). Sentences which are generated or used for philosophical purposes and forced to conform to a self-reflexive mould are, in normal contexts, read in effect *as ending in colons* (or there is a 'but' 'waiting to happen', as in: 'I know I'm a cad and a liar… *but* you must believe me this time when I tell you this: …').

To respecify the question: *Under what ordinary circumstances* would/does it make sense to say something like 'This sentence makes sense'…?

Compare (imagining, perhaps, an explanation of certain technical aspects of grammar to a non-native speaker):

<u>Case 1</u>

A: 'This set of words is an English sentence: "English is not a romance language.".'

<u>Case 2</u>

A: 'This set of words is not an English sentence.'
((pause))
B: 'What set of words do you mean?'
A: 'Oh sorry; *this* set: "English not romance language is a.".'

In Case 2, A's first remark certainly need not be heard as 'self-referring', just because it took a while before A made clear what she was doing with the remark.[7]

Wittgenstein supposed that there are two things that one can equally well do with a sentence such as 'I am lying.' (or 'This sentence makes sense.'). Either one simply excludes it from the language-game as ill-formed (through making, if you like, a useful and utterly reasonable *ad hoc* alteration in the ordinary 'grammatical' rules). Or, if not, then one can stress that there is, if thought about, nothing devastatingly problematic about such sentences (but also nothing special about what they can properly do for one, which is: nothing):

> Is there harm in the contradiction that arises when someone says: 'I am lying.—So I am not lying.—So I am lying.—etc'? I mean: does it make our language less usable if in this case, according to the ordinary rules, a proposition yields its contradictory, and vice versa?—the proposition itself is unusable, and these inferences equally; but why should they not be made?—It is a profitless performance!—It's a language-game with some similarity to the game of thumb-catching. (*RFM*, I, App. III, §12, p.120)

> Such a contradiction is of interest only because it has tormented people, and because this shews both how tormenting problems can grow out of language, and what kind of things can torment us. (*RFM*, I, App. III, §13, p.120)

> It might also be said: his 'I always lie' was not really an *assertion*. It was rather an exclamation. (*RFM*, IV, §58, p.255)

We might add to Wittgenstein's remarks that we sometimes say, for example, 'I am lying. I didn't *really* just get off the bus…' (at the end of a long cock-and-bull story). However, in that case 'I am lying' would refer to the preceding remarks (or

---

[7] We can go further. 'This set of words is not an English sentence,' might according to our analysis now seem to be necessarily 'paradoxical' or misfiring, if presumed to be 'a reference to itself / to the very same set of words'. However, it's possible to imagine a use for this set of words: take the set of words as part of a quiz for instructing second-language students in the use of negation in English. The demonstrative 'This' would be referring to the same set of words, but as part of the language game of a grammar test. The 'paradox' might be a source of amusement, but as part of the quiz the sentence would make sense:

WHICH OF THESE IS A GRAMMATICAL SENTENCE IN ENGLISH?
A: 'This set of words not an English sentence is.'
B: 'This set of words is an English sentence not.'
C: 'This set of words is not an English sentence.'

What this example makes clear is what we have been stressing throughout: the paramountcy of context and the inclusion of the human communicative setting, on *both* 'sides' of that context.

possibly, in some troublesome cases, to the succeeding remarks as well). In other words 'I am lying.' would still not refer to itself.

In sum, we have argued that we have yet to encounter anywhere or anywhen a convincing example of something that we would definitely want, all things considered, to term 'a sentence referring *by* itself *to* itself'. As we shall see, this is of some considerable import when it comes to the philosophy of mathematics, where for certain alleged philosophically-inflected results, sentences that definitely—in and of themselves—refer to themselves are needed.

We are of course not denying that there is such a thing as, for example, the conclusion of a paper referring back to earlier parts of the same paper. There is nothing in principle to stop one from having a practice of taking some pieces of language (or better: some linguistic actions) as being 'self-reflexive'. Indeed, in some instances it seems obvious that this may be the best way to describe things (e.g., the looking up of 'dictionary' in a dictionary, or the occurrence of the word 'orthography' in the discipline of orthography). For we humans can choose, *ceteris paribus*, to make language work in all sorts of different and novel ways for us. We can choose, even, to take a sentence such as 'This sentence has five words.' as 'referring to itself', if it might serve certain purposes to do so (e.g., teaching number words to a child). But that has not yet indicated a sense for a sentence such as 'I am a liar.' insofar as it is supposed 'to refer to itself'. For without even a potential use, we do not yet have meaning or sense. That is to say, it is very hard to see what one could do with a sentence such as 'I am a liar.' *simpliciter*—apart from simply trying to confuse or amuse someone.

In an important sense, then, it is only people who refer, not sentences. This might sound like the U.S. National Rifle Association's infamous and (to us) troubling slogan, that 'Guns don't kill people; people kill people.'—but the NRA's slogan would actually be quite reasonable, relatively unobjectionable,[8] if it were amended to read: 'Guns don't kill people, people kill people ... often/usually/too-easily with guns.' (However, in that case one doubts whether it would have quite the rhetorical and political impact that the NRA hopes for.) The correct analogy then is this: 'Sentences don't refer; people refer … usually with sentences.'[9]

Once one has thought about it for a moment, this should hardly be controversial at all. Sentences do not refer in and of themselves, any more than guns kill *in and of themselves* (thus far, at least, the NRA are right!). People refer by means of using sentences. Some of those sentences already have a clear reference, given a certain human practice or 'language-game'. Some do not; and purportedly 'self-referential' sentences are among their number. Far from having an absolutely obvious interpretation, most 'self-referential sentences' are always liable in practice

---

[8] We say 'relatively', because there are still reasons for worrying about the flat-footedness of the NRA slogan, reasons explored philosophically here (Selinger 2002): https://www.theatlantic.com/technology/archive/2012/07/the-philosophy-of-the-technology-of-the-gun/260220/

[9] Cf. the quotation from Jayyusi (1993) in Footnote 5.

to evoke bizarre reactions, until a community has instituted such practices, i.e., has institutionalised a particular reading of the kind of strings in question.

MATHEMATICAL SELF-REFERENCE

Having dealt with some ‘ordinary’ instances of self-reference, let us now turn to perhaps the most famous example of self-reference, namely Gödel’s Incompleteness Theorem (GIT) and his ‘formally undecidable’ sentence *P* that is supposed to ‘say’ ‘I am not provable.’

The foundations of mathematics: mathematics or philosophy?

Before tackling GIT directly, some remarks about the context in which it was developed are necessary. Gödel’s theorem was a contribution to discussions about the ‘foundations of mathematics’, in particular in the form developed by Hilbert, which is often called metamathematics or proof theory.

Hilbert’s idea was to formalise classical mathematics (e.g., set theory or number theory) as a *formal theory*, i.e., to restate set theory or number theory in a logically more perspicuous way and to make explicit the notion of *formal proof* by writing down all admissible *inference rules* (how to arrive at new propositions out of previously proven ones). The goal was to reduce all of mathematics to a minimal number of *axioms* (the formalisation of the classical theory) and additional inference rules in order to see more clearly and be able to demonstrate that all inferences made within the theory are valid and true (and therefore do not lead to paradoxes). Gödel in 1933 summarised Hilbert’s project thus:

> a perfectly precise language has been invented by which it is possible to express any mathematical proposition by a formula. Some of these formulas are taken as axioms, and then certain rules of inference are laid down which allow one to pass from the axioms to new formulas and thus to deduce more and more propositions, the outstanding features of the rules of inference being that they are purely formal, i.e., refer only to the outward structure of the formulas, not to their meaning, so that they can be applied by someone who knew nothing about mathematics or by a machine. (This has the consequence that there can never be any doubt as to what cases the rules of inference apply to, and thus the highest possible degree of exactness is obtained.) (Gödel, 1995 [1933]: 45)

This quote nicely exhibits that at least part of the motivation of metamathematics was philosophical in nature, namely to provide mathematics with secure foundations (of a roughly Cartesian kind). In fact, Shanker argues that for Hilbert:

> the whole point of 'meta-mathematics' lay in its deliberate *transgression* of the boundary between mathematics and philosophy; viz. in Hilbert's conviction that he could use this tool to solve mathematically what were *au fond* philosophical problems. (Shanker 1987: 224)

Note that within Hilbert's metamathematics (to which Gödel's theorem was a significant contribution) we have two notions of proof: firstly, *formal proofs* within the formal system itself, and, second, *classical proofs,* the proofs of the mathematical theory to be formalised—and all proofs *about* the formal system itself. Gödel's own proof in his famous 1931 paper was of the latter kind: it was a classical proof 'about' formal systems—but not itself a formal proof.[10]

Moving on to Wittgenstein's remarks about metamathematics, we think that his position could be summarised by saying that Wittgenstein did not question the sense of theorems and proofs (e.g., by Hilbert or Gödel) with respect to the former (the study of formal proofs), but wondered whether they had shed any light on the latter (the study of classical proofs). In other words, Wittgenstein had no quarrels with the study of a new mathematical structure called 'formal proofs' (which, in a sense, is not in principle different from, say, algebra, number theory, or analysis). However, Wittgenstein did wonder whether the work of Gödel or Hilbert had really helped at all to clarify any philosophical issues in the neighbourhood, such as the question of what a mathematical proof 'is'—a question that has troubled philosophers since the time of the Ancient Greeks.[11]

Furthermore, the same distinction, that between 'formal' and 'classical' *proofs*, must be made with respect to mathematical *truth*. Tarski developed a notion of mathematical truth (based on model-theoretic arguments) that is nowadays typically accepted by mathematicians, and by a large number of philosophers (e.g., Bays, 2004), working in proof theory or model theory. However, the same basic question applies: has Tarski's definition of formal truth shed light on the philosophical question of mathematical truth that has troubled philosophers since (at least) Plato?[12]

---

[10] We are not denying that Gödel's proof could be formalised (see, e.g., Shankar 1994), but want to emphasize that proofs *about* formal proofs are not themselves formal proofs.

[11] Eric Livingston (1986) pursues his sociological project in *The Ethnomethodological Foundations of Mathematics* from a similar starting point.

[12] Floyd and Putnam (2006) note that the *technical-mathematical* definition of truth, as truth-in-the-model N (based on Tarski) is often substituted for the *philosophical* question of truth. However, as Floyd and Putnam also point out, whether mathematical model theory is able to solve or in any way contribute toward solving the philosophical question of truth is precisely the issue.

Gödel's incompleteness theorem

We now come directly to a discussion of Gödel's famous (First) Incompleteness Theorem (GIT) and Gödel's construction of a formally undecidable sentence *P*. We agree with Floyd and Putnam (2000) that Wittgenstein at no point questioned the metamathematical (in Hilbert's sense) contribution of GIT. In other words, Wittgenstein accepted GIT as a genuine contribution to 'proof theory' (metamathematics). However, Wittgenstein objected to the view that the theorem had an impact on questions about mathematical truth and the nature of proofs 'in general'—including Gödel's own proof of the GIT.[13] In other words, Wittgenstein wondered whether GIT is only a contribution to the mathematics of formal proofs or also a contribution to the philosophy of mathematics, i.e., whether it is only a contribution to the relatively recent branch of mathematics called 'proof theory' or whether it is a contribution to philosophical questions regarding mathematical truth.

Note that the view that Gödel's theorem has implications for philosophy has been (partly) due to Gödel himself. Gödel starts his paper with a 'metaphorical' summary—a summary that is not necessary for the subsequent technical discussion, but a summary that might have led many philosophers astray. As Floyd puts it:

> 'There are true but unprovable propositions in mathematics' is misleading prose for the philosopher, according to Wittgenstein. It fools people into thinking that they understand Gödel's theorem simply in virtue of their grasp of the notions of *mathematical proof* and *mathematical truth*. And it fools them into thinking that Gödel's theorem supports or requires a particular metaphysical view. (Floyd 2001: 299)

In other words, Gödel's summary does not specify what kind of truth and what nature of proof he is talking about ('classical' or 'formal'). As a consequence, some philosophers have taken this summary to be the only 'mistake' in Gödel's paper.[14]

[13] Livingston (1986: 31) seems to be arguing along similar lines: 'The argument that will be made is *not* that a proof of Gödel's theorem does not prove what others have claimed it to prove; instead the origins of the rigor of a proof of Gödel's theorem will itself be examined and the claim advanced that that rigor consists of its local work. Thus this argument points to the primordial character of the activity of doing mathematics over some conception of mathematics-in-itself.' We take Livingston to be asking: Where does the 'truth' of Gödel's proof of the GIT, a proof written in a 'classical' not a 'formal' style, come from?

[14] Helmer (1937: 59) writes: 'As a matter of fact, Gödel did make one "mistake", namely that of writing an introduction to his paper, in which he sketches "the main idea of the proof, without of course making any claims for precision". It is the actual lack of precision in these introductory explanations that has misled Perelman and that may mislead others. It can easily be seen that Perelman's objections are applicable only to these inexact explanatory remarks, and not to the exact formal demonstration given later in Gödel's paper.'

Let us quickly comment on the purely technical content of the Gödel proof. In Gödel's proof it is necessary to talk about all the ('infinitely many') sentences in the formal system. In a way, Gödel extends Cantor's diagonalization procedure (to show that the real numbers have a greater cardinality than the natural numbers) into a procedure for reasoning about formal proofs (hence Gödel's proof idea is sometimes called a 'double-diagonalization procedure').

As Watson (1938), in an article that (according to Floyd and Putnam) was close to Wittgenstein's position, reminds us: all of the problems or questions that arose in the so-called foundation crisis of mathematics were connected to questions of mathematical infinity, especially when treating a process as a totality (see also Shanker, 1987, chapter 5). Again, we have no quarrels with mathematical developments in this area, e.g., extending the concept of number from the finite to the infinite. However, one must bear in mind that we are in the realm of mathematical infinity—and should expect that results may be 'surprising' to the non-mathematical reader.[15]

'I am not provable'

To progress now the arguments advanced in this paper towards their conclusion, let us take a closer look at the 'formally undecidable sentence *P*'. Again, the question is not whether the sentence does the mathematical work assigned to it within Gödel's proof (it does), but what one can say about the sentence—or about what it can itself 'say'.

A first and critical question might be whether 'I am not provable' is the only translation or rendering of the sentence *P*. That is to say, building on our argument in the section on 'Metaphysical lines of projection', above, the question is not whether it can be translated this way (it can) but whether this is the only possible translation.[16] In other words: What ensures that this is the only possible translation? Something in the sentence itself? Are we compelled to read the Gödel sentence as referring to itself?[17]

---

[15] Non-mathematical readers should note that mathematicians have developed a technical definition of infinity. In other words, 'mathematical infinity' is not the same 'thing' as our common-sense notion of infinity (if we have one). So in a sense it is hardly surprising that mathematicians say astonishing things about infinity. (See on this our forthcoming paper on infinity in *Reasons and Empty Persons* (Springer).)

[16] Floyd and Putnam (2000: 628) remark that 'the 'translation' of the famous Gödel sentence P as 'P is unprovable in PM' is not cast in stone'. Rodych (1999: 182) also notes: 'We do not *need* to assume a natural language meaning for '*P*' (e.g., an *English* meaning) to obtain the threatened contradiction, for it is just a number-theoretic 'fact' that an actual proof of '*P*' would enable us to *calculate* the relevant Gödel numbers and thereby arrive at '~*P*' by existential generalization'.

[17] Bays (2004: 206-7) seems to acknowledge this: 'there is nothing in the formal structure of *P*—that is, in *P's* very syntax—which forces us to interpret *P* as '*P* is not provable." However, he goes on to say that '[t]here is a perfectly good—and a perfectly *mathematically* respectable—interpretation of

Secondly, again in line with our earlier arguments, we wonder whether the sentence *itself* can 'say' anything. Or, put less strongly, in which sense does the sentence *P* 'say' 'I am not provable'? Does the sentence itself 'say' this? How? As Wittgenstein remarks:

> Do not forget that the proposition asserting of itself that it is unprovable is to be conceived as a *mathematical* assertion—for that is not *a matter of course*. It is not a matter of course that the proposition that such-and-such a structure cannot be constructed is to be conceived as a mathematical proposition.
>
> That is to say: when we said: 'it asserts of itself'—this has to be understood in a special way. For here it is easy for confusion to occur through the variegated use of the expression 'this proposition asserts something of ...'.
>
> In this sense the proposition '625 = 25 × 25' also asserts something about itself: namely that the left-hand number is got by the multiplication of the numbers on the right.
>
> Gödel's proposition, which asserts something about itself, does not *mention* itself. (RFM, VII §21, pp. 385–386)

Wittgenstein here is drawing our attention to the different forms that 'says of itself' can take, i.e., that 'says of itself' means something different in different contexts of use and that for *this* example the context of use is the practice of mathematics.

In our view, Wittgenstein is also questioning *who* or *what* is doing the 'mentioning' in 'mentioning itself'. *P*, in a sense, is only a string of symbols (on an imagined piece of paper). *P* of itself obviously does not 'know' anything about the natural numbers or about anything at all. Neither can *P* 'do' or (in the normal sense of the word) 'say' anything.

We would argue that paradoxes of self-reference occur only if one looks at such sentences in isolation, i.e., without *specifying* the specific context or the practice in which the sentences occur. It is only as part of a practice that people *take* a sentence to be an example of 'self-reference'. With respect to Gödel's sentence *P*: it can do the required mathematical work (within GIT) without us having to take it as a case of self-reference.[18] It is only for the *philosophical* interpretation of GIT that the sentence has to be treated as an instance of 'self-reference'.

As an ironic summary we might even say: only human decisions produce undecidability (as part of certain practices). If one decides that a certain formula may

---

the language of arithmetic under which *P* expresses the fact that *P* is not provable' (p. 208). For Bays, this is 'the canonical interpretation' (p. 204). We do not question the status of this as a canonical interpretation *within mathematics* (in particular, proof theory), but wonder whether it has—or should have—this status *within philosophy* (i.e., for questions of the kind that troubled Wittgenstein).

[18] See the remark by Rodych (1999) in Footnote 16.

be taken to 'say' that it is unprovable, then one must take responsibility for that decision. The Gödel sentence, in isolation, simply cannot be said simply and definitely to refer to itself by itself (though it must to do the philosophical job it has had assigned it).

Indirect self-reference

The Gödel sentence does of course not simply self-refer. The trick in Gödel's proof lies in the back-and-forth movement between the formal system and the natural numbers. Thus there is apparently no self-reference in and of itself here, but only an indirect one, using some particularly clever working within the previously-defined rules of syntax and interpretation of certain logico-mathematical systems.[19] Gödel mentioned this fact in a very striking way in the course of his own published proof:

> We therefore have before us a proposition that says about itself that it is not provable [in *PM*].*
>
> *Contrary to appearances, such a proposition involves no faulty circularity, for initially it [only] asserts that a certain well-defined formula (namely, the one obtained from the *q*th formula in the lexicographic order by a certain substitution) is unprovable. Only subsequently (and, so to speak, by chance) does it turn out that this formula is precisely the one by which the proposition itself was expressed. (Gödel, 1988 [1931]: 19 and 43)

The key point in all the technical tricks in the Gödelian arsenal comes down to this: let us allow that the arrow from 'This' back to the sentence 'itself' (in the figures drawn earlier in this paper) passes through several intermediate stages and follows a long and devious trajectory before it returns to its target. However, why should this change anything? Why should it make any difference to the logic of the situation if the arrows are long or short? We could easily adapt the diagrams earlier to suit; the very same ambiguities will still arise. In short: indirectness or indirection buys you nothing.

Syntax alone does not generate referentiality (as syntax cannot 'do' *anything*), while semantics gives an infinite regress (as we move from mathematics to meta-mathematics, from meta-mathematics to meta-meta-mathematics, and so on—

[19] Parsons and Kohl (1960: 70), also note that Gödel's sentence, in contrast to the liar paradox, constitutes only indirect self-reference: 'it is shown, by a correlation of formulae with numbers, that the statement of arithmetic which Gödel's formula may be interpreted to express is true if and only if the formula is itself not provable in the system S. It is only in this derivative sense that the formula asserts anything about itself. It is clear that this weak sense of self-reference is quite different from the first two senses and does not, of itself, give rise to paradoxes.'

compare again our three figures earlier). This point is immanent in the Wittgensteinian sense of 'grammar', according to which a syntax not conterminous and simultaneous with a semantics is closer to being a nothing than even to being a something waiting for an 'interpretation' to be imposed on it.

Again, we are seeking to point out that there is nothing in the sentence 'itself' that 'ensures' self-reference. Note that Gödel in the above quotations says 'Only subsequently [...] does it turn out'. Gödel thereby implicitly refers to a human for whom it supposedly 'turns out' thus (as sentences cannot 'notice' or 'see' anything).

In sum: the Gödel sentence *of itself* does not lead to any paradox.

And now one can start to see why even our title, 'Can sentences self-refer?', is misleading in an interesting way. One is inclined to think that our answer to the question is 'No'; but, actually, we wonder whether answering 'Yes' *or* 'No' to this question actually amounts to anything that, all things considered, one really would both want to mean and succeed in meaning. Sentences by themselves cannot do anything at all, we might say; *a fortiori* they cannot self-refer.[20]

CONCLUSION

With respect to the debate between Floyd-Putnam and Bays, what is at stake is the question of whether the *mathematical* contribution to proof theory has settled or even impacted upon the *philosophical* questions about mathematical truth and provability that rightly concerned Wittgenstein. We believe that our arguments provide independent support for the correct conclusion of Floyd-Putnam on Gödel, that it is not a mathematical result, but rather a metaphysical claim, to say that if PM is consistent then some mathematical truths are undecidable in PM. Wittgenstein, Floyd and Putnam (2000: 632) write, does not want 'simply to *deny* the metaphysical claim; rather, he wants us to see how little sense we have succeeded in giving it.'[21]

We believe that the independent support that we have provided for their conclusion comes from the more *general* good reason we have given for believing that the mysteries, paradoxes, and logical results that are thought to follow from the consideration of isolated 'self-referential language' or 'self-reflexive phenomena' are in nearly all cases quite illusory. Instances of self-reference have to be considered as part of specific human practices. Wittgenstein remarks:

---

[20] But these 'cannots' must in turn be worked through as indications of a nonsensicality latent *in one's desires* with regard to one's words. And this is just as true of the desires motivating the philosophical interpretation of Gödel's sentence *P*.

[21] This aspect of our line of thought on Gödel, distinguishing rigorously, as do Putnam and Floyd, between Gödel's mathematical innovations (with which we have no quarrel) and the philosophical consequences alleged to flow from diagonalization, is also in some regards close to Shanker (1987, 1988), as well as to Sayward (2001).

> In the language of a tribe there might be a pronoun, such as we do not possess and for which we have no practical use, which 'refers' to the propositional sign in which it occurs. I will write it like this: Ĭ. The proposition 'Ĭ am ten centimetres long' will then be tested for truth by measuring the written sign. The proposition 'Ĭ contain four words' for example is true, and so is 'Ĭ do not contain four words'. 'Ĭ am false' corresponds to the paradox of the Cretan Liar. –The question is: What do people use this pronoun for? Well, the proposition 'Ĭ am ten centimetres long' might serve as a ruler, the proposition 'Ĭ am beautifully written' as a paradigm of beautiful script.
>
> What interests us is: How does the word 'Ĭ' get used in a *language-game*? For the proposition is a paradox only when we abstract from its use. Thus I might imagine that the proposition 'Ĭ am false' was used in the kindergarten. When the children read it, they begin to infer 'If that's false, it's true, so it is false, etc. etc.' People have perhaps discovered that this inferring is a useful exercise for children.
>
> What interests us is: how this pronoun gets used in a *language-game*. It is possible, though not quite easy, to fill out a picture of a language-game with this word. A proposition like 'Ĭ contain four words' might, for example, be used as a paradigm for the number four, and in another sense so might the proposition 'Ĭ do not contain four words'. *A proposition is a paradox only if we abstract from its use*. (Wittgenstein 1980: §65, our emphasis)

We must re-iterate Wittgenstein's conclusion: 'A proposition is a paradox only if we abstract from its use.' In other words, if Gödel's sentence *P* is taken as an instance of 'self-reference' as part of the 'language game' of formal proofs, then there simply is no paradox (as Gödel has given a technical definition of what 'self-reference' means in this context). However, this means that philosophical consequences are also restricted to this particular context. Consider how Bays (2004, p.210) concludes his critical response to Floyd-Putnam:

> There is a perfectly good—and indeed, a perfectly canonical—interpretation of arithmetic under which Wittgenstein's *P* really does say '*P* is not provable'. Given this interpretation, Gödel's Theorem helps to show that there are 'true but unprovable' sentences of ordinary number theory. Nothing in Wittgenstein's remarks—or in Floyd and Putnam's analysis of those remarks—should lead us to think otherwise.

In saying this, Bays overlooks the importance of his own qualification: 'Given this interpretation ...'. Given that we are operating within proof theory (which means, *inter alia*, accepting certain technical definitions of syntax and semantics) then we could say that there are 'true but unprovable' sentences. What Wittgenstein (and Floyd and Putnam) are questioning is whether we have therefore explained 'truth' and 'proof' within mathematics *in general*.

To think that as a philosopher one *has* to accept that *P* means '*P* is not provable' is to think that one practice (that of formal proofs) can stand as an explication of all other practices. Doing this constitutes an avoidance of responsibility and is exactly what the best Wittgensteinian philosophers, following Stanley Cavell, have been warning against for some years now. For sure, sentences can self-refer—if that is what one wants them to do. But that surety will always be part of a specific, and not compulsory, human *practice*.

REFERENCES

Bays, T. 2004. 'On Floyd and Putnam on Wittgenstein on Gödel'. *The Journal of Philosophy* 101 (4): 197–210.

Cavell, S. 1969. *Must We Mean What We Say? A Book of Essays*. New York: Charles Scribner and Sons.

Cook, J. W. 1978. 'Whorf's Linguistic Relativism: Part 2'. *Philosophical Investigations* 1 (2), 1–37.

Floyd, J. 2001. 'Prose Versus Proof: Wittgenstein on Gödel, Tarski and Truth'. *Philosophia Mathematica* 9 (3): 280–307.

Floyd, J., and H. Putnam. 2000. 'A Note on Wittgenstein's "Notorious Paragraph" About the Gödel Theorem'. *The Journal of Philosophy* 97 (11): 624–32.

Floyd, J., and H. Putnam. 2006. 'Bays, Steiner, and Wittgenstein's "Notorious" Paragraph About the Gödel Theorem'. *The Journal of Philosophy* 103 (2): 101–10.

Foucault, M. 1983. *This Is Not a Pipe*. Berkeley, CA: University of California Press.

Garfinkel, H. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Garfinkel. H., ed. 1986. *Ethnomethodological Studies of Work*. London: Routledge & Kegan Paul.

Garfinkel, H. 2002. *Ethnomethodology's Program: Working Out Durkheim's Aphorism*. Lanham, MD: Rowman and Littlefield.

Gödel, K. 1988 [1931]. 'On Formally Undecidable Propositions of "Principia Mathematica" and Related Systems I'. In *Gödel's Theorem in Focus*, edited by S. G. Shanker, 17–47. London: Croom Helm.

Gödel, K. 1995 [1933]. 'The Present Situation in the Foundations of Mathematics'. In *Collected Works*, vol. 3: *Unpublished Essays and Lectures*, 45–53. Oxford: Oxford University Press.

Greiffenhagen, C., and W. Sharrock. 2006a. 'Logical Relativism: Logic, Grammar, and Arithmetic in Cultural Comparison'. *Configurations* 14 (3): 275–301.

Greiffenhagen, C., and W. Sharrock. 2006b. 'Mathematical Relativism: Logic, Grammar, and Arithmetic in Cultural Comparison'. *Journal for the Theory of Social Behaviour* 36 (2): 97–117.

Greiffenhagen, C., and W. Sharrock 2007. 'Linguistic Relativism: Logic, Grammar, and Arithmetic in Cultural Comparison'. *Language and Communication* 27 (1): 81–107.

Greiffenhagen, C., and W. Sharrock. 2011. 'Does Mathematics Look Certain in the Front, but Fallible in the Back?' *Social Studies of Science* 41 (6): 839–66.

Greiffenhagen, C., and W. Sharrock. Forthcoming. 'Tensions in Garfinkel's Ethnomethodological Studies of Work Programme Discussed Through Livingston's Studies of Mathematics'.

Guetti, J., and R. Read. 1996. 'Acting from Rules: Internal Relations Versus Logical Existentialism'. *International Studies in Philosophy* 28 (2): 43–62.

Helmer, O. 1937. 'Perelman Versus Gödel'. *Mind* 46 (181): 58–60.

Hutchinson, P., R. Read, and W. Sharrock. 2008. *There is No Such Thing as a Social Science: In Defence of Peter Winch*. London: Routledge.

Jayyusi, L. 1993. 'The Reflexive Nexus: Photo-Practice and Natural History'. *Continuum: Journal of Media and Cultural Studies* 6 (2): 25–52.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Livingston, E. 1986. *The Ethnomethodological Foundations of Mathematics*. London: Routledge & Kegan Paul.

Lynch, M. 1985. *Art and Artifact in the Laboratory: A Study of Shop Work and Shop Talk in a Research Laboratory*. London: Routledge & Kegan Paul.

McManus, D. 2006. *The Enchantment of Words: Wittgenstein's 'Tractatus Logico-Philosophicus'*. Oxford: Oxford University Press.

Parsons, C., and H. R. Kohl. 1960. 'Self-Reference, Truth, and Provability'. *Mind* 69 (273): 69–73.

Read, R., and J. Guetti. 1999. 'Meaningful Consequences'. *The Philosophical Forum* 30 (4): 289–315.

Read, R., and W. Sharrock. 2002. 'Kripke's Conjuring Trick'. *Journal of Thought* 37 (3), 65–96.

Rodych, V. 1999. 'Wittgenstein's Inversion of Gödel's Theorem'. *Erkenntnis* 51 (2–3), 173–206.

Sayward, C. 2001. 'On Some Much Maligned Remarks of Wittgenstein on Gödel'. *Philosophical Investigations* 24 (3), 262–70.

Selinger, E. 2002. 'The Philosophy of the Technology of the Gun'. *The Atlantic* (online). Available at: https://www.theatlantic.com/technology/archive/2012/07/the-philosophy-of-the-technology-of-the-gun/260220/ (accessed 31 July 2019).

Shankar, N. 1994. *Metamathematics, Machines, and Gödel's Proof*. Cambridge: Cambridge University Press.

Shanker, S. G. 1987. *Wittgenstein and the Turning-Point in the Philosophy of Mathematics*. London: Croom Helm.

Shanker, S. G. 1988. 'Wittgenstein's Remarks on the Significance of Gödel's Theorem'. In *Gödel's Theorem in Focus*, edited by S. G. Shanker. 155–256. London: Croom Helm.

Sharrock, W., and G. Button. 1999. 'Do the Right Thing! Rule Finitism, Rule Scepticism and Rule Following'. *Human Studies* 22 (2–4): 193–210.

Sharrock, W., and W. Coleman. 2000. 'It Don't Mean a Thing: On What Computers Have to Say'. *Communication and Cognition* 33 (1–2): 83–95.

Sharrock, W., and R. Read. 2002. *Kuhn: Philosopher of Scientific Revolutions*. Cambridge: Polity.

Watson, A. G. D. 1938. 'Mathematics and Its Foundations'. *Mind* 47 (188): 440–51.

Winch, P. 1958. *The Idea of a Social Science and Its Relation to Philosophy*. London: Routledge & Kegan Paul.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

Wittgenstein, L. 1978. *Remarks on the Foundations of Mathematics*. 3rd ed. Oxford: Basil Blackwell.

Wittgenstein, L. 1980. *Remarks on the Philosophy of Psychology*, vol. 1. Oxford: Basil Blackwell.