Perlis on strong and weak self-reference – a mirror reversal

Damjan Bojadžiev

Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana

Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

E-mail: damjan.bojadziev@ijs.si

Fax: (+386 61) 1258-058, 219-385

Phone: (+386 61) 1773-768, 1773-644

URL: http://nl.ijs.si/~damjan/me.html

Abstract: The kind of self-reference which Perlis characterizes as strong (Perlis, 1997), as opposed to formal self-reference which he characterizes as weak, is actually already present in standard forms of formal self-reference. Even if formal self-reference is weak because it is delegated, there is no specific delegation of reference for self-referential sentences, and their 'self' part is strong enough. In particular, the structure of self-reference in Gödel's sentence, with its application of a self-referential process to itself, provides a model of Perlis' characterization of a self. This structure can also be interpreted visually, in a way relevant to self and consciousness, namely as self-recognition in a mirror.

1. Introduction

The view that consciousness has something to do with self-reference appears in various forms. It may connect self-reference only with self-consciousness, in the sense of the apparent truism

consciousness + self-reference = self-consciousness

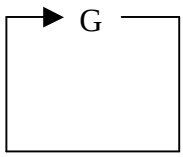
In this vein, Jackendoff says that self-consciousness 'involves (at least) a combination of ordinary consciousness with self-reference' (Jackendoff, 1987, p. 18). The opposite view is that self-consciousness is already involved in (ordinary) consciousness itself: consciousness entails (Gennaro, 1996) or 'requires a large degree of' self-consciousness (Hofstadter, 1980, p. 328). Similarly, in a recent paper in this journal, Perlis suggests that '[self-consciousness] is the most basic form of all [consciousness]' (Perlis, 1997, p. 516). On this kind of view, self-reference is not merely a necessary condition of (self-)consciousness but something much more central to it, closer to a sufficient condition; as Perlis says, 'consciousness is … first and foremost, a special kind of self-reference' (p. 514). A similar orientation is evident in the autopoietic literature: if self-reference is the 'characteristic property of autopoiesis' (Morin, 1981, p. 130), then 'setting consciousness roughly equal with autopoiesis' (Locker, 1981, p. 225) roughly means that self-reference is the characteristic property of consciousness.

However, Perlis' special kind of self-reference, sufficient to, as Locker says, 'found the subject' (Locker, 1981, p. 226), is supposed to be a new, strong form of self-reference, stronger than the traditional forms 'cited and studied, from antiquity to the present' (Perlis, 1997, p. 518). The next section of the paper

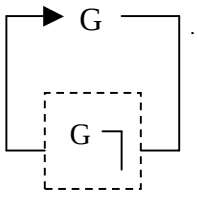
examines this claim and shows that the properties of what Perlis calls strong self-reference are already at work in its traditional forms, which he classifies as weak. In particular, the structure of self-reference in Gödel's sentence provides a model of Perlis' characterization of a self. Interpreting this structure visually then provides a formal model of self-recognition in a mirror, a biologically rare ability thought to indicate self-awareness (sect. 3). The paper concludes by tracing briefly the theme of strong self-reference in philosophical conceptions of consciousness (sect. 4).

2. Strong and weak self-reference

Perlis argues that 'consciousness is synonymous with self' (p. 509) and suggests that 'a self is best thought of as an entity G that can refer to G as that entity doing that very referring' (p. 519). The first part of this definition, the reference of G to G, could be pictured like this:

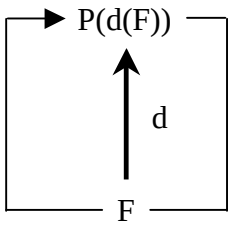


The qualification that G refers to G *as* that entity doing that very referring could then be added like this:

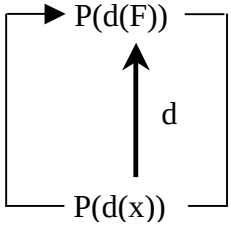


The dashes frame a repetition of the relevant part of the diagram, namely the entity (G) doing the referring (developing the line of reference). The dashed box thus

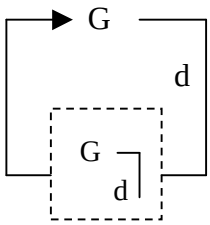
pictures the way in which a self refers to itself: by latching onto the very structure of self-referring. This referring must find its way back to its origin through its own emergence from it, which means that it must somehow involve itself in the way it proceeds to its target/origin. For G to refer to G as the entity doing that very referring, its referring must "turn back on itself " in order to direct itself (back) to its origin. The self-reference of G is thus twofold: G refers to itself, and so does its referring (refer to *it*self, become (part of) its own object). But this double circular structure, with its peculiar relationship between process and object, is also characteristic of formal self-reference. The classic example is Gödel's sentence, which says of itself that it is not provable (in the formal system in which it is formulated). But for the purposes of this paper it is not important what the sentence says about itself but only *how* it says it. Marking what a sentence says about itself by a one-place predicate P, the first, outer layer of its self-reference could be pictured like this*:*



The process d produces the self-referential sentence P(d(F)) from some formula F. The sentence refers to itself through the term d(F), which refers to what d produces from F, which is the sentence itself. This basic loop of self-reference rides on a second, inner layer, which depends on the nature of the process d. In the simplest case, d is reflected in F as well, so that the generic F is P(d(x)):



In this case, d is a process which replaces the free variable in a formula by the name of that very formula. This process, called diagonalization[1], which already has a self-referential character, is further reflected in the formula to which it is applied, in the term d(x), which is a generic description of its results. This inclusion of a process of reference in its object exemplifies the way reference must "turn back on itself" that is necessary for referring to the entity that does that very referring. The application of a self-referential process to itself corresponds to what Perlis says a strongly self-referential entity must do: 'refer to that very referring' (p. 519). The lines of reference in the diagram of the self could thus be labeled with the symbol of diagonalization:



Self-reference through diagonalization thus provides a formal model of Perlis' characterization of a self (and suggests an explanation of his symbol for it).

The reason that Perlis classifies formal self-reference as weak is that it is delegated: 'the actual action of referring is done by an interpreter outside the supposedly self-referential objects (sentences)' (p. 518, fn. 17). But there is no specific delegation of reference for self-referential sentences, and the 'self' part

[1] The reason for this geometrical designation cannot be explained briefly enough here, but see e.g. (Hofstadter, 1979, p. 446). The details of the construction of Gödel's sentence can be found e.g. in (Boolos and Jeffrey, 1980, pp. 170-3). A similar construction is the basis

of formal self-reference is strong enough. As soon as the formal system is interpreted (to refer, in the standard case, to numbers), some of its sentences turn out to refer to themselves, and some of its theorems turn out to express the corresponding statements of self-reference (in so far as they can be expressed within the formal system itself).

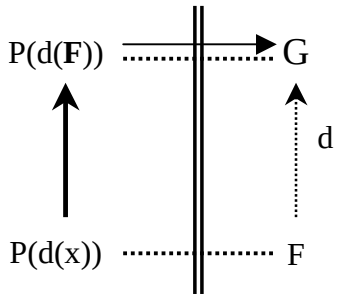
3. Formal and visual self-reference

The parallel between the structures of formal self-reference and Perlis' strongly self-referring self is even closer if reference is interpreted visually, as looking at or seeing. The self then becomes an entity G that can "see itself as that entity doing that very seeing" or, as Perlis says, 'sees itself as a self-seeingness' (p. 523). This way of putting it rightly emphasizes the element of self-recognition, the knowledge that what the self sees/refers to is itself. On the formal side, the interpretation of reference as seeing is even more productive: the self-reference of Gödel's sentence is then comparable to seeing oneself in a mirror. The mirror comes in through a feature of formal self-reference which was left out above for the sake of simplicity: arithmetical self-reference of the kind constructed by Gödel is indirect in that the sentence refers not simply to itself but rather to its own number, the number which belongs to it in some scheme of coding arithmetical expressions as numbers. For, arithmetical expressions as such only refer to numbers, so they must be coded as numbers if they are to be able to refer to themselves (or other expressions). The code thus functions as a numerical mirror which extends the field of reference to what would otherwise

---

for Kleene's theorem (Webb, 1980, p. 214, 234), which generates many kinds of computational reflection that 'takes its own activity into account' (Perlis, 1997, p. 523).

remain outside of it. The details of this interpretation of the code as a mirror are presented e.g. in (Bojadžiev, 1996), but the final diagram of the interpretation is appropriate here in case it all seems too abstract:



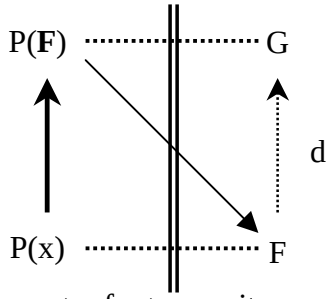
The double vertical line represents the principles of the Gödel code, the numerical mirror in which expressions (left) are reflected (dotted lines) in numbers (right): F is the number of the formula P(d(x)), and G is the number of the sentence P(d(**F**)). Bold type marks the difference between a number (F) and the numeral of that number (**F**): diagonalization now replaces the free variable in a formula (P(d(x))) with the numeral of its number (**F**). This operation on expressions is reflected, through the code, in some numerical function d (dotted arrow on the right) which maps the number of a formula (F) to the number of its diagonalization (G). The term d(x) represents this function, so the term d(**F**) in P(d(**F**)) refers to G, the number of that very sentence. This self-reference is expressible in the equation

(*) d(**F**) = **G**

in which **G** is the numeral of G, and this equation is provable in the system in which P(d(**F**)) is constructed (Boolos and Jeffrey, 1980, p. 173). Reading the self-reference of P(d(**F**)) as "seeing oneself as the entity doing that very seeing" is now straightforward: the sentence refers to/sees its own image, and sees it as its own, belonging to the sentence doing that very referring/seeing. The

provability of (*), which contributes the element of knowledge, narrows the analogy to self-recognition.

The analogy is even closer if diagonalization in its dual role is also interpreted visually. By itself, applied to any formula, it already has an element of looking at the own image, because what is substituted into the formula is a representation of its numerical image:



But the resulting sentence does not refer to, see *its* own image: it only sees, so to speak, the image of its previous, non-referring, unseeing state - as if, on opening my eyes in front of a mirror, I were to see myself with my eyes still closed. But in self-reference, this process of looking at oneself is applied to something containing its representation, so that the look at oneself is part of its object, and part of what is seen by it: literally “seeing oneself as the entity doing that very seeing”, the entity that ‘sees itself as a self-seeingness’ (p. 523). Finally, the representation d(x) of the numerical image of diagonalization relies on the way in which expressions and operations on them are reflected, through the code, in numbers and numerical operations. This suggests a particular, purely formal way of self-recognition, based on noting the parallelism between things and mirror images, e.g. in posture, gesture or movement.

The motivation for this interpretation of formal self-reference is to connect it with other kinds of self-reference thought to be relevant to self and consciousness. Self-recognition in a mirror 'has long been considered a diagnostic indicator for the emergence of a self concept' (Butterworth, 1998, p. 136), and has been proposed as a behavioral, 'objective criterion for testing for awareness of self' (Gregory, 1987, p. 493); cf. (Lacan, 1977, pp. 1-7) on the conception of a mirror stage of development. The development of the ability to recognize the mirror image itself involves the integration of other, more or less self-referential abilities such as proprioception, especially kinesthesia (Sheets-Johnstone, 1998), (self-)perception and cognition (Butterworth, 1998, p. 136). In evolutionary terms, the ability to recognize the mirror image seems to constitute an important cognitive threshold, passed only by a small number of non-human species. The cognitive challenge is to recognize that what appears in the mirror is not another member of the species but an image, and then to recognize that the image is one's own, not by any special mark, which it doesn't have, but by its relationship to oneself.

The mirror recognition test also offers what should be a prime example of what Perlis calls strong self-reference by an action, namely the characteristic action, in Gallup-style experiments (Butterworth, 1998, p. 136), of touching the spot painted on the forehead in recognition of the identity of the mirror image. However, the characterization of strong self-reference by an action (p. 520), which seems to cover any conscious action, does not make it possible to say that reaching for the spot on the forehead is any more self-referential than e.g. reaching for a pencil.

4. Philosophical epilogue

In the philosophical tradition of thinking about the subject, there seems to be a substantial, not merely terminological divide regarding the question whether consciousness can be of itself (Kant's *reine Anschauung*, Sartre's pre-reflective *cogito*) or not (Hume's verdict on introspective attempts to find the self, Hegel's definition of consciousness as that which is opaque to itself, irreflexive Buddhist optics of consciousness). As Toms says, the question is whether consciousness is 'true *self*-consciousness, an act of consciousness knowing itself in its own occurrence' (Toms, 1984, p. 35). One issue here is that it is hard to see how consciousness could "get a grip" on itself, "step behind its own back" and become its own object, if it is always the instrument. As Deikman recently put it, 'awareness cannot be made an object of observation because it is the very means whereby you can observe' (Deikman, 1996, p. 351). Contrary to this, Perlis suggests that awareness can be its own object, 'pure awareness of itself' (p. 523), and gives linguistic examples of strong 'referring that refers to that very referring' (p. 520). This kind of referring again recalls Hegel and his conception of reflexive relationships, reflected in their objects (a relationship to something being at the same time a relationship to that relationship to it). But it was only formal self-reference of the kind exemplified by Gödel's sentence which offered a precise model of such reflexivity, showing *how* the means of observation can become their own object. This aspect of Gödel's work on self-reference, namely the formal construction of a self-referential *sentence*, has received much less attention than the implications of his *theorems* for the puzzles of human and machine reflection. But the sentence itself has perhaps as

much to offer to the studies documented by this journal; as Perlis himself says, though only with reference to the brain, ‘perhaps the diagonal method of Cantor, used so well by him and Gödel and Turing in explicating self-referential mysteries of mathematics and computation, has yet more in store for us’ (p. 524).

Acknowledgments

I am grateful to Joseph Goguen, editor in chief, for his careful reading of previous versions of this paper. His comments prompted me to rethink and improve the presentation, especially the diagrams, and parts of the text as well, by helping me realize what I had actually written, so I could make it coherent. I am also grateful to Don Perlis for supplying a challenging cue, and to Anthony Freeman, managing editor, for handling the communications involved.

References

Baumgartner, Elisabeth *et al* (ed.) (1996), *Phenomenology and Cognitive Science: Handbook* (Dettelbach: Röll Verlag).

Bojadžiev, Damjan (1996), 'Self-reference in Phenomenology and Cognitive Science’, in Baumgartner *et al* (1996), pp. 313-8.

Boolos, George, Jeffrey, Richard (1980), *Computability and Logic* (Cambridge: Cambridge University Press).

Butterworth, George (1998), ‘A Developmental - Ecological Perspective on Strawson’s ‘The Self’’, *Journal of Consciousness Studies*, **5** (2), pp. 132-40.

Deikman, Arthur (1996), ‘‘I’ = Awareness’, *Journal of Consciousness Studies* **3**, (4), pp. 350-6.

Gennaro, J. Rocco (1996), *Consciousness and self-consciousness* (Amsterdam / Philadelphia: John Benjamins Publ. Co.).

Gregory, Richard (ed.) (1987), *The Oxford Companion to the Mind* (Oxford: Oxford University Press).

Hofstadter, Douglas (1980), *Gödel, Escher, Bach: An Eternal Golden Braid* (New York: Random House).

Jackendoff, Ray (1987), *Consciousness and the Computational Mind* (Cambridge: The MIT Press).

Lacan, Jacques (1977), *Écrits - A Selection* (London: Tavistock Publications).

Locker, Alfred (1981), ‘Self-Reference and “Autopoiesis”’, in Zeleny (1981), pp. 211-33.

Morin, Edgar (1981), ‘Self and Autos’, in Zeleny (1981), pp. 128-37.

Perlis, Donald (1997), ‘Consciousness as Self-Function’, *Journal of Consciousness Studies*, **4** (5-6), pp. 509-25.

Sheets-Johnstone, Maxine (1998), ‘Consciousness: A Natural History’, *Journal of Consciousness Studies*, **5 (**3), pp. 260-94.

Toms, Eric (1984), *Relation and Consciousness* (Edinburgh: Scottish Academic Press).

Webb, Judson (1980), *Mechanism, Mentalism and Metamathematics - An Essay on Finitism*, (Dordrecht: D. Reidel Publishing Company).

Zeleny, Milan (ed.) (1981), *Autopoiesis – A Theory of Living Organization* (New York: Elsevier North Holland).