and mutual endorsement of the standards. In the face of deviations from standards, people have to express condemnation in order to make the violated standard salient to all group members (Feinberg 1965; Durkheim 1893). In addition, the expression of blame for norm violations demonstrates that group members care about the norms and the group members protected by those norms. Finally, observers blame norm violators to distance themselves from the deed and avoid being associated with such misdeeds. Thus, in some sense the observers show agency when they blame and praise others' behavior because it expresses their values (usually, shared values). They may even express their values without caring too much for actual responsibility of the actors (i.e., they may not go further than differentiating between coerced and uncoerced behavior).

Moreover, we argue that the assignment of blame or praise for misdeeds also affects the actors' agency. Public condemnation indicates, claims, or even fosters group members' exercise of agency. As observers attribute responsibility to the actors, the actors may also perceive themselves as having agency (or an illusion of agency?). For example, children's agency develops by the guidance of sanctions. Agency may be considered an *ability* (that one could learn) instead of a *habit*. Habits denote what people are accustomed to do, whereas abilities include a normative component that denotes what would count as a correct or incorrect thing to do (Millikan 2000). This normative component specifies when we sometimes succeed in expressing our values and when we fail to express them. As mentioned above, praise and blame direct us thereby in the standard's (valued) direction. In contrast, habits could go in any direction, as they are not necessarily corrected by values. Moreover, by such development of ability over time (i.e., agency-training), we become more reliable in expressing our values in particular situations and apply them to more diverse situations.

As an additional mechanism, we suggest that reminders of our responsibility, such as blaming and praising of certain behaviors, activate the concept of personal agency. Activated concepts also tend to produce concept-related behavior (e.g., the belief that one excels in math enhances math performance, Miller et al. 1975). Activated concepts also change cognitive processing characteristics that lead to the enactment of these concepts (Sassenberg et al. 2017). Accordingly, actors who are held responsible may activate their concept of "being responsible." Thus, before acting, they may think twice, activate their main values, and take precautions to make sure that their behavior conforms to their values. Such a reflection of personal values in turn may lead to a stronger connection of these standards in their cognitive system; they may identify with them and thereby behave more in accordance to them. This is also a social process: it not only involves solitary thinking but also social negotiation and training in justifying behavior in the face of others. This may reward careful action, so that people may arrange their environment in order to avoid known "defeaters" (e.g., temptations). Moreover, being held responsible indicates being watched. This enhances objective self-awareness and thereby a person's own standards become more salient.

The social shaping of agency and responsibility may not always work out completely. Some people may be hard to train or unwilling to develop stable "virtues" (i.e., habits to act according to their own and commonly shared standards). However, this may be irrelevant, as others will still hold them responsible (even if this cannot apply literally) and punish them (e.g., go for incapacitation as a last resort). In addition, people may not want to wait until repeated misdeeds manifest the "negative" values of the actor. There may be an asymmetry in that many positive deeds are necessary to manifest positive values of people, whereas one negative deed can be enough to reveal the negative value of an actor. The extremity of the deed may itself be a clear indicator for moral responsibility (Pauer-Studer & Velleman 2011). In such cases, where the social shaping of individual agency or responsibility may be impossible or come too late, the actor can only be made incapable. However, the general practice of collaboratively shaping agency may not be threatened by this because these examples remain exceptions.

In short, the emergence of agency and responsibility is a social process. Talking to others (including blaming and praising) is a particularly efficient way to develop one's own agency and help others become responsible actors.

Grounding responsibility in something (more) solid

doi:10.1017/S0140525X17000711, e47

William Hirstein and Katrina Sifferd

Department of Philosophy, Elmhurst College, Elmhurst, IL 60126.
williamh@elmhurst.edu sifferdk@elmhurst.edu

Abstract: The cases that Doris chronicles of confabulation are similar to perceptual illusions in that, while they show the interstices of our perceptual or cognitive system, they fail to establish that our everyday perception or cognition is not for the most part correct. Doris's account in general lacks the resources to make synchronic assessments of responsibility, partially because it fails to make use of knowledge now available to us about what is happening in the brains of agents.

Our commentary on Doris's significant book focuses on three areas: (1) Doris's claim that cases of self-ignorance, such as confabulation, are common enough to negate our own judgments of why we did things; (2) Doris's inability to give a good account of synchronic assessments of responsibility; and (3) the disconnect between Doris's account and scientific accounts of human thought and behavior.

Self-ignorance. Doris says that human beings are "afflicted with a remarkable degree of self-ignorance" (précis abstract). But while we certainly at times show self-ignorance, there is no absolute metric that allows us to assess the exact degree of our ignorance compared to our self-knowledge. This opens the possibility for researchers, who feed on a steady diet of examples of ignorance, to overestimate its degree. We need to leave open, for example, the possibility that we are dealing not with phenomena that afflict everyone, but with phenomena that only afflict a minority of people, or even a certain personality type. The scope of Doris's skepticism is also broader than it might appear. One sign that we might be overestimating the amounts of ignorance and error is that we have not been moved to enact major changes in folk-psychology to remove dependence on our capacity for self-knowledge. Doris's view seems to commit us not only to being "routinely mistaken" (précis abstract), but also not ever noticing that we are, and attempting to correct it. Doris seems to be neglecting all those times we *aren't* buffoons.

A comparison with the case of visual perception is illuminating. Even though cognitive scientists have cataloged perhaps hundreds of visual illusions that reveal the seams and flaws of our visual system, the vast majority of our visual perceptions during the day are veridical and serve us quite well. Consider our abilities to visually identify one another. Certainly there are many ways in which the brain systems that achieve this miracle can fail, leading to odd syndromes like prosopagnosia. In the everyday sphere, we have all experienced cases in which we visually misidentified someone. But taken against the overwhelming percentage of correct identifications we make so effortlessly and frequently, these misperceptions are rare. This high rate of effectiveness is due to good equipment.

We think serious cases of ignorance or mistaken self-knowledge are somewhat rare because they typically involve errors at two levels. First, a mistaken impression is created. For instance, it occurs to me that I don't really have to pay back that loan from my friend because he seems to be wealthy, when I would just prefer to keep the money. Then, this error is not corrected (this correction could occur because I note my obligation to repay, or I revise my sense of my friend's situation, or I just realize I am being selfish). The first type of error, where I form a mistaken impression of my own motives, is fairly common; the second, where I fail to correct, or at least *where I fail to correct because I cannot correct*, less so. And in cases where we have the capacity to correct for our mistaken perceptions, using our brain's prefrontal executive processes, it would seem we are responsible for them (Hirstein et al. 2018). For example, a color-blind person can correct for his problem by memorizing the location of the traffic lights. Doris's view amounts to saying that the entire upper level that has been designed into our brains, including the executive processes and consciousness itself, is of little use or import. This level functions precisely to correct basic errors of perception or memory, as can be seen in the case of confabulation (Hirstein 2005). This second level tends to only activate when the stakes are appropriately high, so that examples of our failures where they aren't perceived to be high, such as the case of people failing to put money into the office coffee fund, are not showing our cognitive system at its best.

Synchronic assessments of responsibility. Doris argues that moral responsibility for an act depends upon whether the act in question was an exercise of agency (Doris 2015b, p. 159). Exercises of agency, according to Doris, are expressions of the actor's values; attributions of responsibility turn on whether an actor's values are expressed in an act (p. 159). However, this sort of view faces clear epistemological difficulties, as Doris notes: It will frequently be difficult to determine whether someone holds a value, and actions often seem related to multiple values, some of which may be unknown even to the actor. Plus, "values are expressed over time, and can, oftentimes, only be identified over time," and thus "extended observations" may be required to identify patterns to determine if any particular action is of the sort for which an agent can be held responsible. "If one focuses on isolated events, diagnoses may falter" (Doris 2015b, p. 162). In the end, attribution of responsibility may require first that "a pattern of cognition, rationalization, and behavior emerges, and that pattern is best explained as involving the expression of some value"; and second, a determination that a particular action expresses that value (p. 164). But why in a revolutionary era of neuroscience assume that we must remain forever locked outside the mind and brain of the subject? Doris's account involves the cognitive sciences, but only those that focus on behavior and outward from there, to society. We suggest that connecting his knowledge of the psychological research with neuroscience, via cognitive neuropsychology, would greatly help resolve the epistemic problems involved in discerning what exactly someone's values are.

As it stands, Doris's theory indicates that synchronic assessments of responsibility are often impossible. However, the most common and important responsibility attributions are synchronic. Take, for example, criminal verdicts. Judges and juries do not, and ought not in most cases, focus on past behavior as a means to indicate responsibility for a particular crime.[1] A criminal court is asked to determine whether a defendant held a particular mental state and whether this mental state is causally related to the criminal harm. Such canonical cases of responsibility attribution are considered so secure we use them to deny defendants' liberty and even life. If Doris's theory is correct, and responsibility assessments rest on extended investigations into a person's values, then it would seem our current system of generating verdicts and punishing offenders is likely to attribute responsibility to persons when it has not been proven they deserve blame.

Doris indicates that he is a pluralist about responsibility, and thus "sympathetic" to the possibility that there may sometimes be warranted attributions of other types of responsible agency, including reflectivist agency (Doris 2015b, p. 174). However, he also feels that a pluralistic account must place dialogic agency in an "appropriately prominent" position (p. 175). To vindicate the thrust of his theory with regard to criminal verdicts, Doris should provide an account of how a synchronic act must be related to dialogic agency. Further, this account must explain how a synchronic act can be seen as an expression of such agency without an exhaustive review of the agent's history. But if this were possible, then it would seem that Doris's requirement of "extended observations" would, in most cases, be unnecessary because a less burdensome, synchronic assessment would suffice.

In a similar vein, the reactive attitudes, which Doris acknowledges are important first indicators of responsible agency (Doris 2015b, pp. 23–24), are typically generated in synchronic cases without information about character. They depend on the brain's mindreading (or theory of mind) capacities, through which we attribute motives behind a person's actions, sometimes using fairly few behavioral cues. If these motives are selfish, for example, a strong negative reactive attitude will follow. In the criminal law, we feel stronger condemnation where an agent directly desired criminal harm (committed the act "purposely" under the U.S. Model Penal Code) than in cases where an agent merely ought to have known there was a risk of substantial harm (committed the act "recklessly").

It isn't clear that Doris's weakly proposed pluralism, which encompasses his dialogic view and reflectivism (Doris 2015b, p. 174), can generate many of the synchronic responsibility assessments made in the criminal law. As Doris argues, many culpable actions do not seem connected in the right way to reflective judgments, which are often confabulated. Thus, if extensive investigation of dialogic agency is not done, on what grounds are criminal verdicts generated? For example, in a case where the fire was due to the building owner's forgetting to check the functioning of the water sprinklers, a synchronic assessment of the defendant's conscious mental states with regard to the criminal harm will not secure a responsibility assessment. In our view, only an account that provides a synchronic assessment of capacity for responsible agency, where that capacity is more expansive than just the capacity for conscious reflection, can ground criminal verdicts of negligence.

Personal versus subpersonal. As we noted, Doris chooses to keep his analysis at the personal, rather than the subpersonal level, by using information largely from social psychology. But sometimes, simple knowledge of the person's brain can clear things up. For example, Doris notes that "the valuational account says if your action properly expresses your values, it's an exercise of agency, regardless of whence your values came" (Doris 2015b, p. 30). But what about someone with Tourette's whose outbursts do express his values, but not in a way he wanted? Or a person whose sleepwalking actions do express his values, but are horrible, and which he would never do when awake? In both of these cases, responsibility does not seem to rest with the actor, due to volitional incapacity, despite the alignment of the action with the actor's values. If we could "see" the actors lack of control via evidence of brain function (or dysfunction), we might correct mistaken assessments of responsibility.

Doris searches everywhere for help in attributing psychological states such as motives, including other people (the dialogic part of this theory), except in neuroscience. There is useful information at the subpersonal level, from neuroscience, cognitive neuropsychology, and from historical neurology, that is vital to gaining a full understanding of the relevant phenomena. Neuroscience can provide valuable data regarding synchronic assessments of responsibility. For instance, it might be able to tell whether an action is "done habitually" (i.e., what the neuroscientists call an action done "in routine mode") or done as a result of conscious reflection, which involves quite different and more extensive brain processes. While Doris avows materialism, it is difficult to see how his theory, as stated, can be put into stark, materialistic terms. What concrete things, states, processes, and events do claims about "values," "desires," "plans," "self-awareness," and "the exercise of agency" refer to? We are not done with the project of building a theory of responsibility until we can do that.

NOTE

[1] Federal Rule of Evidence 404(b)(1) states that "Evidence of a crime, wrong, or other act is not admissible to prove a person's character in order to show that on a particular occasion the person acted in accordance with the character."

Getting by with a little help from our friends

doi:10.1017/S0140525X17000723, e48

Enoch Lambert and Daniel C. Dennett
Center for Cognitive Studies, Tufts University, Medford, MA 02155.
Enoch.lambert@gmail.com Daniel.dennett@tufts.edu
http://ase.tufts.edu/cogstud/dennett/

Abstract: We offer two kinds of constructive criticism in the spirit of support for Doris's socially scaffolded pluralism regarding agency. First: The skeptical force of potential "goofy influences" is not as straightforward as Doris argues. Second: Doris's positive theory must address more goofy influences due to social processes that appear to fall under his criteria for agency-promoting practices. Finally, we highlight "arms race" phenomena in Doris's social dynamics that invite closer attention in further development of his theory.

Doris conducts a master class for psychologists on how to extract value from the philosophical debates, and for philosophers on how to use empirical work in psychology to inform their theorizing. In both endeavors, one has to learn how to take the declarations with more than a few grains of salt, which Doris applies judiciously. We heartily endorse what we take to be a major lesson: What we learn from science, while sometimes shocking, need not destroy our confidence in our own practical agency. Rather, by informing our understanding of our agential strengths and weaknesses, science can guide us in discovering and strengthening those practices that foster our agential powers. Of special note is his case that self-ignorance can be crucial to our projects of building and expressing our central values, showing how accurate reflection can actually undermine agency in some situations. He has also done the study of practical reason a great service by setting up a framework for exploring its socially scaffolded nature. In our comment, we aim to contribute to that ongoing project. While we believe Doris is right about the largely social nature of agency, we raise some questions about the skeptical force of the psychology he cites against the role of accurate self-knowledge in our deliberations. We also urge that his own "collaborative-negotiative-dialogical" framework faces significant threats from social psychology – more so than acknowledged.

Doris's critique. First, we wish to question the strength of the case Doris mounts for *global* skepticism regarding the role of accurate self-knowledge in our deliberations. We are more concerned about the size of experimental effects and their implications for everyday decision making than Doris is. It is instructive to recall the reason why so much psychology focuses on surprising effects. Vast swaths of common wisdom concerning self-knowledge prevent psychologists from so much as attempting to confirm things like whether people tend to be accurate about whether they prefer $1,000 to a pin prick, or social praise to ridicule. Finding a new way of generating small, surprising effects may be rewarded in psychology, but it is not clear whether or how the common lore of everyday psychology that psychologists never bother to investigate is undermined by it.

Doris (2015b) dismisses the importance of statistically small sizes partly by saying that known "goofy influences" on behavior indicate an ocean of unknown ones; and partly by saying that such influences may "aggregate" in ways that medical interventions can (p. 64). Our own speculative mechanics of goofy influences suggest a different lesson. If "eyespots and pronouns are in the mix" (to use Doris's nice phrasing), then humans are likely assailed by goofy influences *continuously* (p. 64). The priming and automaticity literatures from across psychology suggest no principles for ruling out much of anything as potentially goofy influence. But if this is so, how do we manage to hold it all together? Why are we not driven every which way by the onslaught of disparate priming stimuli? And how are we able to come by the amount of common knowledge of human psychology that we do? Why can we predict so well what others will do based on "typical" perceptions and desires (which we also attribute to ourselves)? When predicting what the drivers of other cars on the road will do, we justifiably pay no attention to which images on which billboards they recently saw, or the content of the radio advertisement they are hearing, or whether their vehicle interior is leather, or. . . . It isn't that we are in a position to rule out such things ever having some influence on how they drive, whether at a micro-level, or, on occasion, at a life-altering level. But our attributions are sufficiently reliable enough of the time so that it makes no sense to let such influences trigger general skepticism of our usual interpretive and predictive capacities. Similar considerations apply to our own case. It would be silly, for instance, to decide to live as close as possible to the market simply because it would minimize the amount of goofy influence encountered every time we need to do our shopping.

Moreover, Doris ignores the prospect of a gradient between goofy and not-so-goofy, to go along with his valuable gradient between explicit self-reflection and the sort of automatic self-monitoring that gets us relatively gracefully through the day. The fact that pictures of watchful eyes should nudge more honest coffee transactions is striking, but not so striking or upsetting as the non-fact – we wager – that pictures of bicycles or rooftops have the same effect. Doris's richly detailed account of actual decision making suggests that in the real-time hasty triage involved in all but the most portentous moral decisions, a "subliminal" hint about being observed and caught could be just enough to bias the choices made without the choosers' noticing.

Next consider one of the roughly third of test-subjects who detected the switches in the moral choice blindness experiment Doris cites (2015b, p. 139; see Hall et al. 2012). What should such a subject conclude upon learning the results of the experiment? That she got lucky? Why would that be more reasonable than to conclude that, for whatever reason, she was more attentive (or cared more, or . . .)? *Perhaps* she should conclude that her capacity to recognize her own moral positions is more susceptible to error than she would have thought, and so she should keep watch. But it doesn't seem reasonable to conclude that she should be an outright skeptic of her ability to recognize her own morality. And, in general, we urge that individual variation in susceptibility to goofy influences not be swept aside as so much noise. Why is it that goofy influences do not affect some subjects in any given experiment? Are some people who are less susceptible in specific experiments more generally resistant to goofy influences? If so, why? Can any pattern at all be detected in failure to succumb to goofy influence? It seems that such possibilities remain live empirical hypotheses to be ruled out (or in!) rather than assumed. Until we know more about the mechanics of goofy influences, it seems rash to let them *completely* undermine the role of accurate reflection in our deliberative decision making.

Doris's positive framework. Given Doris's conservatism about our everyday attributions of agency and responsibility, it is surprising that he uses psychotherapy as a model for how collaboration and dialogue can facilitate agency. In the history of agential responsibility, psychotherapy has been around for a blink of an eye, and has been employed by a sliver of agents. So it is at best a device for highlighting what aspects of our common practices actually do facilitate agency. Dialogue and "positive alliance" are the agency-facilitating aspects of beneficial psychotherapy highlighted by Doris. But both phenomena are also present in collaborative enterprises where anti-agential forces often prevail. We review some below, but we encourage Doris to say more about what lessons to take from psychotherapy, as well as