Journal of the American Philosophical Association (2015) 601–616 © American Philosophical Association
DOI: 10.1017/apa.2015.24

Moral Responsibility Ain't Just in the Head

ABSTRACT: *In this paper, I dispute what I call* psychological internalism *about moral responsibility, which comprises most classic accounts as well as newer neurobiological ones, and I defend* psychological externalism *about moral responsibility instead. According to psychological internalism, an agent's moral responsibility is determined solely or primarily by her intentional states. I argue that psychological internalism is empirically challenged by recent findings in social psychology and cognitive science. In light of the empirical evidence, I contend that moral responsibility depends on historical and environmental factors to a much greater degree than previously appreciated. Thus, moral responsibility is not just in the head: indeed, it is much less in the head than typically assumed.*

KEYWORDS: moral responsibility, psychological internalism, psychological externalism, moral psychology, implicit bias

1. Introduction

The standard view is that moral responsibility is determined solely or primarily by an agent's current intentional states, such as her beliefs, desires, and values. I call this view *psychological internalism* about moral responsibility (internalism, for short). The opposite view is that an agent's moral responsibility is determined to an important degree by factors other than her current intentional states. I call this view *psychological externalism* about moral responsibility (externalism, for short). From now on, I will use 'responsibility' as shorthand for 'moral responsibility'. I do not intend to discuss other types of responsibility, such as legal, personal, and fiduciary.

Some theorists recognize that factors external to an agent's intentional states may affect her degree of responsibility. For instance, many acknowledge that recent historical events, such as clandestine manipulation by a nefarious neurosurgeon (Fischer 2006), can undermine responsibility. This is a step in the right direction, but too small to break with internalism. The resulting view is still a version of internalism, albeit an impure one.

Internalism and externalism, properly understood, lie on a spectrum. On one end of the spectrum is pure internalism, according to which responsibility is determined solely by the current intentional states of the agent, while external factors play no

This research was supported by the Social Sciences and Humanities Research Council of Canada. I am grateful to John Doris, Ron Mallon, and the anonymous referees for commenting on earlier drafts of this paper.

role. On the other end of the spectrum is extreme externalism, according to which responsibility is determined solely by external factors, while current intentional states play no role. Both ends of the spectrum are implausible. What remains to be seen is where the correct account of moral responsibility is to be found along this continuum.

I will argue that existing accounts of responsibility lie too close to the pure internalist end of the spectrum to be credible. Specifically, I will defend three theses about moral responsibility that undercut strong externalism. First, an agent's earliest history (in addition to her recent history) may affect her degree of responsibility. Second, an agent's environment (in addition to her history) may affect her degree of moral responsibility. Third, external factors may be inculpatory as well as exculpatory. The result of accepting these theses is an account that is so far from pure internalism that it deserves to be called externalism. It may be helpful to view internalism as a species of fundamental attribution error (Harman 1999), in which internal factors are seen as foregrounding responsibility to the neglect of morally relevant external factors. Externalism corrects this bias.

To be more precise about my plans, in section 2 I adduce some prominent examples of externalism from moral philosophy, and in sections 3 and 4 I argue that these accounts insufficiently appreciate the significance of distal historical factors and current environmental factors for moral responsibility. I make this case using a combination of conceptual arguments (viz., doppelganger examples) and empirical research in developmental and social psychology. In section 5 I recruit these arguments to dispute the popular belief that we cannot be responsible for implicit bias. In section 6 I use psychological research to show that, independent of the foregoing arguments, pervasive epistemic opacity and cognitive distortions undermine internal requirements of moral responsibility. Finally, in section 7 I respond to skepticism about the practical tenability of externalism by showing that, in light of the preceding discussion, there is every reason to think that external markers of responsibility are more epistemically tractable than internal markers.

Before proceeding with these arguments, it is worth noting two assumptions that I make in this paper: (1) that moral responsibility comes in degrees and (2) that psychosocial development is influenced by environment. These are, I believe, fairly uncontroversial clams, but they are seldom discussed in moral philosophy, with some notable exceptions (see Faraci and Shoemaker [2010] on degrees of responsibility, and Watson [1987] on formative circumstances). The ensuing arguments will help reinforce these claims by providing comparative examples of responsibility.

I have not yet explained what I mean by 'responsibility', which I take to be an intuitive concept. If pressed, I would describe it in Strawsonian terms (1963) as deployment of the reactive attitudes of resentment, approbation, disapprobation, indignation, and the like. Since this account is compatible with multiple underlying deployment conditions, it illustrates what responsibility consists in without begging the question.

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

2. Psychological Internalism

Historically, moral responsibility has been regarded as an *intra*personal capacity or mechanism, described in intentional terms (via beliefs, desires, or values). The strongest examples of this approach are structural (also known as 'deep-self') views such as Frankfurt's and Watson's. Briefly, Frankfurt (1969, 1971) defines responsibility in terms of the capacity to endorse one's first-order desires decisively, and Watson (1987) defines it as the ability to form evaluative judgments. These capacities are strictly internal to the agent.

Fischer and Wolf present impure internalist views that include internal and external criteria, but the role of external factors is very limited, and the relevance of these factors is insufficiently explained. Fischer (2006) describes responsibility as the capacity for 'guidance control', that is, as moderate reasons-responsiveness, which can be undermined by past coercion such as brainwashing. Wolf (1987) characterizes responsibility as, in effect, reasons-responsiveness plus 'moral sanity', that is, the ability to notice and appreciate moral facts. Sanity can be compromised by such things as deprived childhood circumstances. These external criteria seem reasonable, but they are insufficiently emphasized, fleshed out, or explained. In particular, their relation to the posited internal criteria (reasons-responsiveness and sanity) is not elucidated, making them seem ad hoc. Moreover, these views to do not encompass the whole panoply of relevant moral factors, as they overemphasize excusing conditions. In particular, they seem incapable of attributing responsibility for the culpable acquisition of defects of *moral competency*, that is, of defects in reasons-responsiveness or normative sensitivity. Yet, there seems to be a tangible difference between culpable moral incompetency acquired by negligence, and nonculpable moral incompetency incurred by unavoidable tragedy and tribulations. These views suggest that the buck stops at internal mechanisms: no antecedent analysis is necessary.

If I am right, then internal criteria are neither necessary nor sufficient for responsibility. They are not necessary because someone who lacks relevant internal mechanisms can be responsible for this deficit in cases of culpable negligence, and they are not sufficient because someone who possesses relevant internal mechanisms can nonetheless be excused of responsibility in cases of coercion, duress, and necessity. These familiar inculpating and exculpating conditions will be investigated in the next two sections.

In recent years, Neil Levy (2014: 15) has made a valuable contribution to moral philosophy by developing a neurobiological account of moral responsibility, which defines it as a function of consciousness. He describes consciousness as the capacity to broadcast informational content to a wide variety of consuming systems through the *global neuronal workspace* (GNW), as indexed by self-report. (The neurobiological details are unimportant for our purposes—I will grant them for the sake of argument). This view, which he calls 'the consciousness thesis' (hereinafter, TCT), entails that we can only be responsible for mental states of which we are conscious. This approach adds neurobiological detail to classic accounts of responsibility without significantly altering them.

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

I take Levy's view as embodied in the consciousness thesis to be a modern version of structuralism because it ties responsibility to consciousness and internal coherence. (Elswhere Levy discusses externalist criteria [2008], but the straightforward reading of TCT, independent of his other theoretical commitments, is a structuralist one.) On this view, responsibility can be determined in principle on purely internalist grounds, by reference to neurobiological states. Furthermore, the operation of the GNW ensures relative motivational coherence, which prompts Levy to define TCT as a form of deep-self view. This fits with a structuralist reading. To be fair, Levy mentions something called 'indirect control' twice in his monograph (2014: 3), but he never explains what this amounts to and never discusses external factors that might be relevant independent of their role in GNW processing. If I am right, then the *absence* of moral enabling factors in one's environment and the existence of *unconscious* deficits due to negligence may affect one's degree of responsibility though neither of these factors is susceptible of being processed in the GNW.

One salient aspect of Levy's view deserves special consideration. On TCT, one can never be responsible for implicit bias, since implicit biases are unconscious states on dual system theory, the dominant cognitive model of judgment and decision making (Kahneman 2015). Indeed, the same conclusion seems to follow for Fischer and Wolf, insofar as implicit biases—which are unconscious, automatic, and incongruous with one's explicitly avowed beliefs—are not reasons-responsive. Levy endorses the same interpretation, describing TCT as a neurocognitive gloss on Fischerian guidance control. Since Wolf's view embeds reasons-responsiveness as a necessary condition, she is also saddled with this implication. It is reasonable, therefore, to assume that Levy, Fischer, and Wolf are all committed in principle to the view that implicit biases are exempt from responsibility ascription.

In what follows, I will challenge this by showing that historical and environmental factors can impute responsibility to moral failings, including implicit biases, in case these moral failings stem from culpable negligence. This impugns the necessity of internal criteria for responsibility. While other philosophers have discussed the importance of defeaters (or exculpating conditions) for responsibility (e.g., Doris 2015), few have discussed *inculpating* conditions, such as neglect of available moral resources. These conditions are externalistic since they need not be available to consciousness or reason-responsiveness or other internal mechanisms. Rather, people can be unintentionally and unconsciously negligent due to a long history of moral indifference.

3. Historical Factors

In *My Way* (2006), Fischer constructs a thought experiment to show that moral responsibility is affected by the recent history of an agent. This implies that responsibility is not just in the head. I intend to expand on this argument by showing that moral responsibility is even *less* in the head than Fischer supposes—specifically, it is affected by a longer sequence of historical events, and a broader range of causal factors, than he envisions. I make this case using a combination of conceptual and empirical arguments.

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

First, let's consider Fischer's example. Imagine two psychological doppelgangers, Glum and Plum, both of whom want to commit a murder. Glum, however, has been secretly manipulated by a nefarious neurosurgeon to possess this desire, whereas Plum has undergone a normal course of psychological development. Subsequent to the neuroscientific intervention, Glum and Plum both commit a murder, acting on their current motives, but only Glum's motivational profile has been manipulated. This leads Fischer to pronounce that, 'whereas Plum is morally responsible for killing [the victim], Glum is not. Plum acts from his own, moderately reasons-responsive mechanism, but Glum does not' (2006: 235). Fischer claims that Glum is not responsible, full stop. One might rather think, on a spectrum conception of responsibility (from very responsible to not responsible), that Glum is not *as* responsible as Plum, because he did not autonomously generate his murderous intention. The example supports this inference (implicitly) if we assume that the evil neurosurgeon inherits a majority of the blame for Glum's action. This would explain why Glum's responsibility is diminished compared to Plum's.

Now, we can trace mitigating factors *even further back* than recent psychological manipulation, and we do not need to appeal to science fiction to make the case for historicism even at this level of remoteness. There is now a plethora of empirical research showing that childhood abuse and deprivation reliably generate deficits in moral cognition, which are reasonably construed as constitutive of the capacity for moral responsibility. (This capacity is surely complex and multifactorial, but psychosocial competencies are at its core). Hildyard and Wolfe (2002: 1) find that 'childhood neglect can have severe, deleterious short- and long-term effects on children's cognitive, socio-emotional, and behavioral development'; and while both groups exhibit developmental deficits, 'relative to physically abused children, neglected children have more severe cognitive and academic deficits, social withdrawal and limited peer interactions' (2002: 1). This suggests that severe child abuse, and childhood neglect to an even greater extent, constitute *pro tanto* excusing conditions, commensurate with the type and degree of abuse.

These findings are reinforced by a broader literature in *attachment theory*, spawned by John Bowlby (1969), showing that formative circumstances are crucial for psychosocial cognition and interpersonal attachment—capacities implicated in the comprehension of the reactive attitudes. These empirical findings lend credence to Strawson's famous excuse of 'peculiarly unfortunate formative circumstances' (1963: 79), and Wolf's excuse of 'deprived childhood circumstances' (1987: 382) which otherwise might seem dubious or question-begging. The developmental research indicates that these types of circumstances predictably engender long-term deficits in moral cognition. To be precise, although these circumstances do not *necessitate* moral incompetency, they cause this deficit in probabilistic terms, that is, they render it probable. This causal modality explains disease and disability in epidemiological research, which studies multifactorial conditions in open systems. Moral incapacity is the same type of phenomenon in that it reflects a statistically significant observed correlation between a stimulus and a physiological effect.

This causal relation suffices to excuse moral incapacity because it is exactly the same modality implicated in coercive practices, such as physical torture, that

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

are considered both legally and morally exculpating by nearly universal assent. Waterboarding torture does not *necessitate* confessions—it is physically possible to choose torture and death rather than confess. However, the human propensity to submit in the face of waterboarding gives it the status of an excuse. Childhood abuse and deprivation are responsibility-defeating in exactly this sense (though not necessarily to exactly the same degree): they induce, that is, probabilistically cause, substantial deficits in moral competency.

Fischer does not mention childhood circumstances, and while Wolf regards them as relevant, she does not elucidate the relationship between these factors and the internal mechanisms she describes as necessary for responsibility. This omission generates doubt as to why these circumstances should matter and how they are supposed to undermine responsibility. The psychological research closes this gap: deprived childhood circumstances predictably give rise to significant deficits in moral cognition.

Yet another lacuna in both theories is the lack of attention to *inculpating circumstances* that run deeper than internal mechanisms. By omitting these conditions, theorists risk generating a deflationary bias toward excusing epistemic ignorance. To see this, consider the following doppelganger scenario, which despite being a thought experiment is close enough to real-world scenarios to have unmistakable practical import.

Robert Harris is a famous serial killer, often cited in moral philosophy papers (e.g., Watson 1987; Fischer 2006; Cartwright 2006), who had suffered a notoriously abusive childhood. As an adult, he murdered innocent people without compunction. He seems as good a candidate as anyone for non-reasons-responsiveness and moral insanity, which suggests, prima facie, that he is not morally responsible for what he did. Now imagine a Robert Harris doppelganger—Harris2—who did not experience these (arguably) excusing childhood circumstances or any other extenuating factors throughout his life. Harris2 lacks reasons-responsiveness and moral sanity only because he willfully neglected opportunities to cultivate ordinary moral competency over the course of his life. Although his parents were caring and attentive, and he was by all measures biologically normal as a child, as he grew older he persistently sought out poor role models and corrupting circumstances for personal gratification. Given the facts of the case, it is reasonable to conclude that Harris2 is responsible for his defective moral character and downstream antisocial actions. At the very least, Harris2 is *more* responsible for his moral failings than Harris1. If this is right, then the absence of moral competency in and of itself does not guarantee an excuse: we must dig deeper than current internal states to determine how they came about.

These considerations do not overturn internalist conditions (which still play a limited role in an externalist calculation of responsibility), but they motivate a deeper historical analysis than philosophers tend to undertake, one that examines the distal history of a person's motivational profile. These considerations also corroborate the intuition that external factors matter at all.

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

4. Environmental Factors

Collectivist accounts of moral responsibility offer a lesson to responsibility theorists. They suggest that an agent might be responsible for a collective decision even if that agent does not have autonomous control over that decision—that is, responsibility might outstrip an agent's internal states. Tracy Isaacs (2011) gives the example of the artistic director of the Stratford Theatre Company, who is responsible for an unsuccessful playlist even though her choices as artistic director are constrained by the formal criteria of the company and input from other executive members. This input might shape her judgment in ways that are not epistemically transparent to her. This shows that responsibility in collective contexts does not require full consciousness or guidance control throughout the decision-making process: a person can be responsible for a decision that was constrained by institutional criteria and informed by collective deliberation.

Collectivism thus implies that internal criteria are not strictly necessary for responsibility. In this section, I want to expand on this argument by showing that these criteria are inadequate not only in group contexts, narrowly defined as goal-oriented collectives, but also in cases of individual deliberation in broad social contexts. Rather than enlarging upon extant collectivist arguments, I will make this case more concisely by presenting three scenarios in which the protagonist's psychological states and personal histories are fixed and relevantly similar, but environmental factors vary across cases. This way, if there is a moral difference across cases, it must pertain to the differential external variables.

1. Chum is an alcoholic living in a society where alcoholism is pervasive and generally accepted and in which moral infractions resulting from alcoholism are generally excused. Addictions resources, such as medication and psychotherapy, do not exist. Chum often engages in antisocial behavior due to his alcoholism, such as acting belligerently, instigating brawls, and refusing to seek employment, but he has never been subjected to punitive sanctions.
2. Rum is an alcoholic living in a society where attitudes toward alcoholism are ambivalent. Some people object to alcoholism while others endorse it. In spite of the existence of negative social attitudes, there are no addictions resources. Unlike Chum, Rum is a smug and self-satisfied person, who is indifferent to his alcoholism and its effects on the community. He chooses to surround himself exclusively with like-minded people who endorse his alcoholic behavior and avoids anyone who would challenge him. He often engages in antisocial behavior due to his alcoholism, such as acting belligerently, instigating brawls, and refusing to seek employment, but he has never been subjected to punitive sanctions.
3. Tum is an alcoholic living in an enlightened socialist utopia, where alcoholism is regarded as a public health concern, and effective addictions services are provided by the government at no personal cost to the individual. (They are covered by universal health care).

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

> However, like Rum, Tum is a smug and self-satisfied person, indifferent to his alcoholism and its effects on the community. He only associates with like-minded people who endorse his alcoholic lifestyle and has no awareness of addictions resources because he has never made any effort to seek them out. He often engages in antisocial behavior due to his alcoholism, such as acting belligerently, instigating brawls, and refusing to seek employment, but he has never been subjected to punitive sanctions.

Before evaluating the three cases, we should say something about the psychology of addiction. As R. Jay Wallace (1999) observes, addictive mental states ('A-impulses') inhibit volitional control in virtue of three characteristic features: they are unusually resilient, unusually intense, and connected with pleasure and pain in such a way that satisfying them is particularly pleasant while failing to satisfy them is particularly painful and can cause withdrawal symptoms and death in extreme cases. In this way, A-impulses undermine volitional control—the ability to modify one's motives by an effort of will. That said, there may be a limited degree of volitional control available to addicts, but it is severely compromised compared to the volitional autonomy of the average person.

If this were the end of the story, then addicts would be uniformly excused from blame. But empirical research on rehabilitative programs forces us to reconsider this judgment. There is substantial evidence that even if addicts lack direct volitional control over their A-impulses, they may have *indirect* control over these impulses through the use of external resources. For example, researchers at the Centre for Addiction and Mental Health (CAMH)—Canada's largest teaching addiction and mental health hospital and a World Health Organization Collaborating Centre—find that certain medications are effective but underutilized in treating addictions; cognitive-behavioral therapies, motivational therapy, and brief interventions are efficacious; and clinical support significantly increases recovery rates (CAMH Newsletter 2012). This implies that in the absence of direct control, a person may still have indirect control over her A-impulses through addictions resources. These indirect control mechanisms are *externalistic* in nature but relevant to responsibility.

With this in mind, we can evaluate the three scenarios. To begin, all three alcoholics are relevantly psychologically similar: they all experience volitional impairment to the same extent, and they are all equally ignorant of any negative social attitudes toward addictions or addictions resources in their community. We are also to assume that the protagonists share relevantly similar personal histories. Given these constants, the only significant difference across the three scenarios is the presence or absence of social information and addictions services. It seems that in each scenario, the protagonist's responsibility increases in proportion to the availability of indirect control mechanisms at his disposal, should he choose to use them. Hence, Tum is more responsible for his A-impulses and antisocial behavior than Rum, who is more responsible than Chum. It seems fair to say that Rum and Tum could (and should) have tried to moderate their alcoholism, but

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

instead they chose to rest content with their reckless lifestyle in defiance of their civic duty. Yet, if they had tried to reform themselves, they almost certainly would have made inroads against their addiction. Their inertia suggests that they were, as we colloquially say, willfully ignorant.

Now, one might object that responsibility does not transmit from A-impulses to downstream antisocial behaviors. That is, even if alcoholics are responsible for their A-impulses (due to culpable negligence), they are not responsible for their impulse-driven antisocial behavior. But this response does not fit with our considered judgments in more mundane cases of moral infraction. For instance, if a university student decides to get drunk instead of doing her homework, we hold her responsible for the unfinished homework, not just the drunkenness. (Notably, we hold her responsible for the drunkenness even though it temporarily inhibited her volitional control, revealing a collective practical commitment to historicism on some level). My own explanation for this transitivity of responsibility is that the objectionable act (of getting drunk) and the objectionable omission (of not doing homework) are ontologically coextensive. That is, they amount to the same state of affairs, described in different terms. By parity of reasoning, we can say that long-term neglect of addictions resources is ontologically coextensive with acquiring alcoholism, which is ontologically coextensive with cultivating alcohol-induced antisocial dispositions. The three states of affairs are of an ontic kind. This makes sense on Davidson's theory of action (1980), on which an action is amenable to multiple descriptions, as well as Williams's theory of 'thick' and 'thin' moral description (1985), on which different moral terms can describe the same event. Both of these pictures are compatible with describing *neglecting addictions resources* as *cultivating A-impulses*, and *cultivating A-impulses* as *fostering antisocial dispositions*. The three descriptions are ontologically coextensive but embed increasingly thick moral language.

Note that internalism does not support a differential judgment of responsibility across the three scenarios, since relevant internal states are held constant. Rum and Tum *ignored* available information about addictions. This ignorance, according to the scenarios, amounted to a *lack* of cognitive content: it was not a positive mental state available to their GNW or reasons-responsive mechanism. Rather, the protagonists *lacked* knowledge of available rehabilitation services. However, because they counterfactually could have availed themselves of addictions resources fairly easily, they are to that extent responsible.

This conclusion has important social implications, because it suggests that agents are relatively less responsible in contexts of social deprivation and coercion. There are significant analogues between Fischer's neurosurgeon example and the alcoholism scenarios above. In Chum's case, responsibility is extinguished by Chum's lack of alternative deliberative options. In Rum's case, we might want to say that responsibility is displaced onto government officials, who failed to respond to the public disapproval of alcoholism by providing addictions services. This is similar to finding the evil neurosurgeon responsible for the victim's coerced actions: responsibility is distributed or displaced onto the coercing or constraining agent. Although I cannot here elaborate on the concept of distributive responsibility, it

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

bears analogues to Martha Nussbaum's (1993) distributive account of moral agency (extrapolated from Rawls's theory of distributive justice), according to which governments have a non-voluntary responsibility to foster citizens' basic human capabilities. To the extent that they do not (on one natural interpretation), they are responsible for the social ramifications of this failure, including their citizens' misconduct due to moral incapacity. This is an extrapolation of Nussbaum's view that fits nicely with externalism regarding responsibility and deserves to be investigated further.

5. Implicit Bias

As we saw, TCT implies that no one can be responsible for implicit attitudes, and Fischer's and Wolf's reasons-responsiveness condition seems to have the same implication. But the foregoing considerations provide grounds for challenging this view. We have established that a person can be responsible for unconscious, non-reasons-responsive states if indirect control mechanisms are reasonably available in the person's cultural environment. Is there reason to think that such mechanisms exist for implicit bias?

Such evidence is provided by psychological research. Kelly and colleagues (2010) identify three plausible interventions for implicit racial bias: 'manipulating the immediate environment, self-control, and blocking the development or acquisition of implicit bias' (2010: 459). These methods are supported by burgeoning research on racial cognition, such as Dasgupta and Greenwald's (2001) finding that exposure to admired black and disliked white figures weakens implicit racial bias, even after 24 hours. Moreover, Blair and colleagues (2001) found that generating and focusing on counterstereotypical imagery weakens implicit racial bias. This research has significant public presence; for instance, the Implicit Association Test (IAT) is available online and has been cited by sources ranging from the American Psychological Association to MTV. More importantly, one does not need conscious access to, or guidance control over, these motives to utilize such interventions. If I am interested in mitigating any implicit biases that I might have (whether I am aware of them or not), I can employ exposure techniques, counterstereotypical focusing, implementation intentions (Mendoza et al. 2010), and so on. The pervasiveness of racial bias together with the relative accessibility of remediating measures supports a default assumption to the effect that ordinary people in modern Western society have a duty to mitigate their racial biases. To the extent that they fail to do so, they are presumptively responsible for the motivational and behavioral consequences.

6. Objections from Social Psychology

So far I have argued that when assessing moral responsibility we should place more emphasis on external factors, but now I want to argue that we should simultaneously place *less* emphasis on internal factors because mounting research

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

indicates that we are more susceptible to cognitive distortions than we tend to think, and thus we have less awareness of, and guidance control over, our mental states than we tend to assume. These findings pose two significant practical problems for internalism. First, if we do not know when people are acting on conscious, reasons-responsive (as opposed to unconscious, automatic) motives, then we cannot reliably attribute responsibility on internalist grounds. Second, if we have less direct control over our mental lives than we tend to think, then the psychological research has deflationary implications for internalism: we are not as responsible as we like to assume. This implication is not something most theorists would accept since attributing responsibility is an important social practice that helps to regulate interpersonal relationships. Significantly diminishing its scope (if this is even psychologically possible) would undermine these relationships and risk letting too many people off the hook.

To see why the research has these implications, consider a small sampling of the literature on cognitive distortions. Unfortunately, there is not space here to offer a robust literature review, but a few salient examples should suffice for our purposes. A more extensive catalog can be found in John Doris's book on (lack of) reflective agency (2015). Here is a selected list of cognitive distortions:

> Bateson and colleagues (2006) found that people were three times more likely to donate to an honor box when the payment instructions were paired with a picture of eyes as opposed to a picture of flowers. Tversky and Kahneman (1981) found that subjects' response to a hypothetical epidemic depended on whether the proposed intervention was described as 'preventing deaths' or 'saving lives' (demonstrating *framing bias*), even though the result was the same in both cases. Pelham and colleagues (2002) report that women are 18 percent more likely to move to states resembling their first name, and 36 percent more likely when the state is a perfect match, such as Virginia or Georgia. Uhlman and Cohen (2005: 475) described hypothetical male and female candidates for the position of police chief, in alternate conditions, as 'formally educated' and 'streetwise' and found that subjects strongly preferred the male candidate regardless of which predicate was given. Bertrand and Mullainathan (2004: 991) report that employers who were sent identical résumés, except that half were headed by 'White-sounding names' and half were headed by 'African American-sounding names', responded to twice as many résumés with White-sounding names, and the result was no different for employers who had advertised an employment equity policy.

While one might question the ecological validity of these experiments, they have all been replicated many times—for instance, the eyes-versus-flowers study, to my knowledge, has been replicated using alternatives to the honor box by Haley and Fessler (2005), Burnham and Hare (2007), and Ernest-Jones and colleagues (2011). The results are also highly statistically significant in each case. Doris (2015)

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

compiles a much broader inventory of research and draws the same primary conclusion, namely, that we have less reflective awareness and autonomy than moral philosophers tend to assume. The additional claim that I am making here is that this poses a unique problem for responsibility scholars, most of whom assume internalism. If we have less reflective control over our motives than internalists suppose, then we are less responsible than they suppose on their own terms. Not only do philosophers not fully grasp the deflationary implications of their view in light of this research, they likely would not endorse these implications since responsibility theorists tend to err on the side of conventionalism: they design theories to capture common practice, not to rebuke it (notable exceptions being Manuel Vargas 2013, Levy 2011, and to a lesser extent Levy 2014). So the research on cognitive distortions unearths implications of internalism that would rankle many internalists.

A second problem for internalism is that it is exceedingly difficult to attribute responsibility accurately on an internalist model, given that our effective motives are often inferred from murky introspective data or else entirely epistemically opaque. In each of the social psychology experiments mentioned above, the subjects' motives (responsiveness to eyes, framing bias, implicit bias) were introspectively unavailable. Furthermore, from a third-person perspective it is difficult to distinguish these opaque motives from conscious, reasons-responsive motives, except in the context of meticulously designed social psychology experiments constructed specifically to track these distortions. In everyday life, we can only conjecture at others' motives, and if our introspective capabilities are as unreliable as the research indicates, then these conjectures are likely to be *even less accurate* than first-person reports. If we see someone placing a donation into an honor box, the natural assumption is that the donor is motivated by deliberate and transparent personal reasons, not a reflexive response to the presence of eyes. The effective motive in this scenario is opaque to every natural (that is, uninformed by theory) perspective. This illustrates why we cannot invest much faith in our attributions of internalist responsibility, which are at best weak inferences and at worst speculative conjectures extrapolated from our own distorted introspective judgments.

Peter Carruthers (2010) takes an even stronger approach to introspection, saying that there is no such thing: rather, all moral judgments (including judgments of responsibility) are interpreted by the mindreading faculty. He also disputes the existence of consciousness on the GNW model for similar reasons (2012). I have instead granted the (bare) existence of these mechanisms but have disputed their reliability as criteria of responsibility. If there is such a thing as consciousness, but we cannot reliably discern whether a particular motive is conscious or unconscious, then we cannot reliably, fairly, or effectively attribute responsibility on internalist grounds. This motivates an alternative account (i.e., externalism) that has independent plausibility. Carruthers and I are in agreement that psychological research prompts either radical revisionism about responsibility or an alternative basis for judging responsibility. I have offered a sketch of the latter.

One of the explanatory advantages of externalism is that externalist criteria—that is, indirect mechanisms of control—are much less epistemically problematic than internal criteria, as they are features of a person's environment rather than

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

features of a person's mind. The cultural availability of indirect control mechanisms is epistemically accessible both from a first-person and a third-person point of view because these mechanisms are located in a person's cultural environment, in a limited geographical region and limited time frame. They can be unearthed with a reasonably thorough degree of scrutiny. By comparison, internal criteria, such as whether a particular mental state was conscious, are extremely obscure from every possible vantage point. Externalism therefore has a substantial practical edge.

7. Trouble with Tracing?

Manuel Vargas (2005: 287) has highlighted problems with what he calls 'tracing approaches', which trace responsibility to historical factors. In his broader body of work (2013), he admits a degree of historicism—less than I allow here—and does not mention current environmental factors. Since my view is (in one sense) a more extreme example of tracing than he espouses, I should respond to his worries. But since there is little space remaining, I must give an adumbrated response based on a broader project (Ciurria 2014, 2013).

Vargas's objection is that tracing accounts, which locate (some degree of) responsibility prior to the moment of deliberation or action, are epistemically inadequate because it is surpassingly difficult to identify relevant 'prior moments' (287) when an agent could have acted autonomously and forestalled a certain outcome. This worry is compounded by the fact that 'epistemic powers tend to degrade very rapidly when they have to be projected more than a little into the future', which implies that people cannot be expected to foresee the distal consequences of their actions and thus cannot be held responsible for 'the full range of downstream effects that would flow from' their choices (2005: 227). Vargas calls this constraint on foreseeability, or foreknowledge of distal effects of one's choices, the 'knowledge condition'. It has been suggested to me that this counts against my view.

My response to this objection is twofold. First, I have just argued that internalist accounts face a distinctive epistemic challenge regarding the identification of an agent's effective motives, that is, motives that effectively precipitate an agent's behavior. In light of our introspective impotence and susceptibility to cognitive distortions, there is little chance that introspective self-monitoring will yield more reliable data than third-person fact-finding for responsibility. At least external 'traces', in the form of historical and environmental defeaters and enablers, are available to third-person observation and interpersonal corroboration. They are in this sense akin to scientific data, which is observable and quantifiable. The question is how far we have to dig for such factors, which depends on how much external evidence we take to be sufficient for an attribution of responsibility. My position on this matter is that we do not need conclusive proof of responsible agency; modest evidence is sufficient. Returning to the example of childhood deprivation, although these circumstances do not necessitate moral incapacity, they *probabilistically condition* moral deficits, and in this sense they constitute coercion; thus, these circumstances provide *sufficient reason* to suspend a judgment of responsibility. Of

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

course, not all relevant data will be discernible even with an extensive search, but the quantity of reliable external data available for moral reasoning is greater than the quantity of reliable internal data.

The second part of my response is that Vargas is worried about the foreseeability of the consequences of choices, but my account does not require evidence of foreseeability. Rather, it requires evidence of inculpating and exculpating factors that are counterfactually available to an agent—that is, factors an agent could have utilized under ideal epistemic conditions—and these factors are third-personally observable. They are thus much more easily traceable than foreseeability since they are determinate features of a person's past and current environment, not speculative moments of unfettered choice. Since externalism does not require that a responsible agent be *aware* of enablers and defeaters, it does not require evidence of reflective autonomy. This follows from the fact that a person can be responsible for gross negligence. Thus, the tracing requirement on externalism is much weaker than for Vargas. It does not hinge on any volitional states of persons although it assumes (as any reasonable theory must) that responsible agents are persons, with minimal moral agency. That said, persons may neglect to exercise their capacity for moral agency, in which case they are (on my view) responsible for negligence and for the consequences of that negligence (*ceteris paribus*). To evaluate culpability for negligence, we need only track morally relevant resources in a person's current social context and past environment. These include such things as available social discourses, social groups, and educational resources, which can often be located with a standard Google search.

Granted, externalism must take into account many factors, but not so many that the search is impracticable. We merely need to execute a reasonable inquiry into likely defeaters and enablers and weigh these factors against one another to arrive at a balanced estimation of responsibility. This is a task that legal fact finders often undertake for courts; and if we are to give a person a fair *moral* hearing, then we should make a similar effort to acquire as much data as possible. We also have independent reason to place less importance on subjective motives, which are maligned by epistemic obscurity and cognitive distortions. Indeed, courtrooms are increasingly moving away from witness testimony due to the documented unreliability of perceptual reports (see Gazzaniga 2005: 120–42).

Theorists may require more or less evidence for a judgment of responsibility, but my thesis does not need to settle this matter. The thrust of my argument is that internalist accounts are inadequate insofar as they either reduce responsibility to internal criteria that are epistemically problematic or insufficiently appreciate and fail to explain the significance of external factors for responsibility. Externalism mitigates these problems.

8. Toward an Externalist Account of Moral Responsibility

In this paper, I have developed an externalist account of moral responsibility by identifying factors that are external to the intentional states of agents and yet affect

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

their degree of responsibility. These factors are social and environmental variables that counterfactually enable or incapacitate moral competency. Conversely, I have argued that internal states are unreliable indicators of responsibility because our effective motives are epistemically obscure from both a first-person and a third-person standpoint. This does not mean that internal states are completely dispensable, but it implies that they must be given far less weight than theorists tend to grant. Responsibility surely requires some intentional states, but consciousness and reasons-responsiveness are neither necessary nor sufficient. They are not necessary because we can be responsible for moral incompetency due to neglect, and they are not sufficient because we can possess moral competency but be excused in the presence of defeaters.

These claims support the position that responsibility, so to speak, ain't in the head, to a much greater extent than theorists tend to assume.

MICHELLE CIURRIA
WASHINGTON UNIVERSITY IN ST. LOUIS
mciurria.academia@gmail.com

References

Bateson, M., D. Nettle, and G. Roberts. (2006) 'Cues of Being Watched Enhance Cooperation in a Real-World Setting'. *Biology Letters*, 2, 412–14.

Bertrand, M., and S. Mullainathan. (2004) 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination'. *The American Economic Review*, 94, 991–1013.

Blair, I., J. Ma, and A. Lenton. (2001) 'Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery'. *Journal of Personality and Social Psychology*, 81, 828–41.

Bowlby, J. (1969) *Attachment*. 2d ed. New York: Basic Books.

Burnham, T. C., and B. Hare. (2007) 'Engineering Human Cooperation: Does Involuntary Neural Activation Increase Public Goods Cooperation?' *Human Nature*, 18, 88–108.

CAMH Newsletter. (2012) http://www.camh.ca/ en/hospital/ about_camh/newsroom/ news_releases_media_advisories_and_backgrounders/current_year/Pages/What's-the-best-way-to-treat-problem-alcohol-use.aspx; accessed September 2, 2015.

Carruthers, P. (2012) 'Moral Responsibility and Consciousness'. *Journal of Moral Philosophy*, 9, 200–28.

Carruthers, P. (2010) 'Introspection: Divided and Partly Eliminated'. *Philosophy and Phenomenological Research*, 80, 76–111.

Cartwright, W. (2006) 'Reasons and Selves: Two Accounts of Responsibility in Theory and Practice'. *Philosophy, Psychiatry & Psychology*, 13, 143–55.

Ciurria, M. (2014) 'Moral Responsibility: Justifying Strawson and the Excuse of Peculiarly Unfortunate Formative Circumstances'. *Ethical Theory & Moral Practice*, 17, 545–57.

Ciurria, M. (2013) 'Situationism, Moral Responsibility, and Blame'. *Philosophia*, 41, 179–93.

Dasgupta, N., and A. Greenwald. (2001) 'On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals'. *Journal of Personality and Social Psychology*, 81, 800–14.

Davidson, D. (1980) *Essays on Actions and Events*. Oxford: Clarendon Press.

Doris, J. M. (2015) *Talking to Ourselves*. Oxford: Oxford University Press.

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press

Ernest-Jones, M., D. Nettle, and M. Bateson. (2011) 'Effects of Eye Images on Everyday Cooperative Behavior: A Field Experiment'. *Evolution and Human Behavior*, 32, 172–78.

Faraci, D., and D. Shoemaker. (2010) 'Insanity, Deep Selves, & Moral Responsibility'. *Review of Philosophy & Psychology*, 1, 319–33.

Fischer, J. M. (2006) *My Way: Essays on Moral Responsibility*. Oxford: Oxford University Press.

Frankfurt, H. (1971) 'Freedom of the Will and the Concept of a Person'. In G. Watson (ed.), *Free Will*, 2d ed. (Oxford: Oxford University Press), 322–36.

Frankfurt, H. (1969) 'Alternate Possibilities and Moral Responsibility'. In G. Watson (ed.), *Free Will*, 2d ed. (Oxford: Oxford University Press), 167–76.

Gazzaniga, M. (2005) *The Ethical Brain: The Science of Our Moral Decisions*. New York: Harper Perennial.

Haley, K., and D. M. T. Fessler. (2005) 'Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game'. *Evolution and Human Behavior*, 26, 245–56.

Harman, G. (1999) 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error'. *Proceedings of the Aristotelian Society*, 99, 315–31.

Hildyard, K. L., and D. A. Wolfe. (2002) 'Child Neglect: Developmental Issues and Outcomes'. *Child Abuse & Neglect: The International Journal*, 26, 679–95.

Kahneman, D. (2015) *Thinking, Fast and Slow*. Toronto: Anchor Canada.

Isaacs, T. (2011) *Moral Responsibility in Collective Contexts*. Oxford: Oxford University Press.

Kelly, D., E. Machery, and R. Mallon. (2010) 'Race and Racial Cognition'. In J. M. Doris and The Moral Psychology Research Group (eds.), *The Moral Psychology Handbook* (Oxford: Oxford University Press), 111–46.

Levy, N. (2008) 'Counterfactual Intervention and Agents' Capacities'. *The Journal of Philosophy*, 105, 223–39.

Levy, N. (2011) *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.

Levy, N. (2014) *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Mendoza, S. A., P. M. Gollwitzer, and P. M. Amodia. (2010) 'Reducing the Expression of Implicit Stereotypes: Reflexive Control through Implementation Intentions'. *Personality & Social Psychology Bulletin*, 36, 512–33.

Nussbaum, M. (1993) 'Social Justice and Universalism: In Defense of an Aristotelian Account of Human Functioning'. *Modern Philology*, 90 (supplement), 46–73.

Pelham, B. W., M. C. Mirenberg, and J. K. Jones. (2002) 'Why Susie Sells Seashells by the Seashore: Implicit Egotism and Major Life Decisions'. *Journal of Personality and Social Psychology*, 82, 469–87.

Strawson, P. F. (1963) 'Freedom and Resentment'. In G. Watson (ed.), *Free Will,* 2d ed. (Oxford: Oxford University Press), 72–93.

Tversky, A., and D. Kahneman. (1981) 'The Framing of Decisions and the Psychology of Choice'. *Science*, 211, 453–58.

Uhlmann, E. L., and G. L. Cohen. (2005) 'Constructed Criteria: Redefining Merit to Justify Discrimination'. *Psychological Science*, 16, 474–80.

Vargas, M. (2005) 'The Trouble with Tracing'. *Midwest Studies in Philosophy*, 29, 269–91.

Vargas, M. (2013) *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Wallace, R. J. (1999). 'Addiction as Defect of the Will: Some Philosophical Reflections'. In G. Watson (ed.), *Free Will,* 2d ed. (Oxford: Oxford University Press), 424–52.

Watson, G. (1987) 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme'. In F. Shoeman (ed.), *Responsibility, Character, and the Emotions* (Cambridge, UK: Cambridge University Press), 256–86.

Williams, B. (1985) *Ethics and the Limits of Philosophy*. London: Fontana.

Wolf, S. (1987) 'Sanity and the Metaphysics of Responsibility'. In G. Watson (ed.), *Free Will*, 2d ed. (Oxford: Oxford University Press), 372–87.

https://doi.org/10.1017/apa.2015.24 Published online by Cambridge University Press