Review of John M. Doris, Talking to Our Selves.

Elisabeth Pacherie

Doris, John M. *Talking to Our Selves*
Oxford: Oxford University Press. Pp. 264, 2015.

Elisabeth Pacherie
Institut Jean-Nicod (ENS/EHESS/CNRS, PSL Research University)

Draft version. For purposes of quotation please consult the published version: Pacherie, E. (2017). Review of John M. Doris, Talking to our selves. *Ethics*, 127, 3: 772-777. https://doi.org/10.1086/690076

The topic of *Talking to Our Selves* is human agency and moral responsibility. John Doris argues that what the sciences of the mind tell us about what drives human conduct and how humans function as agents calls into question important strands of philosophical theorizing about human agency and moral responsibility. Doris does not advocate blanket skepticism about agency and moral responsibility. His view is not that empirical research shows that the practice of treating people as morally responsible agents is fundamentally misguided. Rather than on the practice itself, this research cast doubt on the main theoretical justifications philosophers have proposed in support of this practice.

The main target of Doris's criticism is reflectivism, a standard philosophical approach to agency that characterizes it in terms of deliberation informed by accurate self-awareness. The picture of human cognition and behavior emerging from decades of research in psychology suggests, however, that we lack accurate knowledge of what we do and why. We are thus faced with two options: either we retain the reflectivist conception of human agency and are forced to conclude that the exercise of human agency is the exception rather than the rule, or we jettison reflectivism and propose an alternative theory of human agency that provides more solid foundations for our practice of treating people as morally responsible agents. Doris goes for the latter option and develops a theory of human agency he characterizes as anti-reflectivist, valuational, collaborativist and pluralist.

The book is organized into two parts, each with four chapters. The first part offers a detailed argument that reflectivism, as a theory of human agency, lacks the resources needed to deflect the skeptical challenges it faces. The second part develops an alternative theory of human agency meant to answer these skeptical challenges.

I examine the two parts in turn, saying, in brief, what each of the chapters is about and then considering some objections one might have to Doris's arguments.

In chapter 1, Doris sets the stage for his enquiry into moral responsibility. Empirical investigations into the workings of the mind suggest that rather than as a unified system the mind should be seen as motley collection of at best only loosely integrated systems and subsystems. This empirical picture of psychological anarchy makes three vexing problems of self central to moral psychology – identity, continuity, and agency – appear even more intractable. If the disintegration of the mind brings with it a disintegration of the self, then skepticism about moral responsibility looms: if agents as unified loci of moral praise and blame vanish, moral responsibility vanishes with them. Doris proposes to resist this skeptical temptation but to do so while embracing "bare-knuckle

naturalism". His brief is thus to offer an empirically adequate theory of human agency that answers the question: "under what circumstances, if any, are human beings able to function as morally responsible agents?" This is a tall order, but Doris gives himself a little breathing space by assuming that the question of moral responsibility is independent from the question of freedom: an account of moral responsibility need not require freedom from causal constraint.

Chapter 2 discusses reflectivist accounts of agency. The preoccupation with reflection is, Doris notes, a central feature of the Western philosophical tradition. The idea that reflection is what separates humans from non-human animals finds deep echoes in many areas of philosophy, ranging from epistemology, to political philosophy and moral psychology. Doris is first and foremost concerned with reflectivism about agency, at its most obvious in the Kantian tradition, but also common in other traditions. Approaches to moral responsibility generally assume that attribution of moral responsibility to someone for their behavior depends on whether this behavior is an exercise of agency. Hence, the tight connection between accounts of human agency and accounts of moral responsibility. It is also commonly held that a behavior is an exercise of agency when it is self-directed. The question then becomes, What does it take for behavior to be self-directed? The reflectivist answer to this question can be schematically characterized , Doris proposes, as the doctrine that "the exercise of human agency consists in judgment and behavior ordered by self-conscious reflection about what to think and do" (p. 19), a doctrine whose corollary is that "the exercise of human agency requires accurate self-reflection" (p. 19).

On a reflectivist account of agency as reflective self-directedness, agency can be imperiled either through failure of deliberation – the behavior is not the product of deliberation or is the product of invalid reasoning or through failure of self-awareness – the agent is mistaken about his motives for acting. Doris puts forward three claims. First, empirical evidence shows that both types of failures are very common, thus throwing doubt on reflectivism interpreted as the view that reflective self-reflection is the most commonly observed form of agency. Second, we usually lack sufficient warrant for attributing the exercise of reflective agency, including on occasions of practical importance, thus throwing doubt on an interpretation of reflectivism as the view that reflective self-direction is practically dominant. Finally, he claims that it is possible to develop an account where exercising agency does not require reflection, thus creating trouble for reflectivism understood as the conceptual thesis that reflection is necessary for the exercise of agency. While the arguments for these three claims are left to later chapters, Doris immediately proceed to fill the vacuum created by his rejection of reflectivism about agency by outlining an alternative, valuational account of self-direction, that locates the exercise of agency in the expression of an actor's values and where values are "associated with desires that exhibit some degree of strength, duration, ultimacy and non-fungibility, while playing a determinative-justificatory role in planning" (p. 28).

Chapters 3 and 4 spell out the empirical and epistemological challenges to reflectivism. In chapter 3, Doris presents a wealth of empirical evidence identifying influences on behavior that are both unconscious and unexpected. For instance, if you have an honor box system for coffee at the office, placing an image of eyes near the box may help you avoid bankruptcy, as this makes people much more likely to pay their contribution (the Watching Eye Effect). Another example it the Pronoun Effect: people tasked with circling

pronouns in a text featuring first personal person pronouns are more likely to identify with "collectivist" values when asked afterwards to complete a values questionnaire then people who tasked with circling pronouns in a version of the text where the plural pronouns have been replaced by singular pronouns. Implicit and egotism biases are also well-documented. What these various effects have in common is that they are supposed to involve unconscious, effortless processing rather than conscious, effortful processing of the kind implicated in reflective deliberation. Together with this partition of human cognition into automatic and analytic processes comes the possibility of what Doris calls incongruence, where the outputs of automatic processes conflict with the output of simultaneous analytic processes or would conflict with the output of analytic processes were they engaged. This forms the basis for Doris's skeptical challenge. Instances where the causes of behavior would not be recognized by an agent as reasons for that behavior, were she aware of them at the time of performance, constitute defeaters. When defeaters obtain, agency doesn't. Unless defeaters can be ruled out, we are on the road to skepticism about agency.

Doris considers two moves open to the reflectivists to try and deflect the skeptical challenge. The first move would involve challenging the force of the empirical evidence Doris musters. There are replication issues with a number of findings on priming or implicit effects. Scientific journals may be biased towards publishing studies showing the presence of surprising effects rather than studies showing their absence. The size of the surprising effects in question tends to be small and the effects themselves short-lived. Doris acknowledges all these issues, but takes it that the empirical evidence is so bountiful that even applying the most drastic criteria to screen experimental findings leaves us with effects that are together larger than reflectivist theories would wish. The second move open to reflectivism is what Doris calls the "triage" response, that allows that reflective-self direction may be relatively infrequent but insists that when the situation demands it, people reflect effectively. Doris' rejoinder is that situations where the stakes are high enough that people should reflect before acting are also typically emotionally charged and thus also situations in which unruly automatic processes may be at full play and taint reflection.

Chapter 4 considers the reply to the skeptical challenge that appeals to our experience of our agency as evidence for reflective agency. Doris uses the literature on confabulation to block this reply. He argues that two of the four main factors in clinical confabulation, deficient self-awareness and motivation, are also common, in milder form, in healthy subjects. We have little insight into the actual causes of our behaviors and we readily confabulate explanations for what we do. Furthermore, the explanations we come up with have a motivational basis, often reflecting a tendency towards self-enhancement. The conclusion then is that our experiences of agency are not the reliable guides to our actual agency they would need to be to support reflectivism.

Have the prospects of reflectivism been steamrolled into dust by Doris' skeptical argument and the mass of empirical evidence behind it? Reflectivists might want to welcome these empirical findings rather than looking at them with a jaundiced eye. If, like Socrates, they are much taken with the Delphic Maxim, "Know Thyself!", then, with him, they are probably well aware that self-knowledge does not come for free and that the first step on the path to self-knowledge is the acknowledgment of self-ignorance. The sciences of the mind tell us that this self-ignorance is deeper that previously thought and that unconscious influences on our behavior are multifarious. But a reflectivist might

think that to be forewarned is to be forearmed. Doris notes that healthy people do not suffer the deficiencies in self-control characteristic of clinical confabulation. A reflectivist might argue that when so forewarned, we can exercise self-control to keep unwanted influences at bay when deliberating about what to do. As a real life example, some years ago my research institution decided that people sitting on its hiring committees would be systematically briefed on implicit gender biases in science. This policy was aimed at reducing the gender gap observed in certain disciplines and appears to have met with some success. Thus, reflectivists may take the empirical evidence mustered by Doris not as a cause for despair but as an opportunity to become better deliberators.

I now turn to Doris' positive account. Doris proposes a valuational approach to morally responsible agency, which locates the exercise of agency in the expression of an actor's value. His main purpose in the second part of the book is to show that this account of morally responsible agency is well equipped to meet the skeptical challenges that spell the doom of reflectivism.

In chapter 5, Doris rejects another hallmark of the Western philosophical tradition, individualism, defending instead collaborativism. Collaborativism maintains that optimal human reasoning is substantially social: people reason best when they reason in interaction with others rather than in isolation. Thus, substantial cognitive achievements, such as sciences and technology but also moral reasoning, are socially embedded. The first step in his positive argument is that collaborativism is not limited to reasoning but extends to agency. Sociality facilitates agency. For instance, dialogic interactions, such as talk therapies, can facilitate the expression of an agent's values.

The second step of his argument, developed in chapter 6, is that self-ignorance, together with sociality, can promote agency rather than hinder it. First, self-ignorance, when it takes the form of positive illusions of self can motivate us to act in ways that enhance agency. For instance, my believing that I am a good athlete when I am not may motivate me to train hard and indeed become a good athlete. Second, agency may be facilitated by self-ignorance and confabulation in tandem with collaboration. Absent accurate awareness of what makes us behave as we do, the exchange and negotiation of rationalizations can help us make sense of our behavior. Finally, these rationalizations provide the material for the self-narratives we construct in interaction with others to interpret our behavior and these self-narratives shape in turn our behavior. Because people are motivated to reduce cognitive dissonance between attitudes and behavior, self-narratives and behavior will tend to converge, thus insuring that behavior expresses the values endorsed by the self-narratives. This dialogic conception of agency forms the core of Doris's positive account.

Two chapters on responsibility and self conclude the book. Chapter 7 argues has the resources needed to address skeptical concerns about responsibility and can ground a normatively robust account of responsibility. While Doris recommends pluralism about agency and responsibility, and concedes that reflectivism way be warranted for some attributions of agency and responsibility he maintains that a dialogic conception of agency makes better sense of most attributions. Chapter 8 goes back to the problems of the self identified in chapter 1 and argues that the dialogic approach can be extended to continuity and identity.

Value and collaboration are the two pillars on which Doris builds his theory of agency. Let me briefly consider some concerns one might have regarding the solidity of these pillars. Doris extolls the virtues of collaboration, but collaboration may not be an unmixed blessing. While our scientific achievements may be taken as evidence that collaboration facilitates optimal reasoning, social psychology also has a rich store of counter-examples. The phenomenon known as groupthink, thought to have led to such fiascos as Pearl Harbor, the Bay of Pig Invasion or more recently the Invasion of Iraq, is a case in point. Rather than reducing biases, collaboration may exacerbate them when the desire for harmony and cohesion in the group leads group members to minimize conflict and suppress dissenting viewpoints. The success of science may be explained in part by the fact that in the scientific arena collaboration is peppered with a good dose of competition. Even if we agree with Doris that the exercise of agency is a substantially interpersonal phenomenon, it is unclear whether social interaction systematically facilitates optimal agency. Dialogic rationalizations of behavior may have more in common with the dynamics of groupthink than with the dynamics of science. Doris may think this is of little import. Even if these rationalizations are little more than social confabulations, once incorporated in our self-narratives, they start shaping our behavior and become expressions of our values.

This takes me to the second pillar of Doris theory, values. Doris claims that behavior is an exercise of agency when it is self-directed rather than having external sources and that self-directed behavior is behavior expressive of the actor's values. However, one can regret that he doesn't spell out in more detail what values are, what makes behavior expressive of values and how we can tell whether behavior is indeed self-directed as opposed to externally directed. In particular, it isn't clear to me that his approach has the resources to identify instances of self-directed behavior with the confidence needed to rule out defeaters and dispel skeptical challenges. Since, on the approach, self-directed agency is compatible with self-ignorance, the subjective reports and rationalizations of actors cannot be taken at face value and cannot form the basis for attributions of self-directed agency. Doris's answer to this epistemological problem is that we should consider extended behavioral processes rather that isolated behaviors are that attributions of agency are warranted when patterns of behavior and rationalizations emerge and are best explained a involving the expression of some value. But one may wonder whether this approach really gives us sufficient warrant for agency attributions. Suppose we could travel back in time and observe the life or ordinary Soviet citizens in the 1950's. We would certainly find strong trends in their behaviors and rationalizations, and, following Doris's method of attribution, would seem to be warranted in attributing them self-directed agency expressive of communist values. But is it really the case? Some Soviet citizens were no doubt bona fide communists; others might have been mere opportunists doing and saying what it took to become part of the soviet elite, and fear of the gulag was probably the main force driving the behavior of quite a few. Given that in all three cases the observable patterns of behavior would have been the same, we cannot with confidence attribute to them self-directed agency expressive of communist values, we cannot even attribute to them self-directive agency, whether expressive of communist values or not. Rather than being self-directed, their behavior could well have had its source in features of their oppressive social environment. When the soviet composer Dmitri Shostakovich was asked why he had such a passion for football, his answer was that the stadium was the only place where he could be himself and openly express his emotions. Of course some of us are lucky

enough not to live under totalitarian regimes. Even in more open societies, however, social pressure to conform remains present even if in milder forms and creates smaller-scale versions of the "Soviet attribution problem". It is not so clear that from temporally extended patterns of behavior we can infer with assurance whether behavior is self-directed and if so what values it expresses.

Talking to Our Selves raises a series of fascinating challenges to the defenders of reflectivism and makes a strong case that philosophers should come to grips with the wealth of psychological findings relevant to their inquiries. The positive account Doris proposes shows that taking these findings seriously can inspire bold, new views of human agency and moral responsibility. This well argued, thought-provoking and, too rare a quality in philosophy writing, very entertaining book makes for a very stimulating and enjoyable read.