Introduction

Sanford C. Goldberg

1

In the nineteen sixties and seventies a new view began to emerge, first in the philosophy of language, then subsequently in the philosophy of mind, having to do originally with the theory of reference. Prior to this time, and in accordance with a particular (not uncontested) reading of Frege, it had been common to assume that who or what one was referring to, when one used a linguistic expression to refer to someone or something, was determined by *the satisfaction of certain criteria*. The criteria in question were assumed to derive either from the meanings of one's expression, or from the cognitive significance that speakers attached to that expression, or perhaps from the referential intentions speakers have in mind as they purported to refer.[1]

But in the sixties and seventies various authors began to question this *satisfaction-theoretic conception of reference determination*.

There were two main sources of dissatisfaction. One is a dissatisfaction with the implications of the satisfaction-theoretic conception for modal logic. As both Ruth Barcan Marcus and Saul Kripke were to recognize, the view that reference proceeds by satisfaction appears to have unacceptable implications for the semantics of sentences involving proper names (and other referring expressions). The other source of dissatisfaction has to do with the overly intellectualist (and individualistic) assumptions of the satisfaction-theoretic conception. If that conception were correct, using an expression to refer would require a

[1] Admittedly, this is not an entirely happy way to put the point, for two reasons. First, I am ignoring the distinction between speaker reference and semantic reference. Second, I am assuming something that many theorists will resist, namely, that it makes sense to speak of reference in connection with e.g. predicates. Still, in the interest of brevity I persist in this highly informal way of speaking, trusting that the more technical uses of 'refers' as well as the various distinctions that can be drawn will be irrelevant to the contrasts I am hoping to bring out, and that in any case the technicalities will be well known to the readers who are most involved in these debates.

speaker to have identifying knowledge of the referent, and to be disposed to express this knowledge by using the referring expression itself. Many theorists argued that these conditions are not jointly satisfied in some cases; illustrations included cases in which the expressions themselves were common names with standard references (Kripke 1972, Evans 1973), natural kind terms (Putnam 1975), definite descriptions used in a referential way (Donnellan 1966), or demonstratives (Kaplan 1975, 1979; Evans 1975, 1979).

Thus began the so-called "externalist" revolution in the philosophy of language. In place of the satisfaction-theoretic conception of reference determination, various authors proposed "causal" or "historical" theories of reference, on which reference determination proceeds by way of the causal or historical antecedents of the use of a given expression. (Early versions were proposed by Chastain (1975) and Stampe (1977), though Barcan Marcus (1961, 1971, 1972), Putnam (1975), Kripke (1972, 1977), Donnellan (1966, 1968, 1979), Evans (1973, 1975, 1979), and others published influential relevant early work on this as well.) What these views have in common is the idea that in specifying who or what the expression refers to – alternatively, in specifying who or what the speaker refers to when she uses that expression on an occasion – the theorist must ineliminably appeal to items or properties in the speaker's "external" environment. Where the satisfaction-theoretic account had it that the reference is determined by way of the satisfaction of criteria, the "externalist" account has it that the reference is determined in some more or less complicated way by appeal to the environmental object or property that plays the relevant role in the history of the use of that expression. In some cases, the relevant role is the history of the use of that expression in the linguistic community itself: a name, for example, might be thought to refer in virtue of its being part of a name-using practice that can be traced back to an original baptismal act, where the person or thing named in the original act is the referent of the name as it is used by anyone participating in this practice; or a predicate might be thought to refer in virtue of its being part of a sophisticated practice of the "division of linguistic labor" on which ordinary speakers defer to experts, where the experts themselves employ the predicate in scientific theorizing (in which case the reference might be the most salient natural property responsible for their use of the predicate). In other cases, the relevant role is simply the object or property that elicited this referring use by the speaker on this occasion: a demonstrative such as 'this' or 'that' will refer to whatever object or property the speaker was attending to as her target (whatever properties she happened to think that object satisfied). These views require further refinement, of course; but it has

struck many that this is the sort of project one ought to embrace in the theory of reference.

Not long after "externalist" views about linguistic reference began to emerge, a number of authors extended the "externalist" analysis from language and linguistic (or speech act) reference to thought. In retrospect, the extension from language to thought is natural. Many mental states, such as thoughts, have representational content. It is natural to think of the content of a mental representation as how the mind, when in that state, represents the world to be. It is also natural to think that how the mind represents the world to be depends on at least two things:[2] *who* or *what* is being thought about, and what is being attributed to who or what is being thought about. An externalist account of both dimensions is natural. Consider for example perceptual thought. In perceptual thought, who or what is being thought about will be determined in part by one's causal-perceptual relations to one's environment; and (there will be cases in which) what properties one ascribes to that item in thought is determined in part by the "external" properties with which one oneself has been in causal-perceptual contact in the past. Nor is the externalist analysis limited to perceptual thought. For insofar as one expresses one's thoughts in language, it would seem that an "externalist" account of the content of one's thought is appropriate whenever an "externalist" account of the reference of one's linguistic items is appropriate. If so, externalist analyses are relevant far beyond the case of perceptual thought. Views of these kinds, which we might label "attitude externalism," were developed by Burge (1979), McGinn (1982), and McDowell (1986).

Even as attitude externalism was being developed, various authors had a suspicion that the epistemic implications of this doctrine might be far-reaching. Here I highlight two potential problems, both of which purport to bring out these implications in connection with a subject's self-knowledge of her thoughts. Following the taxonomy first introduced by Martin Davies (2000) and subsequently developed by Jessica Brown (2004a), I will call these problems the "achievement problem" and the "consequence problem." Both assume that a thinker's judgments regarding her own occurrent thoughts enjoy a special ('first-personal') authority: one can know from the armchair what one is thinking merely by reflecting on one's thoughts. If this is so, then the assumption of attitude externalism appears to give rise to two problems,

[2] "At least": there may be more that goes into this. Perhaps there is the "how": *how* what is being thought about is being thought about. What I say above is consistent with the relevance of the "how."

both of which are nicely summarized by Sarah Sawyer in her contribution to this volume:

> According to the achievement problem, if [attitude externalism] were true, then we could not achieve the kind of privileged access to the contents of our psychological states that we think we have. According to the consequence problem, the hypothesis that we do have privileged access to the contents of our [externalistically] individuated psychological states apparently has the prima facie absurd consequence that we can have broadly a priori knowledge of the environmental conditions which serve in part to individuate those states. (*Chapter 4, this volume*)

A good deal of the early work developing attitude externalism attempted to respond to these worries.

Burge's influential 1988 aimed to show that, at least when it came to a certain range of thoughts, the achievement problem was not a problem at all, owing to the self-verifying nature of a class of judgments he called "*cogito*-like" judgments. These were self-ascriptions of the form '(With the very thought) I am thinking that p' or '(In this very judgment) I hereby judge that p'. Burge noted that these judgments will invariably be true simply in virtue of being made; and he noted that the assumption of attitude externalism does not jeopardize them, for the simple reason that the very content being self-ascribed is a constituent in the thought or judgment itself, so whatever conditions individuate the first-order thought or judgment – the thought or judgment *that p* – will also individuate the higher-order (self-ascriptive) thought or judgment – the thought or judgment that *I myself am thinking/judging that p*. The result is an account of a class of judgments that manifested what Burge called our "basic self-knowledge." While I think it is fair to say that most people in the debate thought that Burge's account of basic self-knowledge was sound, not everyone agrees. Several chapters in this volume address the soundness, scope, and details of this account. (See for example those by Brueckner, Fernandez, and Haukioja.)

Two claims underlie Burge's account of basic self-knowledge. These claims (which Burge defended at some length) pertain to the nature of our knowledge of our thoughts. First, he claimed that to know one's thought, when one is thinking that p, is to know that one is thinking that p. Second, he claimed that one can know that one is thinking that p, even though one can't distinguish the thought that p from other thoughts that one would have had, if one had grown up in some counterfactual situation. This combination of claims has been challenged, in ways that go to the heart of our conception of self-knowledge. To begin, several authors have questioned the significance of the sort of knowledge one has when

one (merely) knows that one is thinking that p. Following the influential work of Paul Boghossian, these authors have begun to think that we should expect more from an account of self-knowledge of one's thought. Such authors have argued that we should expect a theory of self-knowledge of thoughts to vindicate *the transparency of thought content*: the idea that, for every thought a thinker can think, she can tell from the armchair when she is thinking a thought with that content, and so can discriminate that thought from all other thoughts with distinct contents. Boghossian (1992a, 1994, 2011) had argued that any theory that violates transparency will represent subjects as not always in a position to tell from the armchair whether their reasoning is valid, and so will jeopardize the "*a priority* of logical ability" (in Boghossian's words). Others, following Boghossian, had argued that the violation of transparency will give rise to an inability to capture the agent's point of view (see Wikforss 2006, 2008a; but see Goldberg 2002 for a contrary view). Since it is widely acknowledged that attitude externalism is incompatible with this principle of transparency, Boghossian draws the lesson that such theories are not acceptable; and his paper has generated a lively debate regarding the status of the demand for transparency. Many of the chapters in this volume address this question. (See for example the chapters by Boghossian, Sainsbury and Tye, Wikforss, Gertler, Jackman, Ebbs, and Goldberg.)

It is also worth remarking that, even if it is sound, Burge's account of basic self-knowledge only covers a restricted domain of the phenomenon of self-knowledge. (Burge himself was under no illusions on this score.) We might want to know how to extend an account of "externalist" self-knowledge beyond the knowledge manifested in the class of cogito-like judgments. Burge himself (1996) sought to do so; and several of the chapters in this volume also seek to ask about the nature of first-person authority more generally. (See for example those by Ebbs, Wright, Stalnaker, and Jackman.)

Turning next to the consequence problem, this problem was first developed by Michael McKinsey in his influential 1994. (For this reason the problem is sometimes labeled "McKinsey's paradox," and the recipe for generating the so-called paradox is sometimes called "McKinsey's recipe.") When it is presented as a 'paradox', the problem consists in the fact that three propositions, all of which are thought (at least by proponents of attitude externalism) to be plausible, appear jointly inconsistent. The first is the principle of first-person authority itself, according to which, for all thoughts that p, whenever S thinks that p, S knows from the armchair that she thinks that p. The second is the doctrine of attitude externalism itself: for some attitudes, being in that attitude (bearing an

attitude to that particular content) requires the existence of some "external" condition. The third is an anti-skeptical thesis, to the effect that one can't tell from the armchair that external-world skepticism is false. These three propositions are thought to be jointly inconsistent on the assumption that, as a philosophical thesis, attitude externalism itself can be known from the armchair (if it can be known at all). Thus it would appear that a subject who thinks that water is wet is in a position to deduce from the armchair that external-world skepticism is false: she can combine her knowledge of her thought, together with her knowledge of the truth of attitude externalism, to derive the existence of some "external" condition – of which the falsity of external-world skepticism is a trivial implication.

Various replies to this worry have been presented in the literature. Of these, two types are prominent. One allows that the reasoning behind the paradox is both valid and involves premises knowable from the armchair, but it denies that epistemic warrant can be "transmitted" across this sort of inference (Wright 1986, 2000, 2003; Davies 2000, 2003a, 2004; and Sawyer 2001, 2006; see also Brown 2004a for a critical discussion). The other denies that the premises are all knowable from the armchair. Most replies of this type argue that one cannot tell from the armchair, regarding one's thought, that it is a thought whose availability requires the existence of some "external" condition (see McLaughlin and Tye 1998a, 1998b; Goldberg 2003a, 2003b, 2003c, 2007). In this volume, several authors suggest other replies to the McKinsey paradox (see the chapters by Sawyer and Haukioja).

2

The foregoing characterization of the debate enables us to situate many of the chapters in this volume in the discussion of externalism's epistemic implications. But what motivated this volume was not merely the hope that there are new things to say about these topics. I hope that these contributions make interesting and novel contributions to those older discussions. But this hope is not groundless; there are reasons to think that there should be new things to say about these older debates. In particular, in the last two decades several developments in the philosophy of mind, the philosophy of language, and epistemology make it worthwhile to reconsider semantic externalism's implications for self-knowledge and skepticism. In the philosophy of mind and language, these developments include various versions of 2D semantics, the emerging popularity of so-called transparency views of self-knowledge (owed originally to Gareth Evans, and defended more recently by Alex Byrne

(2003, 2010, 2011)), the development of a new theory of concepts (the originalist account in Sainsbury and Tye 2014), and recent developments in the theory of understanding. In epistemology, these include various doctrines pertaining to the semantics of 'knows' (such as contextualism, contrastivism, and pragmatic encroachment), a renewed focus on anti-luminosity arguments (deriving from the work of Williamson (2000)), and developments in the epistemology of testimony as well as in the epistemology of understanding. These contributions bring these novel developments to bear on the older debates.

In this respect, this volume aims to contribute to the literature in two ways. First, it aims to update the literature pertaining to the (semantic and epistemic) implications of "externalist" views of mind and language. Second, it aims to bring these debates regarding the implications of the "externalist" views in mind and language to other issues in contemporary philosophy of mind and language. I will now describe the individual contributions themselves.

3

I have grouped the chapters in this volume into three main groups corresponding to their main thematic orientation. In the first part are those focusing on the mechanisms of self-knowledge and the semantics of ascriptions of self-knowledge. In the second are those focusing on issues of transparency and the nature of one's grasp of one's own concepts. In the third part are those focusing on metasemantics and the nature of content itself.

3.1

The volume begins with the four chapters focusing on the mechanisms of self-knowledge and/or the semantics of ascriptions of such knowledge.

In "Luminosity and the KK Thesis," Bob Stalnaker explores the scope of Williamson's anti-luminosity argument. After arguing that phenomenal properties are not the best place to challenge Williamson's (2000) anti-luminosity argument, Stalnaker argues that purely epistemic or doxastic properties fare better. In this he follows Selim Berker, who had argued (in Berker 2008) that the anti-luminosity argument succeeds only regarding conditions that are constitutively independent of thought and judgment. In the light of this idea, Stalnaker considers a KK thesis, according to which knowing that p puts one in a position to know that one knows that p. Such a principle renders knowledge that p luminous in Williamson's sense, and so this

KK principle should be susceptible if Williamson's anti-luminosity argument is fully general. And yet, Stalnaker argues, Williamson's anti-luminosity argument does not undermine such a KK thesis – Williamson's own claim to the contrary notwithstanding. Whereas Berker's argument against the generality of the anti-luminosity argument lead him to reject that safety (as Williamson construes it) can be captured in terms of a margin-for-error principle, Stalnaker's argument calls into question both Williamson's margin-for-error principle as well as Williamson's safety condition on knowledge. Although Stalnaker himself does not put the point this way, the significance of his argument for those with an interest in externalism and self-knowledge is straightforward: Stalnaker's argument paves the way for externalists to continue to insist on a limited form of luminosity in the sort of *cogito*-like judgments Burge had highlighted in his account of basic self-knowledge. After all, according to Burge's account of basic self-knowledge, the conditions judged to obtain in *cogito*-like judgments are not constitutively independent of the judgments themselves.

In "Some Questions about Burge's 'Self-Verifying Judgments'," Tony Brueckner takes up Burge's account of basic self-knowledge and defends it against a recent objection presented by Finn Spicer. Spicer (2009) had sought to show that Burge's argument for the doctrine of basic self-knowledge, and by extension his case for the compatibility of attitude externalism and first-person authority, both fail for failing to square with a necessary condition on substitution within propositional attitude contexts. In response, Bruckner (2011) sought to show that Spicer's (2009) objection posed no threat to Burge's account. In his contribution to this volume, Brueckner seeks to further advance that debate, as well as our understanding of the *cogito*-like thoughts on which Burge's account had focused. Brueckner's aims are two: first, to show that Spicer's (2009) objection had rested on an unacceptably strong construal of sameness of truth conditions in propositional attitudes; and second, to explore several possible mechanisms by which the *cogito*-like thoughts Burge explored might be self-verifying.

In "Self-knowledge: the Reality of Privileged Access," Crispin Wright aims to address several criticisms (presented in Snowdon 2012) of his earlier account of first-person authority. Wright's view is that first-person judgments regarding's one's occurrent states of mind have three features that need to be accounted for: they are *immediate* (in the sense that they are not based on any evidence), *authoritative* (in the sense that they presumed to be epistemically secure, and enjoy this status so long as there are no reasons for doubt on this score), and *salient* (in the sense that they cover the whole range of the mind's occurrent states – no mental

state "eludes awareness"). Wright agrees with Snowdon in two criticisms: first, that Wright's earlier taxonomy of the states regarding which we have authoritative self-knowledge – the phenomenal and the propositional – oversimplifies matters; and second, that his earlier claim to the effect that first-person judgments of the relevant sort are "groundless" was itself confusing (and in some respects confused). Even so, Wright argues that his earlier account of such self-knowledge, as largely an artifact of our language game, remains defensible. If he is right in this, he has paved the way for externalists in the philosophy of mind to endorse an "avowability" conception of self-knowledge that goes beyond what Burge had offered in accounting for what he had called basic self-knowledge. (Wright himself does not draw this conclusion, but it seems clear from what he does say that he would endorse it.)

In "Contrastive Self-Knowledge and the McKinsey Paradox," Sarah Sawyer argues that the contrastive account of knowledge provides the basis for a response to the McKinsey argument for incompatibilism, and makes clear how it is that we can have authoritative self-knowledge of our own thoughts. After developing a contrastive account of self-knowledge, she goes on to diagnose the failure of McKinsey's argument as a failure to appreciate the relevant contrasts. The basic point is that a subject who has the relevant sort of self-knowledge knows that (for example) she herself is thinking that water is wet as opposed to thinking that grass is green; but this is compatible with it's being false that she knows that she herself is thinking that water is wet as opposed to thinking that q (where 'that q' is the BIV-analogue proposition). Since the success of the McKinsey argument trades on the subject's having the latter sort of knowledge, the contrastive account of self-knowledge thus provides a way to endorse the compatibility of externalism and self-knowledge, without having to embrace the unacceptably strong anti-skeptical implications of the McKinsey argument.

3.2

The next part of the book groups together the chapters that focus mainly on the topic of transparency and/or the nature of a thinker's grasp of her own concepts and thoughts.

In his "Further Thoughts on the Transparency of Mental Content," Paul Boghossian returns to the doctrine of transparency, according to which "When our faculty of introspection is working normally, we can know a priori via introspection with respect to any two present, occurrent thoughts whether they exercise the same or different concepts" (p. 98). It has long been acknowledged that externalism about mental content is

incompatible with transparency in this sense. Following Tyler Burge, many authors (including Falvey and Owens (1994), Schroeter (2007, 2013), and Goldberg (1999)) have embraced this implication and have tried to argue on independent grounds that the doctrine of transparency is false. In their recent book *Seven Puzzles of Thought* (2014), Sainsbury and Tye defend a view of concept individuation which similarly has the implication that transparency is false, and they offer their defense of this implication. Boghossian examines their defense and finds it wanting. His main concern, anticipated in several of his earlier seminal papers on this topic (1992a, 1994, 2011), is that without transparency we appear to have no way to construe what it is for a thinker to be rational in the inferences she draws. Sainsbury and Tye (2014) had proposed that we do so by appeal to the reasonableness of higher-order belief and judgment (regarding e.g. the validity of one's own inferences). Boghossian points out, first, that such a view is only applicable, if at all, to creatures who can have higher-order beliefs, and second, that issues of the reasonableness of second-order beliefs appear to raise precisely the same questions all over again. While Sainsbury and Tye had anticipated this worry and tried to spell out how we might assess this reasonableness in terms of how inferences "strike" subjects, where this need not involve explicit belief or judgment at all, Boghossian concludes that such a view needs to be fleshed out before it can be assessed. The conclusion is that work remains for any externalist who hopes to live with the rejection of transparency.

In their reply to Boghossian's contribution, "Counting Concepts: Response to Paul Boghossian," Mark Sainsbury and Michael Tye respond to Boghossian's concerns. They acknowledge that their (2014) "originalist" theory of concepts implies that there can be cases in which thinkers "behave as if they were mistaken about how many concepts their thoughts involve; or about whether a pair of thoughts is contradictory; or about whether an argument fails to be valid through equivocation." But they respond that originalism has the resources to construe such thinkers as rational nevertheless. For, they contend, it "can be *rational* to so behave." Their explanation is by way of the principle (R), according to which "Thinkers who believe contradictions, or who incorrectly assess the validity-status of simple arguments, or who make fallacious inferences from simple arguments, are irrational, unless they have an excuse." And Sainsbury and Tye go on to say that the sorts of cases to which Boghossian appeals are cases in which the subject does have an excuse. In the Paderewski case, given the subject's background (justified) beliefs, the subject is rational in believing that there are two Paderewskis (even though in fact there is only one). In the 'water' case, the subject is rational