Self-Knowledge

EDITED BY
Anthony Hatzimoysis

OXFORD
UNIVERSITY PRESS

Contents

Notes on Contributors

DORIT BAR-ON is a Professor of Philosophy at the University of North Carolina, Chapel Hill.

JOSÉ LUIS BERMÚDEZ is Professor of Philosophy at Washington University in St Louis, where he is also Director of the Philosophy-Neuroscience-Psychology program and Director of the Center for Programs in Arts and Sciences.

SVEN BERNECKER is Professor of Philosophy at the University of California, Irvine.

ANTHONY BRUECKNER is a Professor of Philosophy at University of California, Santa Barbara.

ALEX BYRNE is a Professor of Philosophy in the Department of Linguistics and Philosophy at the Massachusetts Institute of Technology.

GARY EBBS is Professor of Philosophy at Indiana University, Bloomington.

ANDRÉ GALLOIS is Professor of Philosophy at Syracuse University.

BRIE GERTLER is Associate Professor of Philosophy at the University of Virginia.

ANTHONY HATZIMOYSIS is Assistant Professor of Philosophy at the History and Philosophy of Science Department of the University of Athens.

RAM NETA is Associate Professor of Philosophy at the University of North Carolina, Chapel Hill.

DAVID OWENS is Professor of Philosophy at the University of Sheffield.

R. M. SAINSBURY is Professor of Philosophy at the University of Texas at Austin.

CHARLES TRAVIS is Professor of Philosophy at King's College London.

CRISPIN WRIGHT, FBA, FRSE, is Professor of Philosophy at New York University, and Director of the Northern Institute of Philosophy at the University of Aberdeen.

Introduction

Anthony Hatzimoysis

Self-knowledge has had an enduring appeal to philosophical inquiry. It is hard to think of a major philosopher, from ancient times to the present, who refrained from pronouncing on the nature, the importance, or the limitations of one's knowing of oneself as oneself.

If there is a marked difference between ancient and modern views of the matter, it appears to lie in the fact that, for the ancients, self-knowledge is primarily a goal to be achieved, whereas for the moderns it is mainly a puzzle to be resolved. Even for the ancients, though, the goal held little value for those who failed to appreciate what might be involved in knowing the cognitive, volitional, or affective states of one's soul.

Take, for instance, Socrates. In one of the lesser known dialogues, the *Charmides*, Socrates invokes a distinction between reflexive and intransitive states, and applies it in separating a cognitive state's knowledge of itself from our knowledge of ourselves. According to the reasoning deployed there, a state is presented to its owner in a perception, which grounds the owner's knowledge of the state's content, as well as of the occurrence of the state in the owner's soul. The received wisdom that the others are barred from looking into what someone has inside is put into question, as is the claim that one can come to know much about oneself just by looking inward, while shutting out everything and everyone else. The suggestion of dissociating what the state is of from one's awareness of the state's occurrence is firmly resisted, as is the view that one can develop one's good-aiming dispositions without possessing *sophrosune* (best translated as 'sound-mindness' or 'temperance', depending on the conversational context), the ability to securely grasp one's own veridical perceptions, beliefs, and affections. And that was a first philosophical stamp on the significance of self-knowledge: to know that you know what you know and what (you think that you do but, like most of Socrates' interlocutors) you don't know, is a condition of developing the virtues necessary (and, according to Socrates, sufficient) for leading the excellent life.

No sooner is a claim put forward in that dialogue, though, than it is unseated by doubts about its exact meaning, hidden assumptions, or counterintuitive implications.

Hence, the above précis does not aim to offer the definite rendering of some theory of self-knowledge; after all, Socrates would be the last to claim that he possessed a 'theory' on this matter, or that the proper aim of an inquiry into how we may know ourselves, is that of theory-construction. The exegetical issues that arise from a close reading of a classical text might be numerous, but are not for the present. It is simply worth noting that, from the beginning of systematic engagements with the question of self-knowledge, it was well appreciated that there are no short cuts to a neat philosophical answer.

What makes self-knowledge such a perplexing phenomenon? The essays commissioned for this volume seek to deepen our understanding of self-knowledge, to solve some of the genuine (and to resolve some of the spurious) problems that hold back philosophical progress on that front, and to assess the value of some classic moves in the debate over the epistemic status of self-ascriptions. Some of the chapters discuss features of self-knowledge that appear to account for its unique—and, in that sense, peculiar—status; some advance straight proposals on how to solve a number of crucial issues; and others take a step back to consider the terms in which we set the questions to which a philosophical theory of self-knowledge is to provide the answer.

Perhaps part of the philosophical peculiarity of self-knowledge is due to the fact that we appear to possess a peculiar method of cognitive access to ourselves. The peculiarity at issue, though, might be nothing more than a matter of a quite specific, if not singular, employment of an otherwise metaphysically innocent capacity of conscious awareness. Philosophical disputes about the exact mechanism of how such a capacity operates, are still some way from being resolved; yet, as it is testified by the majority of the chapters in this volume, there is wide agreement that our access to at least some of our own states is 'privileged'. Ram Neta opens the discussion by exploring what exactly is meant when it is claimed that we enjoy 'privileged access'. He sets up his analysis by inquiring into the reasons why it is generally thought that we have privileged access to only some mental states (such as believing, imagining, or feeling) and not to others (such as knowing, remembering, or seeing). After surveying the main answers given in the literature, he argues that they fail to account for the issue at hand, either because what they state is false, or because they address the wrong kind of question. The answer Neta puts forward seems to account exactly for the phenomena under consideration, without rendering inoperative any of the substantive theories regarding the particular mechanisms of privileged access.

While Chapter 1 throws its net as wide as possible, Chapter 2 aims to dive a little deeper by defining its field of exploration to accounts of self-knowledge that take seriously the lessons of the representationalist approach to mental content. Sven Bernecker makes a strong case for two rather bold claims. first, that one cannot know a priori that one knows a particular proposition as opposed to being incapable of having any knowledge states; and, second, that if one does have knowledge states, one can have a priori knowledge of their specific contents and, hence, one can know a priori that one knows a particular proposition as opposed to some other proposition.

To ground his claims, Bernecker works through a number of crucial distinctions concerning the types of a priori knowledge available from a first-person perspective, the kinds of second-order knowledge that pertain to knowing one's own states, and the notions of propositional content that should inform our account of self-knowledge. Thus, although that chapter has as its declared aim to address a specific question from within a particular approach to mental content, its ramifications for a number of fundamental issues make it required reading for those interested in the epistemology, the metaphysics, and the semantics of self-ascriptions.

Much of the philosophical debate about self-knowledge operates on the assumption that there are several things we can claim to know about ourselves without appealing to empirical evidence. More precisely, it is often stated that we possess the capacity for at least minimal self-knowledge, whereby we can each know what thoughts our own utterances express, without waiting for the deliverances of empirical inquiry. Gary Ebbs undertakes to clarify the exact meaning of that assumption and to show how, when properly understood, that assumption relates to a common view of epistemic possibility. The standard analysis of epistemic possibility implies that if we suspend all of our empirical beliefs, we can each make sense of actually being in any possible world that is subjectively indistinguishable to us from the actual world. The idea that certain things might be subjectively indistinguishable to us plays a pivotal role in the debate over the pros and cons of externalist or anti-individualist theories of mental content. Through a rigorous discussion, Ebbs shows that there is no way to reconcile anti-individualism and minimal self-knowledge, with the standard analysis of epistemic possibility.

The papers by Bernecker and Ebbs lend support, in their own distinctive ways, to an externalist account of mental content that stays true to the immediacy and authority that appears to characterize claims to self-knowledge, as expressed in one's present-tense ascriptions of one's own mental states. For externalism, mental content is conditioned by constraints external to the subject, so that most concepts are available to a subject only when he, or his speech community, has had causal interaction with items lying in the extension of the relevant concepts. However, the assertion that we enjoy some sort of immediate or a priori access to our current mental content sounds to some philosophical ears discordant with content externalism, according to which the content of a mental state is dependent on the interaction between a subject and the world. Such an interaction is a matter of contingent historical fact, whose existence cannot be (dis)confirmed by any sort of a priori reasoning operating in the restricted area of one's own psychological states. Mere reflexion, for instance, cannot tell us whether I, or my speech community, have been in the appropriate kind of causal contact with water. Yet, from the introspectively available psychological fact that, e.g., I believe that water is wet, and the externalist claim that I could not believe this, unless I or others in my community have had a history of interaction with water, it appears that a factual conclusion about the world can be drawn, to wit that I or others in my speech community have had a history of interaction with water. The combination of

externalism about mental content, with unobjectionable claims about ordinary self-ascriptions of that content, generates a major paradox, whose solution appears to call either for a radical revision of the externalist view of content, or for an outright denial of the immediacy and increased epistemic security that seems to characterize psychological self-ascriptions. In Chapter 4, Crispin Wright marks a different route in this terrain by questioning the idea that the validity of a piece of reasoning from warranted premises, commits us to the truth of its conclusion: to be thus committed, the reasoning should transmit epistemic warrant; but as Wright's exploration of different epistemic scenarios shows, there are several cases where the kind of warrant enjoyed by the premises fails to transmit to the validly drawn conclusion. Hence—to return to our example of reasoning from my self-ascribed belief about the wetness of water—even if the premises of the reasoning are warranted by some sort of introspective, or otherwise a priori, method 'from our armchair', and the reasoning employed is valid, it should not move someone antecedently open-minded about the conclusion, to the conviction that the conclusion is true.

Wright's paper addresses some crucial issues in the philosophical understanding of warrant. It thus presents an example of how the specific problems that surround the topic of self-knowledge, instead of being approached as peripheral cases to which ready-made theories can be applied, may themselves illuminate some fundamental issues in epistemology. As far as the epistemology of self-knowledge is concerned, Wright reinforces the externalist cause, by indicating that ordinary self-knowledge of intentional states remains available 'in the armchair' even if we are persuaded of externalism about intentional content, not because there is prior empirical reason for confidence in its presuppositions, but because, given the context of ordinary self-ascriptions of such content, no empirical reason, or indeed evidential reason, of any kind is required.

If the above analysis is correct, externalism as such does not render illicit the employment of 'armchair methods' for knowing oneself. Introspection, after all, as conducted from the comfort of our armchair, has been considered one of the major contenders in the explanation of how we acquire knowledge of our mental activities; yet, in the long debate over the legitimacy of the very idea of 'perceiving inward', one of the things that remains unclear is whether the appeal to introspection can explain the privileged access that is distinctive of—or peculiar to—self-knowledge without postulating any epistemic abilities in addition to those needed for knowledge of other subject matters. In Chapter 5, Alex Byrne confronts this issue head on by putting forward an economical theory of self-knowledge which appeals to nothing beyond rationality, normal intelligence, and conceptual capacity, while retaining some of the best of empiricist insights into how ordinary experience can be a reliable (though not infallible, or incorrigible) source of knowledge about our own thought activities. Byrne's handling of those issues is significant for at least two reasons. On the one hand, he spells out the major—and rather unexpected—implications of a coherent, empirically grounded, non-extravagant theory of how we know that (and what) we

currently think—arguing inter alia that one can know that one is thinking about some thing by 'hearing' oneself speak about that thing in inner speech. On the other hand, he articulates epistemic rules for the formation of belief that invoke the world-directedness of our attention in responding appropriately to questions that pertain to our knowledge of our own belief states. In particular, his employment of the so-called transparency stricture, that a rational agent gets herself in position to answer the question whether she believes that some thing is the case not by searching inside her, but by putting into operation whatever procedure she has for answering the question whether that thing is the case, may turn the tables on those who contend that introspection ought to limit our mind's sight to the world within.

The transparency method enjoins us to discover our beliefs by turning our attention outward—not inward to the states of our soul, but to the worldly facts that the contents of our states represent. Yet, that 'discovery' might in practice be more a matter of deciding on, or forming a new judgment about, what we (are rationally required) to believe, than a genuine disclosure of beliefs that are already in place. Nevertheless, appeals to the transparency method are currently on the rise, along with increased differentiation on how exactly that method should be understood. Through a lucid discussion of the main categories of the relevant cognitive states, including occurrent judgments, implicit dispositional beliefs, and ordinary dispositional beliefs, Brie Gertler argues that, despite its current popularity, the method of transparency fails to explain our privileged access to our own beliefs. On her diagnosis, the apparent plausibility of the transparency method derives from a rationalistic conception of belief, which envisions a constitutive tie between beliefs and evidence. One of the wide-ranging implications of Gertler's analysis is that even on a highly rationalistic conception of belief, the method of transparency will not reveal one's beliefs.

The prospects of a rationalistic conception of belief-formation have been well explored in the philosophical work of André Gallois. In Chapter 7, Gallois takes us a step back and reconsiders some basic claims in the literature on self-knowledge, claims whose very familiarity might make us overlook the fact that it is not, after all, so clear how exactly they should be linked to each other. The claims under consideration concern (i) Moore's Paradox (instantiated in sentences of the form: p, and I do not believe that p), (ii) one's beliefs about one's own belief states, and (iii) beliefs as to whom one believes oneself to be. His treatment of the interrelation between those three issues passes through a penetrating discussion of the transparency condition, and brings to the fore the significance of the 'augmentation capacity', i.e. the agent's capacity to retreat from something being so, to her simply believing it is so; an agent in want of 'augmentation capacity' lacks the ability to represent the fact that she believes something as a fact about her psychology—and that ability is necessary for having an integrated, and rationally revisable, view of the world. Gallois draws the various strands of his discussion together by maintaining that the knowledge acquired through the rational exercise of augmentation, while required for having genuine

knowledge of oneself as oneself, delivers at most self-knowledge that is 'minimal', since it leaves completely open the nature of the self known.

The philosophical debate over the nature of self-knowledge has to a large extent been conducted in terms of how we communicate in discourse our mental states. In recent years, Dorit Bar-On has articulated a comprehensive account of first-person present tense mental ascriptions, that purports to answer several of the epistemic, semantic, and metaphysical questions pertaining to self-knowledge. Her approach is quite distinct on a number of fronts. First, it does not take for granted the privileged access-thesis subscribed to by so many of the participants in this debate; secondly, it affirms the semantic continuity between self-ascriptions and ordinary propositions, thus resisting the naively expressivist contention that avowals are a kind of linguistic substitute for reflexive behaviour, such as crying or winching, devoid of truth-evaluable content; thirdly, it rebuts the epistemic model of introspectionism, which, in both its Cartesian and materialist version, renders avowals subject to all sorts of failings that beset reports drawn on the basis of external observation, and thus eliminating the distinctive security that characterizes first-person self-ascriptions of presently occurring mental states. In Chapter 8, Anthony Brueckner presents a critical study of Bar-On's neo-expressivist theory, by scrutinizing her employment of the idea of 'expressing a mental state', by querying her treatment of self-verifying assertions, and by questioning her theory's ability to account (in an informative and non-circular manner) for the phenomenon that beliefs about contingent matters of fact, regarding one's mental states, require no ordinary evidence and are not open to epistemic assessment. In her response, Bar-On faces up to Brueckner's challenge by explicating how her approach illuminates two important characteristics of avowals, the immunity to error through misidentification of the person having or undergoing the self-ascribed mental state, as well as (what is rather more contentious) the immunity to error through misascription of the mental state's content. She also addresses some of Brueckner's specific complaints concerning the epistemological force of the notion of expression. While acknowledging their significance, she maintains that, if stated at the appropriate level of discussion, her theory provides substantial, non-deflationary responses to those complaints.

One of the philosophers who revolutionized our approach to mental content, and whose probing remarks set the terms for much of the contemporary discussion of the meaning of self-ascriptions, is Gottlob Frege. In "Viewing the Inner", Charles Travis undertakes the difficult task of extrapolating Frege's lessons for philosophy of mind, through a most careful delineation of the Fregean constraints on having ideas, entertaining thoughts, and exposing oneself to the risk of error that is involved in judging things being, in the world at large, as well as with one's own mental life, in a particular way. Through an examination of how the Fregean analysis may be employed in elucidating visual appearings, Travis specifies what exactly counts as a 'viewing' of either the 'outer' or the 'inner', and to what extent—and in what precise sense—a person's experience is shareable by others.

Frege remarked that everyone is presented to himself in a special and primitive way in which he is presented to no one else. However, that remark seems to jar with the unobjectionable claim that several things said of me by me using the first-person pronoun "I", can be effortlessly understood and accurately reported by you either by using the third-person pronoun "he"—when your report about me is addressed to a third party—or the second-person pronoun "you"—when you talk to me about myself. How should we account for the alleged distinctiveness of the first-personal way of being presented to oneself, without rendering the sense of "I", and of the sentences expressed in the first-person, private and unsharable? In Chapter 11, José Luis Bermúdez addresses this question, first by showing the inadequacies of the standard answers encountered in the flourishing field of theorizing about indexicals, secondly by pointing to the limitations of attempting to explicate the relevant phenomena just by analyzing the fact that the subject is in receipt of information that is immune to error through misidentification, and thirdly by proposing a novel way of doing justice to the insight that the sense of "I" reflects a distinctive way of being presented to oneself. On Bermúdez's proposal, what an understanding of a use of "I" requires is that one has the capacity to locate the utterer in space. The distinctive way in which I am presented to myself can be systematically related to the distinctive way in which I am presented to others, because I possess a distinctive ability to locate myself in objective space.

In his discussion of those issues, Mark Sainsbury acknowledges some positive aspects in the above proposal, but he argues that the whole project within which Bermúdez's account operates is misguided: possessing or devising some method for identifying the utterer of a sentence expressed in the first-person is in no way necessary for making proper sense of an "I"-utterance. Sainsbury argues that all one needs to know to understand "I", whether as a user of it or as a hearer, is simply that English speakers should use "I" to refer to themselves. In defending this prima facie modest claim, Sainsbury explicates why there is no special ego-distinctive content that supposedly holds the key to the puzzle of self-knowledge, while he puts to rest some of the main philosophical worries concerning the use of "I" for self-reference.

The final chapter in the volume revisits the moment where the modern philosophical thinking on those issues began. Descartes's *Meditations* had a catalytic effect on subsequent philosophical approaches not only for the substantial answers it provides to a host of fundamental questions, but also for its analysis of how awareness of oneself is involved in grounding one's claim to indefeasible knowledge. A pair of conditions are particularly significant for the realization of the meditating process: first, that there exists a method by which one can directly control one's cognitive and volitional states; second, that, in meditating, one must conceive of oneself as the agent of one's own meditation. The question raised by David Owens is whether there is a form of mental activity that can satisfy both specifications. Through solid argumentation that critically addresses some of the most prominent contemporary discussions of that issue, Owens shows why none of the major candidates for the task of Cartesian meditating, including plain reasoning, deliberation, reflective maintenance, and reflective reasoning, can

actually fit the bill of offering rational self-control over one's beliefs and decisions, while being aware of oneself as the agent of one's meditating.

I would like to express my thanks to Sven Bernecker for his sound advice at various stages of this project; José Luis Bermúdez and Brie Gertler for their comments on the draft of the "Introduction"; Peter Momtchiloff and Catherine Bennet for their editorial support; and the anonymous readers of the Press whose extensive critical remarks and detailed recommendations were of substantial help in preparing the final version of the volume. I would like above all to thank the authors who graciously responded to my critical queries and revised their drafts in light of the referees' reports. Socratic *sophrosune* (understood as both temperance and sound-mindness) is, perhaps, the chief virtue exercised by all parties involved, for seeing a multi-author project to its successful completion.