Philosophia (2019) 47:335–342
https://doi.org/10.1007/s11406-018-0006-6

Belief and Commitment: Commentary on Annalisa Coliva, *The Varieties of Self-Knowledge*, London: Pallgrave Macmillan (2016)

Aaron Z. Zimmerman[1]

Published online: 26 July 2018

In search of the indubitable, Descartes tried to doubt everything he could. He found what he was looking for in his famous "cogito ergo sum." I cannot coherently doubt that I am "cogitating," as doubt is itself a form of cogitation. And it should have gone without saying—even though it didn't—that cognition requires a cogitator. I cannot coherently doubt the existence of my mind.

In fact, Descartes' could not intentionally doubt as much as he could without representing himself as doubting as much as he could. His representation of himself as "doubting as much as much as I can" was the "intention in action" (Searle 1983) that sustained Descartes in his ruminations. And his knowledge that he was at least trying to doubt as much as he could "encoded" a large body of more substantive self-knowledge to which we are now privy. Biologically, no animal can experience doubt without having developed a nervous system capable of impeding some process of "sensorimotor coupling" so as to focus more of its neuro-cognitive resources on an evaluation of its sensory inputs and their effects. In other words, animals who doubt have learned to check themselves. So a person's bare knowledge that she is doubting, when coupled with some instruction in evolutionary psychology, will allow her to infer a great deal about herself: that she is an animal with a complex nervous system which somehow enables her to represent goals and various obstacles to their achievement.

How much of this self-knowledge is amenable to philosophical analysis? Though Descartes did write a great deal about the origins of our minds and contributed a great deal to the science of his day, his *Meditations* were focused on the "concept of doubt itself," where this might be equated with normal use of "doute" and synonymous expressions in the social life of Europeans in the first half of the seventeenth century. Then, as now, we can understand expressions of doubt and identify our own doubts as such without reflecting much on the evolutionary origins of our minds. Nor need we reflect on the characteristic function of doubt and attributions of doubt to regularly report

✉ Aaron Z. Zimmerman
aaronzimmerman@ucsb.edu

[1] Department of Philosophy, UC Santa Barbara, Santa Barbara, CA, USA

our doubts and intentionally attempt to assuage the doubts of others. We might put this by saying that we don't need science to know what "doubt" *means* even if we do need science to know what doubt *is*. If we ignore a century of doubts about the epistemological substance of any analytic-synthetic distinction, and endorse this way of conceptualizing the matter, we might then join Descartes in thinking that the premises of his cogito are classically philosophical in their *a priority*. Everyone we will credit with understanding "doubt" or "doute" knows that doubt is a form of thought. So everyone who speaks our language is in a "position" to see that the attempt to doubt that one is thinking undermines itself. If you succeed in doubting that you are thinking, you therein insure the truth or accuracy of the very proposition whose truth you subjected to doubt. One who reasons as Descartes did does not eschew such doubts on the basis of perception or an analysis of the results of scientific experimentation. Descartes' rejection of the most extreme forms of self-doubt was non-observational and positively pre-theoretic.

Of course, epistemologists since Descartes have asked how certain it is that we are thinking. Does this certainty vastly exceed our certainty that we are animals? Famously, *Descartes'* certainty that he was thinking so greatly exceeded *his* certainty that he was an animal that he inferred that he *wasn't* an animal from his certainty in his thinking, once he conjoined that certainty with a set of metaphysical auxiliary hypotheses robbing matter of the ability to think. More recently, Saul Kripke (1980) has argued that our sensations cannot be states of our nervous systems because sensations are "directly" known whereas neurological states must be known under a description or "mode of presentation." And David Chalmers (1996) has argued that pain and other "qualitative" aspects of our mental lives must be nonphysical because we can imagine zombies who are physical duplicates of us but who lack all "qualia," and what is imaginable must be in some sense possible. In all these cases, epistemologists have sought to draw heavy-duty metaphysical conclusions from theses articulating the security and directness of self-knowledge.

In this wonderful new book, Annalisa Coliva broadens our focus from Descartes' self-verifying foundations for knowledge and Kripkean demonstrations of dualism to examine self-knowledge more generally, reflecting the evolution this topic has assumed in the literature generated by analytic philosophers over the last twenty years. I think this is a positive development and adds considerably to the depth of Coliva's analysis. There is more to know about yourself than what you are thinking and feeling at the present moment. And Coliva has a number of meaningful things to say about both our relatively automatic forms of self-knowledge and those forms of self-understanding that are "cognitive achievements" and so largely stripped of the certainty Descartes sought and found in his cogito. Coliva advances a "pluralistic" account of self-knowledge, and I think we should all be pluralists in Coliva's sense.

However, there is another contrast with Descartes et al. that is worth worrying over. Whereas Descartes, Kripke, and Chalmers begin their arguments for dualism with a set of epistemological premises, Coliva begins her epistemological analysis of self-knowledge with a metaphysics of mind, distinguishing sensations from perceptions, perceptions from beliefs and other "propositional attitudes" (e.g. desires and intentions), and arguing for emotion as a distinctive category with similarities to both sensation and belief. A person's psychology is supposed to include all of these things, as well as character traits, quirks of personality and other more obviously "dispositional" features of her profile. The "varieties" of self-knowledge featured in the book's title are the varying methods people use to knowledgably characterize facts about these varying components of their minds. And though it stands to reason, as Coliva argues, that our access to our "first-order" minds is as heterogeneous as these minds themselves, I wonder whether Coliva has adequate argumentative support for certain crucial components of her metaphysics of mind, drawn, as they are, in advance of those insights she draws from epistemological reflection.

More specifically, though Coliva employs terms drawn from our folk psychology when articulating her metaphysics of the mind, she does introduce a "technical" distinction—really a meta-technical distinction—that is crucial to her epistemological project. For Coliva argues that, "on reflection and contrary to what mainstream philosophy of mind seems to hold, our notion of an intentional mental state is not univocal" (2016: 27).[1] To be clear, Coliva agrees with the "mainstream" that there is a class of intentional mental states, and she thinks that they are all appropriately characterized as propositional attitudes that cannot be possessed by subjects who lack the concepts we use when expressing them, reporting them, or attributing them to others in speech. (According to Coliva, a dog cannot believe that her owner is home if she lacks the concept "owner," nor want her owner's leftovers if she lacks the concept "leftover.") Indeed, because they are supposed to share this descriptive essence, Coliva says that the mental states that populate the extension of "intentional mental state" are more unified than those in the extension of "jade," which we now know to be bifurcated between two different gemological kinds: jadeite and nephrite (Coliva 2016, 37; cf. LaPorte 1996). Nevertheless, despite their acknowledged metaphysical unity, Coliva draws on work by Akeel Bilgrami to argue that intentional mental states come in two varieties: *dispositions* and *commitments*, which are supposed to differ in kind from one another. Metaphysically, commitments are "within our direct control," whereas dispositional propositional attitudes are not. Moreover, commitments result from reasoning and "assessments" of evidence whereas dispositions do not. And these metaphysical differences are supposed to ground or explain a set of social and normative distinctions. We hold each other responsible for our commitments, but not our dispositional propositional attitudes, and Coliva thinks this is exactly how we should proceed.

Now I have some doubts about Coliva's metaphysical distinction between dispositions and commitments. For one thing, it cannot be denied that control comes in degrees. (I can control both my left hand and my right, but because I am right-handed, I can better control the latter than the former.) Perhaps the more control we have over a state of our minds, and the more effort we have spent in generating and sculpting that mental state, the more prone we are to thinking of its content as something to which we are committed. But whatever "maximal control" might entail, it seems unlikely to me that we have it over the vast majority of our commitments, whether these be doxastic commitments to the truth of various propositions or practical commitments to the value of those states of affairs we work to achieve. For example, is my commitment to the

[1] See too "Our notion of an intentional mental state is not univocal" (2016: 188). I guess I do not have the mainstream concept of "intentional mental state." I do not think beliefs and intentions are "propositional attitudes" because I think many other animals have beliefs and intentions without being able to construct propositions. (Admittedly, many contemporary philosophers of mind think they can consistently attribute propositional attitudes to animals incapable of communicating in sentential terms. See, e.g. Fodor's (1980) language of thought hypothesis seemingly embraced by Coliva (2016: 170).) Though I have argued that the propositional attitude analysis is at best misleading (Zimmerman 2018), this doesn't prevent me from evaluating Coliva's claim that the conception evoked by "intentional mental state" when it is used by mainstream philosophers is too coarse, and that progress could be made by participants in the mainstream discussion were they to agree to distinguish dispositions from commitments in the way Coliva suggests.

value of my marriage genuine if I am not disposed to honor and respect my wife? Not if "commitment" is understood in anything like its usual sense.

Of course, we can use "commitment" to denote a speech act rather than a state of mind. According to this usage, a groom commits himself to honoring and respecting his wife by assertively uttering, "I promise to honor and respect you," with his eyes on his bride during the course of a ceremony to which there are ample witnesses. Indeed, he therein commits himself to the duties of a husband even if he has no intention of honoring or respecting his wife and is hatching plans to betray her while performing his vows. Moreover, it is true that by performing this speech act, our groom opens himself to the kinds of criticisms Coliva identifies as definitive of commitment. By saying what he does in apparent seriousness, the man presents himself to both his bride and the assembled audience as someone who is genuinely committed to playing his part in an honorable and respectable marriage. So if he is not in fact committed to the marital project to which he has committed himself in speech, he can be criticized as insincere: a lying fraud. But this reaction just further highlights the substantive, disposition-constituted nature of commitments when they are understood as mental states rather than speech acts. The groom hasn't really committed himself to the project to which he has committed himself in speech. Commitments are dispositions of a sort.

Perhaps, then, we should focus on the degree to which I can *dispose* myself to honor and respect my wife when assessing the propriety of "holding me responsible" for lacking this commitment. Since dispositions may be more or less amenable to our control, the supposed distinction "in kind" that Coliva draws between commitments and dispositions seems to me to be superimposed upon a more messy set of distinctions in degree between various sets of dispositions, some of which constitute our intentional mental states and some which do not (cf. Zimmerman 2018, ch.2). So while I agree with Coliva that there are differences between, say, beliefs and intentions on the one hand and character traits like courage and honesty on the other, and I agree that these metaphysical differences help explain why we have better first-person access to our intentions and beliefs than we have to our character traits, I think this is best characterized as a difference in the kinds of disposition necessary for the possession of beliefs and intentions when these are compared with those that constitute character traits. As I have argued in a number of different publications, if you believe something at a given time, that entails that you are then disposed to act and reason on that information when you are paying close attention to some stimulus to which that information is relevant and exercising as much control as you can over your response to it. In contrast, our character traits are revealed by our reactions to stimuli just as well as they are by those attentive, controlled actions we perform in light of them. (According to a common conception of character that can be traced to Aristotle, if your first instinct is to lie, you are not yet honest in character, even if you suppress that response for the sake of the virtue in question. Such a subject is "continent," but not yet virtuous.) In favorable circumstances, you can be fairly certain about how you would act were you exercising attention and successfully controlling your mind and body. It's because these conditional dispositions are sufficient for belief and intention that we have whatever degree of first-person epistemic authority we in fact possess over these states of mind. But it would be unwise to place much certainty in a claim about how you would react to a given stimulus were it introduced. So you are rarely as authoritative about your character. In my opinion, the epistemic asymmetries Coliva sets out to explain have their source in differences of these kinds.

Though there is at least one passage in which Coliva seems to accept a picture like this one on which commitments are dispositions of a sort (2016: 208), the bulk of her analysis follows a different route. For, again, the requisite distinction between commitments and dispositions is supposed to be one in kind rather than degree, and once she has this distinction in hand, Coliva sets out to use it to defuse ongoing debates within the philosophy of mind over the nature of self-knowledge: debates she expertly surveys in the volume's subsequent chapters. Coliva levels detailed criticisms against the inner sense theories of Armstrong and Lycan; the inferentialist theories of Gopnik and Cassam; "simulationism" as differently elaborated by Goldman and Gordon; the expressivist accounts offered by Wittgenstein and Bar-On; and a host of more rationalist theories of self-knowledge defended by Burge, Peacocke, Moran and Fernández. A general diagnosis of contemporary theories of self-knowledge emerges from these critiques: Dispositions cannot be known in a distinctively first-person way, which is marked by phenomena Coliva labels "groundlessness," "authority," and "transparency." But commitments can be known in this distinctively first-person way. We can know what propositions and values we are committed to without observation or inference, our commitments are more or less directly accessible to us when we have them, and we are consequently better situated to report our commitments than are the other people we know. So while our friends and acquaintances are rarely (if ever) able to correct our claims of commitment, nothing similar holds for dispositions. Indeed, we are supposed to be able to know a priori that a person can know her commitments without grounds in an authoritative and transparent manner whereas dispositions cannot be known in this manner. To argue for this last claim, Coliva adapts considerations advanced by Sydney Shoemaker (1996): if a subject lacked accurate beliefs or knowledge of her commitments, this would either compromise her possession of psychological concepts or her rationality. There is no such thing as "brute" self-blindness with regard to our commitments, and we must reject any theory of self-knowledge that implies otherwise.[2]

[2] Coliva's positive view also includes the claim that psychological concepts are "drilled into us" as children "blindly" (2016, 191–209). For example, we are supposedly conditioned to replace assertion of a proposition with self-attribution of belief in this proposition: replacing "P" with "I believe P" in appropriate circumstances. I agree with Colvia that a minimally rational agent will not fall prey to Moore's paradox insofar as she will not assert "P" unless she is prepared to attribute to herself the belief that P. (When asserting or judging P one represents oneself as believing p, a representation one undermines by asserting one does not believe what one asserted. Self-undermining assertions and judgments are obviously defective. They are the flip-side of self-verifying assertions and states of mind like "I am speaking," Descartes' cogito, and the belief that one has at least one belief.) I agree, moreover, with Coliva and Bar-On that self-ascription of belief and assertion of what is believed are metaphysically realized in the same discursive mental state and likely have the same neurological realization in humans. But I nevertheless reject Coliva's account of concept acquisition for several reasons. First, we need introspective knowledge to detect the circumstances that warrant the replacement she identifies: We only "replace" a simple assertion with a report of belief when we somehow detect that we are not entirely certain of what we are disposed to assert. (A competent speaker does not replace "A bus is going to hit you!" with "I believe a bus is going to hit you.") How can children detect that they are less than certain of what they are prepared to assert without employing some more or less substantive form of introspection? More generally, I think Coliva's developmental story is impugned by the trajectory of contemporary cognitive science, which supports the view that children bring their full cognitive resources to the task of learning psychological vocabulary, cognitive resources that already include knowledge of their own minds and the minds of others. Indeed consensus is emerging that psychological knowledge evolved in various species whose members remain incapable of grasping psychological vocabulary. See Zimmerman (2018, chapter 3) for details.

Can Coliva wield this apparatus to dissolve the contemporary philosophical debate over self-knowledge? If mainstream analytic philosophers read this book, and they adopt the recommended metaphysics of mind, along with its embedded distinction between commitments and dispositions, will they stop arguing over how we know our intentional mental states? Will philosophers agree that our commitments are known to us without grounds, but with an authority and transparency absent in our judgments about our dispositions, and then move on to debate other matters?

Though it would be nice to have some more agreement with my peers on the nature of self-knowledge, I am more concerned with the application of Coliva's project to our folk psychology. So we must ask, do "belief," "desire," and "intention," as we actually use them to explain ourselves to one another in the course of our social lives, have the same ambiguity Coliva has posited for "intentional mental state" as used by mainstream analytic philosophers? Though I think a positive answer is essential to the positive theory of self-knowledge that Coliva advances in this work, I don't see where she's argued for one in the text.[3]

To see what is at stake, we need to examine cases in which a subject's avowals and self-reports come apart from other aspects of her behavior. So consider a case in which someone who identifies and is identified as a "white person" is accused of thinking black men are more dangerous than white men. Suppose that this subject, W, asserts in reply that black men are no more dangerous than white men, and self-ascribes belief in this proposition by saying "I believe black men are no more dangerous than white men," citing the superficiality or incoherence of racial distinctions as evidence for her first-order claim. Is it possible that W is wrong in thinking that she believes that black men are no more dangerous than white men?

The only possibilities for mistake in this case that Coliva is willing to acknowledge concern linguistic misunderstanding and insincerity. For example, suppose that W tells her friend to be careful walking alone in the evening in a predominantly black neighborhood, but does not offer similar advice when her friend is set to walk through an otherwise comparable neighborhood that is mostly white in population. And suppose that when she is "called out" on this, W justifies herself by saying that black men are in fact more violent than white men. In this case, we would rightly conclude that W doesn't know what "dangerous" means. How can she believe that black men are no more dangerous than white men if she also thinks that black men are more violent than whites? More commonly, a subject who says that she believes in racial equality or religious equality, but who also asserts that blacks are more dangerous than whites or that Muslims are more dangerous than Christians, might be said not to know what "racial equality" and "religious equality" really mean to those of us who are challenging that subject's claim to believe in these things. Since people rarely (if ever) mean exactly the same thing by a given expression, I think linguistic misunderstanding is a more important source of challenges to self-ascriptions of belief than Coliva lets on.

But suppose that we have somehow figured out that W means what we do by "dangerous." Might she still be mistaken in judging that she believes that black men are no more dangerous than white men? According to Coliva, she cannot be mistaken in

[3] In fact, her official statement is that "belief" is univocal but that "commitments and dispositions are two species of the same genus 'belief'" (2016: 123, fn.46). But I'm not sure how this is supposed to square with her repeated claim that usage of "intentional mental state" is equivocal.

this as long as she is *sincere* in her assertion. But what is sincerity? If we define a sincere assertion as one motivated by *belief* in the proposition asserted, we can agree with Coliva that sincerity of assertion entails the truth of a self-ascription of belief in what is asserted. But this "insight" borders on triviality. Of course, if the subject believes what she is saying then she is right that she believes it. But we don't need an intricate philosophical analysis to deliver this result.

So let us suppose, instead, that sincerity in assertion is defined differently, as, say, automatic or unthinking assertion, or as assertion unaccompanied by the intention to deceive. Then I think that W *can* be justly accused of speaking falsely when she says she believes that black men are no more dangerous than white. For it may turn out that though she sincerely asserts this proposition in the sense of "sincerity" we have adopted, and sincerely asserts that she believes it in this same sense, she does not believe what she therein claims to believe.

Now, if "commitment" is understood as a speech act, W cannot be wrong in thinking she is committed to the proposition that black men are no more dangerous than white men. By asserting this in seeming sincerity, she has put it on the record—she has committed to it in the sense at issue—and she can be criticized by her audience if she fails to act and reason on the basis of what she has herein claimed to believe. Indeed, this is the analysis Coliva gives of "self-deception" more generally (2016: 198–200). The self-deceived spouse is indeed committed to the proposition that her husband is faithful and so believes in the commitment sense what she says she does when she self-attributes a belief in his fidelity even if her habit of opening her husband's mail and checking his phone for suspicious texts shows that she lacks this belief in the dispositional sense (and so believes (or suspects) that he is unfaithful in this same dispositional sense).

But what if commitment is understood in its psychological sense, as when we ask whether our groom really is committed to the marriage to which he has committed himself in speech? Then it seems to me that W might be mistaken in thinking that she is committed to the equi-dangerousness of black and white men. For if this proposition is not poised to direct her actions and deliberations, and she is not disposed to use it to guide and shape her actions, *even when she is aware of its relevance to what she is doing, and so able to bring it to bear on her actions and deliberations*, she does not believe it. W could not say to her black, male friend, "Well I believe that you are no more dangerous than my white friends in that I am committed to this proposition, even if I lack any disposition to incorporate this truth into my thoughts and behavior." For this would be to confuse the speech theoretic sense of "commitment" with the psychological sense that I have been trying to distinguish.

But what of Coliva's ambiguity thesis? If W takes pains not to react with more fear toward black men than white, if she masters her racist fears when deciding where to live, who to hire, and who to date, might this show that she believes the target proposition in the commitment sense of "believes," while nevertheless lacking it in some dispositional sense insofar as she might still be disposed to experience more fear toward black men than white? Though I do not like this way of thinking of the matter, I acknowledge that it is perfectly coherent. It goes along with the distinction between "conscious" and "unconscious" beliefs and the use of folk psychological concepts more generally to describe sub-personally generated perceptual and affective response. Alternatively, we might argue that "belief" is univocal and that a person who acts and reasons on a proposition really does believe it even if some of her reactions belie those beliefs. If she adopts this alternative understanding, W might say to her black, male friend, “I believe you that you’re not dangerous, I am certain of this, even if my startle response to your face is more pronounced than my startle response to white faces” (Amodio et al. 2003). According to this alternative taxonomy, which I favor, W can say this to her friend without equivocation, there being no sense of “belief” in which her being disposed to startle more at black faces than white entails that she believes that black people are more dangerous than white.

To my mind, both of these conceptualizations of belief are compatible with folk usage, intuition and science. Use of “belief” is loose enough that ambiguity can be posited or denied by adopting one of these philosophies of mind or the other. So the choice between these differing conceptual schemes is a pragmatic one that should be informed by an assessment of the differing consequences that would attend their adoption. But our freedom of choice on this issue should not blind us to the very real distinction between commitments as acts of speech and commitments as facts of mind. Nor should it obscure the substantive nature of our knowledge of our own commitments and the variety of cases in which this knowledge fails because we judge ourselves to be committed to the truth of a proposition we do not in any sense believe.

References

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink responses and self-report. *Journal of Personality and Social Psychology, 84*, 738–753.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Coliva, A. (2016). *The varieties of self-knowledge*. London: Pallgrave Macmillan.

Fodor, J. (1980). *The language of thought*. Cambridge: Harvard University Press.

Kripke, S. (1980). *Naming and Necessity*. Cambridge: Harvard University Press.

LaPorte, J. (1996). Chemical kind term reference and the discovery of essence. *Nous, 30*, 112–132.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Shoemaker, S. (1996). *The first-person perspective and other essays*. Cambridge: Cambridge University Press.

Zimmerman, A. (2018). *Belief: A Pragmatic Picture*. Oxford: Oxford University Press.