




The Agency-Last Paradigm: Free Will as Moral Ether

Geoffrey S. Holtzman¹ 

Received: 1 February 2017 / Revised: 18 May 2018 / Accepted: 22 May 2018
© Springer Science+Business Media B.V., part of Springer Nature 2018

Abstract I argue that free will is a nominal construct developed and deployed *post hoc* in an effort to provide cohesive narratives in support of *a priori* moral-judgmental dispositions. In a reversal of traditional course, I defend the view that there are no circumstances under which attributions of moral responsibility for an act can, should, or do depend on prior ascriptions of free will. Conversely, I claim that free will belief depends entirely on the apperceived possibility of moral responsibility. Orthodoxy dictates an agency-first thesis, according to which free will is necessarily antecedent to moral responsibility. However, I present a number of arguments against this view, and in favor of an agency-last stance, according to which the concept of free will is dependent upon that of moral responsibility. I provide further support for my case in the form of new empirical evidence regarding the stable mode of inference used to attribute free will across moral contexts. These experimental results can be interpreted to imply the deflation of one of the longest-standing veridical paradoxes in experimental philosophy. Furthermore, the sole conceptual scheme found to be capable of modeling the experimental results is also capable of illuminating several classic works in the analytic philosophy of moral agency.

Keywords Cognitive science · Conceptual analysis · Compatibilism · Determinism · Experimental philosophy · Free will · Incompatibilism · Individual differences · Metaethics · Metaphilosophy · Moral psychology · Moral responsibility

On February 17th, 2004, the state of Texas administered a lethal injection to Cameron Todd Willingham for pouring gasoline all over his home, blocking the exit with a refrigerator, and setting fire to the house while his three daughters slept inside. Multiple investigators reported evidence of “mineral spirits” and “puddle patterns all over the place” from lighter fluid, which led from under the girls’ beds to the front door of the

✉ Geoffrey S. Holtzman
geoffreyholtzman@gmail.com

¹ Department of Psychology, Franklin & Marshall College, P.O. Box 3003, Lancaster, PA 17604, USA

house. Willingham's own lawyer has since said that over twenty pieces of "evidence showed that he was one hundred percent guilty" (Grann 2009). One neighbor testified that as the fire blazed, Willingham calmly moved his car down his driveway to keep it away from the flames, and waited until the authorities arrived to "put on a show," purportedly feigning a level of emotion that led a police chaplain to conclude that Willingham "was in complete control." Supposing that these events were the inevitable result of natural laws, the invariable unwinding of a clockwork universe—that they took place in a universe that adheres to the philosophical thesis of *causal determinism*—might we have to relinquish the thought that Willingham really was in control of his behavior, that he even had the choice to behave otherwise?

1 Motivating the Concept of Agency & the Concept of Motivated Agency

Why do philosophers care about free will? The impetus to preserve or willingness to forgo a viable notion of agency is generally motivated by *a priori* attitudes toward the need to maintain or ability to do without a robust concept of moral responsibility. But this only shifts our inquiry to an analogous one about why people care about moral responsibility. This shift suggests that moral responsibility may be a more fundamental concept than free will, which in turn raises important new research questions for metaphilosophy and moral psychology.

Normally, people assume that moral responsibility presupposes agency, but that picture may be perfectly backwards. Instead of facilitating meaningful blame and praise, free will may only be an instrument of blame and praise, a secondary and inessential construct through which we rationalize immediate dispositions to seek retribution and yield to obligation. This is roughly the opposite of what I have been taught, and I initially found the idea nearly impossible to fully wrap my head around. Nonetheless, I have found this perspective to yield a great deal of explanatory power. If we consider that free will may only be a vestigial heuristic for *explaining* moral responsibility in the absence of more sophisticated neuropsychosocial explanations—rather than a necessary condition for formally *establishing* moral responsibility—then it becomes unclear if the existence of free will is relevant to moral inquiry at all.

1.1 Topical and Methodological Purview

In §2, I discuss a widely studied puzzle of moral-agentive cognition,¹ which I use as a window into the current moral psychological landscape. Against this backdrop, I contrast my own views on moral-agentive judgment in practice (psychological questions) and in theory (philosophical questions). I identify in §3 a number of matters arising from previous attempts to understand the puzzle, and offer a way to circumvent the logical inconsistencies encountered on the approaches traditionally taken by philosophers and psychologists. I do so by proposing a change in perspective, which I

¹ By moral-agentive cognition, I mean the processes through which the moral agency of others is evaluated. I deliberately use the opaque term 'moral agency' and its conjugates in all propositions that refer ambiguously to moral responsibility and free will, but which fail to distinguish between the two. My reasoning for employing this uncommon catchall is to flag and avoid the potential pitfalls addressed in §4.

more fully outline and compare to its theoretic competitors in section §4. In §5, I take a decidedly empirical turn, in order to test the predictive accuracy and relative fit of each theoretical paradigm² of explanation. In that section, I operationalize the proposed new approach, as well as its theoretic competitors, as predictive models of moral-agentive cognition, and test each of these models against experimental data of ordinary philosophical judgments. The discussion of these results, which constitutes §6, serves to explain how the model I defend can reduce to triviality several psychological pseudoproblems arising from the purportedly “paradoxical” (Sinnott-Armstrong 2008) nature of ordinary beliefs about moral responsibility and free will.

Finally, I discuss in §7 how the *prima facie* renegade theory I defend is in fact anything but the radical departure from tradition that it may at first appear to be. On the contrary, the *agency-last paradigm* I embrace is uniquely capable of providing a unified philosophical perspective from which we can more easily develop a coherent picture of several of the most prominent metaethical arguments in twentieth-century philosophy. Rather than debating whether or not free will exists, my interest here is in considering the question of whether, from a strictly moral perspective, free will matters. My thesis is that regardless of whether or not free will exists, its existence—something about which we can neither have certain knowledge nor certain doubt, and which we could never come to know through experience (Kant 1785/1998)—may be largely irrelevant to the domain of moral inquiry, the domain with which most people who entertain questions of agency are primarily concerned.

2 A Puzzle of Moral-Agentive Cognition

Why do we tend to view others as autonomous agents—as endowed with the freedom of choice (Nahmias 2006), as behaving intentionally (Knobe 2004), and as possessed of causal powers (Alicke et al. 2011)—in proportion to the extent to which they have ‘done wrong’? With greater moral transgressions comes greater—or at least more certain—moral responsibility. This stands to reason, as the responsibility one shoulders by killing thousands in cold blood far exceeds the moral burden undertaken by lighting up a joint or smoking a cigarette in a no-smoking zone. But does it also stand to reason that greater moral transgressions are indicative of greater levels of volition, deliberation, and instrumentality?

2.1 Type and Token Thought Experiments

Consider a deterministic universe, in which...

...scientists figure out the exact state the universe was in at the time of the big bang, and figure out all the laws of physics as well. They put this information into

² I use the term ‘paradigm’ in the fourth sense listed in the *Oxford English Dictionary*: “A conceptual or methodological model underlying the theories and practices of a science or discipline” (2005/2014). I reserve the word ‘model’ for discussions of statistical tests of observed moral-agentive attributions, which I analyze with a technique called *structural equation modeling* (Kline 2011).

a supercomputer, and the computer perfectly predicts everything that has ever happened and ever will happen. In other words, these scientists prove that everything that happens has to happen exactly that way because of the laws of physics and everything that's come before.

Now, suppose that in such a universe...

...someone commits a crime.

With both these premises in mind, consider two questions:

Was this person free to choose³ whether or not to commit this crime?

How morally responsible is this person for committing this crime?

Next, consider a second scenario, adapted from the actual case of Cameron Todd Willingham, the triple-murder-arson suspect who ultimately received the death penalty for the crimes of which he was accused (Grann 2009). Suppose that in this same deterministic universe...

...a man named Todd has taken to abusing his daughters. In order to cover up this abuse, he pours gasoline all over his home one morning while his wife is out shopping, lights the house on fire, and successfully murders his daughters while remaining unharmed himself.

Here, we can ask:

Was Todd free to choose whether or not to commit this crime?

How morally responsible is Todd for committing this crime?

2.2 The Source of the Puzzlement

Even when asked to assume determinism, people are more likely to make *attributions of⁴ moral responsibility* ('MR') (Nichols and Knobe 2007) and *of free will* ('FW') (Nahmias et al. 2005) to agents who have committed more severe moral transgressions. Especially puzzling is the fact that for some *types* of act (e.g., crimes), most people tend to deny *MR* and *FW*, yet they usually assert *MR* and *FW* for certain *token* acts of those very same types (e.g., Todd's criminal infanticide). Previous researchers have generally

³ Throughout this paper, I treat 'free will' and 'free choice' as interchangeable. If this gives the reader the impression that I (or the participants in the experiment) do not know how one is supposed to use the term 'free will,' then the reader is beginning to get my point.

⁴ I use *MR*, *FW*, and *ACT* to refer to the psychological constructs, rooted in personal perception, of which participant ratings are an indicator. In light of recent evidence and arguments for 'interpretive diversity' (Nichols and Ulatowski 2007), I think it is important to distinguish these constructs from the potentially mind-independent, objective, real phenomena with which philosophers are typically concerned.

defended one of two interpretations of the effects of the *affective, concrete*, or morally *transgressive* nature of an act (*‘ACT’*) on attributions of moral agency.

3 Normative and Deflationist Accounts of the Puzzle

What do people really believe about the compatibility of free will and causal determinism, and are their beliefs contradictory? Historically, experimental philosophers have adopted what I will call the *normative approach*, taking results like these to show that concrete, affect-laden details can bias attributions of moral agency. Because these normative theorists believe that these results reveal a logical contradiction in ordinary attributions of moral agency across contexts, many researchers have taken such studies as evidence of a kind of “abstract/concrete paradox” (Sinnott-Armstrong 2008). From this point of departure, these researchers have sought to develop *error theories*.⁵ Error theories are characterizations of the *situational factors* that select for faulty processing and attribution of moral agency (Leslie et al. 2006), and of the *cognitive processes* that lead to these errant judgments. But there is a rift between error theorists who think that *ACT* inhibits the production of competent, accurate attributions of moral agency, and those who believe that it actually facilitates accurate moral-agentive attribution.

One camp of normative theorists, perhaps guided by their own *a priori* commitment to the philosophical thesis that genuine moral agency is incompatible with determinism, insists that the intrusive presence of certain details leads people to erroneously attribute inflated levels of moral responsibility and free will. These expressed opinions, they argue, actually bely people’s true, “naturally incompatibilist” beliefs (Kane 1999). Members of the other camp, who are sometimes motivated by an overt interest in promoting the belief that moral agency is compatible with moral responsibility and free will (Vohs and Schooler 2008), argue instead that it is the absence of affective details that biases people, causing them to mistakenly mitigate the levels of moral responsibility and free will they assign in the presence of deterministic factors.

Opposed to the normative approach is the *deflationist approach*,⁶ which is typically defended by skeptics about experimental philosophy. Such skeptics believe that experimental philosophy has nothing to contribute to philosophy as traditionally construed, and that empirical approaches can only reveal the unstable “folk intuitions” (Kauppinen 2007) of philosophical novices who have been confronted with stripped down cases and denied sufficient time or resources to adequately reflect on them. While normative theorists find results like those discussed here to be interesting and surprising, deflationists raise doubts about the philosophical import of such findings (Cappelen 2012), and about the broader, potentially insidious assumption that

⁵ One exception to this rule is the *Norm Broken, Agent Responsible (NBAR)* theory (Mandelbaum and Ripley 2012), which focuses on transgressions rather than emotion/cognition or abstract/concrete distinctions. Most of my criticisms do not apply to these sorts of value-neutral approaches, and I think that projects like these may be particularly informative.

⁶ It would be misleading to contrast deflationists directly with error theorists. Many (and I suspect most) deflationists believe that philosophical questions have objectively correct and incorrect answers. On this view, competent and errant responses might be thought to correspond directly to these correct and incorrect answers, respectively. The primary difference between (realist) deflationists and normative theorists, then, is only the philosophical import each attaches to her preferred error theory.

“intuitions are likely to be reliable and should form the building blocks for sound moral judgments” (Sunstein 2005). From a skeptical, if somewhat reductive perspective, such findings might be seen as nothing more than specific examples of the general principle that upsetting events influence ordinary judgments. If concrete, affective details only influence ordinary attributions of moral agency through domain-general processes—that is, through processes capable of biasing all sorts of cognition, not just moral-agentive cognition—then the effects of *ACT* on *MR* and *FW* might not be of any uniquely moral or philosophical significance.

3.1 Shortcomings of the Normative and Deflationist Approaches

But neither the normative nor the deflationist approach is entirely well-founded, and both interpretations impugn the so-called ‘folk intuitions’ of outsiders to the analytic philosophy community without sufficient justification for doing so. Each approach purports to tell us what to make of the variance in harshness and leniency with which people make moral-agentive attributions in different situations. But it does not follow from the fact that in different situations people tend to respond differently (*more or less* harshly or leniently) in their moral-agentive attributions, that in some of these situations they tend to respond differently than they should (*too* harshly or leniently). To claim that instability in the application of agentive concepts across cases can properly be characterized as logical inconsistency in those concepts’ application is to make two unwarranted assumptions.

First, there seems to be no way of deciding which of these two ostensibly inconsistent applications is errant, other than by reference to one’s own *a priori* attitudes toward the compatibility or incompatibility of moral agency with causal determinism; the ‘error theory’ to which a researcher subscribes will inevitably be biased by her own philosophical beliefs. The reason for this is straightforward. The assertion that a person has responded differently than she should have implies certain propositions about the range of responses she could have given which would have been more acceptable. But these propositions amount to little more than claims about whether or not full causal determinism may be the sole factor that decides moral agency, even in the presence of other moral and agentive factors. As such, any theory premised on an assertion of inconsistency (as opposed to mere instability) would have to presuppose the very same highly contentious philosophical framework it sought to crown as the one most people really believe—either compatibilism, or incompatibilism.

As an illustrative example, imagine that an empirical scientist wishes only to defend a ‘psychological reading’ of one of these error theories, which only describes behavior and the component processes leading to it. This scientist, we can assume, wishes to eschew anything that might be construed as a ‘philosophical reading,’ a reading that extends interpretation to the metaphysical or normative-ethical domains. In other words, she intends only to describe the processes by which people arrive at different philosophical conclusions in different cases, and to indicate the goal-directed success or unsuccess of those processes in each case. Still, her error theory must characterize certain practical judgments as either representative or non-representative of subjects’ competent, unbiased beliefs, and therefore as resulting from cognitive processes taking place under ideal (controlled) or non-ideal (confounding) conditions. But characterizations of this latter kind require her to embrace some set of *a priori* views toward the

relevance or irrelevance of certain *ACT* details to competent, unbiased moral-agentive judgment. Such views are tantamount to *a priori* assumptions about the normativity or deviancy of the circumstances under which these acts are committed, and in which these moral-agentive assessments are made. From this, it follows that some of our researcher's *a priori* assumptions will have to be ones about the moral normativity or deviancy of the acts committed and judged.

Second, in taking the instability of moral-agentive ascriptions across situations to be logically inconsistent, researchers usually attribute these supposed mistakes to the presence or absence of emotional or affective processes. But the tacit assumption that the key factor is affective or other processes *per se*, rather than the component of affect (or other aspects of mentation) central to and perhaps even constitutive of (Prinz 2007) judgments of moral agency, is unwarranted. If it turns out that the details of a moral transgression (or the way in which that transgression is presented) affects either *MR* or *FW* in a way that is essentially indirect—that is, which only occurs due to some causal link between *MR* and *FW*—then studying these indirect effects could reveal fundamental facts about the conditional relationship between the ordinary language concepts of moral responsibility and free will. Such findings could reveal attributions of free will to be largely inert in the initial assignment of moral culpability. If this were true, we would have reason to doubt that agentive ascriptions are anything more than moral honorifics rendered *ex post facto*.

3.2 The Need for a New Approach

Ultimately, the philosophical project of deciding which moral-agentive evaluations are appropriate in a given situation is inseparable from the project of deciding whether a given situation is conducive to making appropriate moral-agentive evaluations. In order to develop any error theory, a researcher must make *a priori* assumptions about the liberty or illiberty of agents, and about the moral normativity or deviancy of acts committed and judged. The unjustified and, I have argued, seemingly false presuppositions made by both kinds of error theories necessitates an alternative explanation of the effects of *ACT* on moral-agentive cognition. This demands that we shift our research focus away from trying to decide which judgments of moral agency are right and which are wrong, and turn instead to a new set of questions about the relationship between agency and moral responsibility.

The relationship between moral responsibility and free will is usually taken to be analytic (that is, true by definition), so focusing on the nature of the logical connection between these concepts is not only useful from a philosophical perspective, but might also help us develop and test a psychological model of judgments of moral responsibility and free will. This, in turn, might tell us something about the analytic interdependence of these two philosophical concepts. Rather than inquire as to the individual differences that underwrite moral and philosophical disagreement and misunderstanding (Holtzman 2013), my purpose here is to explore an even more perplexing question. This is the question of whether we first assess free will, and on the basis of this information build our understanding of an agent's moral responsibility, or if we first make attributions of moral responsibility, and only later construct agentive narratives on the basis of these attributions.

4 Paradigms of Moral-Agentive Relata

What are the common conceptual schemata that allow us to coherently debate and discuss such opaque phenomena as liberty, willfulness, and agency? Here, I begin to address the focal problem of how it is possible to infer moral responsibility and free will from ordinary observations of behavior. In particular, I am interested in what causes (or allows) people to draw these inferences even when they assume causal determinism (one version of which I have laid out in §2).

The thesis that causal determinism is compatible with free will is, for reasons that may be obvious, referred to by philosophers and psychologists as *compatibilism*; and the thesis that the two are not compatible is called *incompatibilism*. For reasons that may be equally obvious, the thesis that causal determinism is compatible with moral responsibility is also referred to as *compatibilism*, and the thesis that the two are not compatible is called *incompatibilism*. But as illustrated by way of parable in Frankfurt (1969), and explicitly elucidated by Fischer (1982), compatibilism and incompatibilism about causal determinism and moral responsibility are distinct from compatibilism and incompatibilism about causal determinism and free will.

In coining the terms *compatibilism* and *incompatibilism*, Slote (1969) intended them to refer to only the theses that we might call *agentive compatibilism* and *agentive incompatibilism*. For Slote, these were views about free will, which did depend on “one’s evaluation of certain moral issues,” but also on a number of other factors, including “the force and significance of certain similes, analogies, and diagrams.” They are broader versions of the compatibilist theory that William James (1884) called *soft determinism*, and the incompatibilist position he dubbed *hard determinism*, more inclusive because they do not require the assumption that determinism is true (an assumption which, it just so happens, James rejected). But in time, Slote’s words have replaced Strawson’s (1963) *optimism* and *pessimism* and begun to lead a double-life, moonlighting as shorthand for *moral compatibilism* and *moral incompatibilism* while keeping their jobs as monikers for *agentive compatibilism* and *agentive incompatibilism*. It should be clear, then, that any thoroughgoing understanding of “Folk Intuitions About Moral Responsibility and Free Will” (Nahmias et al. 2005) must recognize the distinction and relationship between these two kinds of compatibilist thesis—ones concerning free will, and ones concerning moral responsibility.⁷ This recognition naturally requires us to understand the distinction and relationship between free will and moral responsibility. The purpose of this section is to develop that understanding.

4.1 The Identity Paradigm

In their watershed paper, “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions,” Nichols and Knobe warn that “one might maintain that determinism is compatible with moral responsibility but not with free will” (2007). But

⁷ It should be acknowledged that Nahmias et al.’s 2005 paper, contemporary reactions to which kicked off the experimentalist debates between compatibilists and incompatibilists, did in fact measure both constructs. It may be due in part to the unfortunate similarity in ratings of *MR* and *FW* for the particular cases tested by Nahmias and his colleagues that many researchers have since begun to assume that the two types of attribution are—or can at least be reported as—more-or-less identical.

with rare exception (e.g., Feltz 2013), it has become *status quo* for psychologists and philosophers to treat ‘free will’ as synonymous with ‘moral responsibility’ (e.g., Baumeister 2008; Paulhus and Carey 2011; Rose and Nichols 2013; Feltz and Cokely, 2009). While this *identity paradigm* may be expedient for communicating research to the public in the sexiest way possible, it makes a mess out of some of the best-known metaphysical arguments of the last fifty years. The issue is not merely that the *identity paradigm* stands in contrast to certain philosophical positions that some of us might like to defend; the issue is that the paradigm itself renders entire debates, and every contrasting position within those debates, utterly incoherent.

Fischer’s *semicompatibilism*, developed from the premise that we can “separate compatibilism about causal determinism and moral responsibility from compatibilism about causal determinism and freedom to do otherwise” (1987), would be far less influential if he had endeavored on the *identity-paradigmatic* project of trying to separate compatibilism about moral responsibility from, well, compatibilism about moral responsibility. Van Inwagen’s assertion that we possess “the free will required for moral responsibility” would be a fallacious case of affirming the consequent, were we to think of him as arguing that people have the moral responsibility required for moral responsibility (1983). And Pereboom’s denial of “whatever sort of freedom is sufficient for moral responsibility” (2001) would read as miserably circular to believers in free will, if all he really was denying was the sort of free will sufficient for free will.

Philosophers have done an excellent job of identifying the ways in which moral responsibility and free will are theoretically orthogonal. As experimentalists work to identify the ways in which attributions of these concepts are made, it is important not to lose sight of their theoretical distinction, as this distinction could be expected to dictate their use in practice. Any satisfactory account of moral agency or the attribution thereof should recognize this distinction. Therefore, one of three possible non-identity accounts can be expected to provide greater insight than the *identity paradigm*.

4.2 The Common Causes Paradigm

One possibility is that attributions of free will and moral responsibility are independent and immediate responses to certain kinds of perceived acts. On this view, their covariance (or tendency to co-occur) belies their common dependence on some third variable or set of variables. The puzzle at hand suggests that the culprit, should this view be correct, is their shared ancestry in the affective, concrete, or morally transgressive nature of an act. Within this *common causes paradigm* (Fig. 1a), *MR* and *FW* are not conditional on one another. Unfortunately, this approach raises a glaring question: If moral responsibility and free will are not causally linked, why should affective, concrete, and moral factors influence judgments of free will?

The *common causes paradigm* demands an independent path for this influence, but it is not clear why we should expect such a path to exist. It seems reasonable that a person’s moral responsibility at some time T_2 depends on just what she has done at some earlier time T_1 , but it does not seem possible that the free will an agent possesses at some earlier time T_0 could depend on what she has not done and will not do until some later time T_1 . The *common causes paradigm*, it can be seen, unyokes moral responsibility and free will too much.

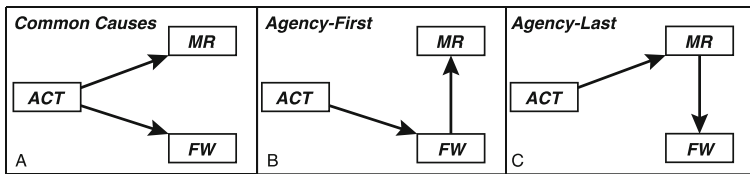


Fig. 1 Model specification: theoretic paradigms of moral-agentive relata. Panels show three competing paradigms of moral-agentive entailment relations. **a** The common causes paradigm takes the correlation between MR and FW to be a spurious “third variable” phenomenon. **b** The agency-first paradigm assumes that free will is assessed prior to judgment of moral responsibility, and therefore can affect its assessment but not vice-versa. **c** The agency-last paradigm, advocated here, is based on the theory that judgments of moral responsibility are immediately influenced by the details of moral transgressions, and attributions of free will are essentially conditional on these earlier judgments of moral responsibility

4.3 The Agency-First Paradigm

Most philosophers have endorsed, at least implicitly, an *agency-first paradigm* (Fig. 1b). This approach introduces a level of parsimony by treating MR and FW as serial rather than parallel judgments, which may be appropriate insofar as moral responsibility and free will are serial phenomena. By definition, an agent’s free will at some time T_0 can only play a causal role in an event which occurs at some later time T_1 , and she cannot be morally responsible for the consequences of that event until some even later time T_2 , once those consequences have manifest. In keeping with this sequence of events, the *agency-first paradigm* contends that the association between attributions of moral responsibility and of free will is due to a conditional relationship thought to hold between them: To whatever extent ACT informs FW, a portion of these effects will be indirectly transmitted to MR via their influence on FW. Because this sort of conditional process in moral-agentive cognition is perceived by many to be a desirable reflection of the relationship between the metaphysical concepts of moral responsibility and free will, it is the relationship most commonly assumed by social scientists (Baumeister 2008) and philosophers (Dennett 1984) to hold between MR and FW. But this assumption is based on fallacious reasoning.

Belief in the priority of agency over responsibility is driven primarily by commitment to a widely-held dogma regarding temporal priority, not logical antecedence. This dogma dictates that an agent must freely choose to commit an act prior to its occurrence, and therefore prior to being responsible for its occurrence, in order to be morally responsible for that act. While this may be true, logical priority and temporal priority are two very different matters, and in the case of necessary preconditions, the arrow of time and the arrow of entailment will always head in opposite directions. To say that moral responsibility must *follow* free will is to say that free will must *follow from* moral responsibility. Therefore, the fact that a person must freely choose to commit an act before she can be morally responsible for its consequences does not license us to infer moral responsibility from free choice; instead, the fact that a person is morally responsible licenses us to infer that she has acted freely. Philosophers who wish to maintain that free will is a necessary precondition for moral responsibility are therefore in no position to defend the *agency-first paradigm*.

4.4 The Agency-Last Paradigm

On the contrary, philosophers have unwittingly developed a conception of moral responsibility and free will that implies an *agency-last paradigm* (Fig. 1c). The potential philosophical implications of the suggestion I have just made are not insignificant. The concept of moral responsibility plays, at the very least, an important causal role in individual and group behavior. For this reason, even those who believe it to be little more than an instrumental construct have reason to pay it heed, in the interests of social regulation and social explanation. Free will is also thought to play a role in social cognition, in substantiating attributions of moral responsibility. But the *agency-last paradigm* suggests that whatever roles free will might play in social cognition, establishing moral responsibility from the get-go is not one of them. If so, its deployment *ex post facto* as a source of moral justification can only underwrite circular arguments.

4.5 From Psychological Models to Conceptual Paradigms

Traditionalist readers, who might be expected to embrace the *agency-first paradigm* and perhaps also to eschew *metaethical naturalism* (see Prinz [in preparation](#)), may at this point be unimpressed by the claims I have just made about inference. In fact, some readers sympathetic to my views will have recognized that inferential priority is not the same thing as logical priority. But because people generally take the relationship between free will and moral responsibility to be analytic—and in particular, because philosophers who care about free will justify their interest on the basis of this analytic relationship—inferences about moral agency are strictly logical ones.

In the substantive sense in which many philosophers aspire to speak about free will, a person cannot, *ex post facto*, choose past courses of actions for which she is already morally responsible, anymore than she might foresee past events for which others were morally responsible. Conversely, a person can only be known by others to have partaken in some event of her own free will *ex post facto*, just as foresight can only be confidently attributed in retrospect. Thus, the assumption that free will can be established independently from moral responsibility is based on a gross misunderstanding of the basic logic of moral causation. Philosophers who wish to maintain that free will is a necessary precondition for moral responsibility are therefore the last people who should want to defend the *agency-first paradigm*.

If free will has no empirical correlates in daily life, no explanatory role in theory, and no instrumental value in social cognition, then it is unclear why we should concern ourselves with that concept at all. The purpose of the next section is to test the hypothesis that attributions of free will are secondary to judgments of moral responsibility, and do not play an essential role in the understanding or regulation of normative judgment.

5 Experiment

From where do we derive the apperception of free choice? Here, my goal is to show that the variance in *FW* across moral transgressions can be accounted for entirely by the

variance in *MR* across those transgressions. I also hypothesize that, conversely, the variance in *FW* fails to explain a significant amount of the variance in *MR* across cases. Together, these two findings would show that *MR* and *FW* are distinct, yet not entirely independent judgments. Moreover, such findings would demonstrate that the influence of *ACT* on *FW* can be understood entirely in terms of moral considerations of a transgression, so long as we accept the view that free will falls out of—rather than factors into—moral responsibility. To test the predictive accuracy and relative fit of each paradigm, undergraduates were recruited to participate in an experiment.

5.1 Participants

Participants were selected from introductory philosophy classes in the university-wide required core curriculum at Brooklyn College, and were told that participation would not affect their course grades. Responses were delivered with no identifying information, and pen-and-paper surveys were proctored by a professor who was not their instructor, in order to preclude any form of actual or perceived coercion. All instructors of classes from which participants were recruited indicated that free will, moral responsibility, determinism, and compatibilism had not yet been addressed in their courses. All students who passed a comprehension check⁸ and filled out every response of interest were included in the analysis ($n = 228$). Of these, 42% were female; 5.3% over age 30; 46% White, 19% Asian, 17% Hispanic or Latino, 10% Black, and 8% Other or Mixed.

5.2 Methods and Materials

First, participants were asked to imagine the deterministic universe described in §2. To check that between-group differences in *MR* and *FW* were due to the experimental manipulation rather than *a priori* differences in willingness or ability to assume a deterministic universe, participants were asked whether they thought that our own world is deterministic in a way similar to the world they read about. Overall, 17.5% of participants indicated that our own world is deterministic in the same way as the world described in the vignette. There was no significant difference in rate of *a priori* determinism between groups ($t_{226} = .347, p = .729$).

Next, participants were presented with one of the two crime vignettes described in §2.⁹ Participants were then asked to indicate whether or not the agent freely choose to commit his crime, and were then prompted to rate the agent's level of moral responsibility on a Likert-type scale from 1 ("Not at all responsible") to 7 ("Absolutely responsible"). These procedures again used the materials presented in §2. *FW* was always collected first, in order to increase the likelihood that participants would

⁸ After reading about a deterministic world, participants were asked whether our world is subject to the very same deterministic principles. Participants who responded 'Yes' to this question, and who also indicated that the agent in the vignette was 'Not at all responsible,' were excluded from all analyses. Presumably, these participants (a) did not understand the task, (b) did not take the task seriously, or (c) rejected the very concept of moral responsibility *a priori*, regardless of the (as-yet unread) case at hand. Thus, their responses would not reflect perceptions of the relationship between moral responsibility and free will, and could therefore not be used to adjudicate between any of the possible models of moral-agentive cognition.

⁹ Vignette materials are presented verbatim, in offset, italicized text in §2.1.

evaluate free will before considering moral responsibility, thereby increasing the probability that the proposed *agency-last* model would be rejected.

5.3 Results

Conceptual replication Initial tests successfully replicated previous findings in the literature. As in these earlier studies, there was a significant relationship between the vignettes participants read, and the levels of moral responsibility they assigned ($t_{226} = 9.64$, $p < .001$). Participants who read about the gruesome triple-murder-arson tended to view its perpetrator as much more morally responsible for his actions ($M_{114} = 6.535$, $SE = .116$) than participants who were asked to judge a generic criminal ($M_{114} = 4.404$, $SE = .188$). Participants who read of this more heinous crime also demonstrated significantly higher odds ($OR = 2.74$) of accusing its perpetrator of freely choosing to commit his crime ($\chi^2_1 = 10.058$, $p < .005$).

Differentiation Further analyses supported the hypothesis that moral responsibility and free will were treated as distinct concepts (Fig. 2). When participants were collapsed across experimental conditions, there was a medium-sized (Cohen 1988) correlation between *MR* and *FW* ($r = .445$, $p < .001$). But as anticipated, the relationship between these ascriptions was moderated by the vignette participants read, $F_{1, 224} = 6.313$, $p < .05$. Although their correlation was medium-sized among participants who read about the child-killer ($r = .486$, $p < .001$), it was small for those who read about a generic criminal ($r = .247$, $p < .01$).

Mediation A path analysis (Hayes 2013) was then conducted to test the hypothesis that *MR* mediated the effect of *ACT* on *FW*.¹⁰ This prediction was supported by the finding that there remained no significant main effect of *ACT* on *FW* (Fig. 2). To rule out possible mediation in the other direction, a second path analysis was conducted against the more traditional view that *FW* mediated the effect of *ACT* on *MR*. This analysis found no evidence of *agency-first* mediation. As predicted, the direct effect of *ACT* on *MR* remained significant ($z = 9.01$, $p < .001$) even after accounting for potential indirect effects of *ACT* on *MR* (Baron and Kenny 1986).

Evaluation Finally, fully-specified *agency-first* and *agency-last* models were tested, and were compared to one another and to a *common causes* model (Table 1). Only the *agency-last* model predicted a pattern of responses that did not differ significantly from the actual data collected (Likelihood-ratio χ^2). That model was also the only one to provide an acceptably close absolute fit to the data after adjustments for parsimony were made to each model (RMSEA; Browne and Cudeck 1993). Comparisons between models were then made with a parsimony-adjusted index of relative fit, the Bayesian Information Criterion (BIC). These comparisons found that the *agency-last* model provided a better relative fit than either of the other two nested models, and was an even better fit than a *just-identified* model (Kline 2011) in which all variables were correlated with one another (BIC = 1459.405). The *agency-last* model was also the only

¹⁰ Coefficients of dichotomous and continuous variables were made comparable by the method first recommended by MacKinnon and Dwyer (1993) and expanded upon by Hayes (2009).

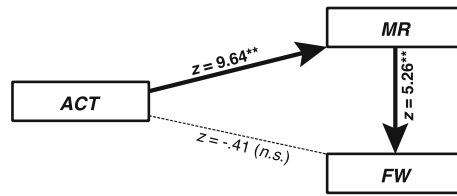


Fig. 2 Parameter estimation: agency-last mediation analysis of moral-agentive attributions. Differences in free will attribution between conditions were totally mediated by judgments of moral responsibility. After controlling for differences in MR, participants who read details of a grotesque arson actually made slightly less frequent attributions of free will than those who only read only of a generic crime, though this effect did not approach statistical significance ($p = .463$)

one to provide an acceptable improvement in fit over a null model (CFI; Hu and Bentler 1999). The *common causes* and *agency-first* models both left unacceptably large correlations between variables unexplained (SRMR), but on the *agency-last* model, less than 1% of the unaccounted for (residual covariance) between *ACT* and *FW* could have been due to a common factor.

6 Discussion

How do people actually think about moral responsibility and free will? The experiment suggests two clear answers to this question. First, people generally conceive of moral responsibility and free will as related but decidedly different constructs. In the sample population, attributions of free will were much less susceptible to the influence of moral and emotional considerations than were judgments of moral responsibility. Put another way, the extent to which *MR* and *FW* came apart depended on differences in the acts under consideration. Because these two moral-agentive concepts diverged in this way, it can be inferred that they are not identical. Therefore, any account of how people

Table 1 Hypothesis evaluation: fit between theoretic paradigms and observed attributions

	Agency-first	Common causes	Agency-last
χ^2	69.403*	41.25*	.538 (<i>n.s.</i>)
RMSEA	.548*	.420*	.000 (<i>n.s.</i>)
BIC	1523.379	1495.226	1454.513
CFI	.459	.682	1.0
SRMR	.150	.112	.012
Residual covariance	.443	.285	-.008

All fit statistics favored the *agency-last* model, and disfavored the *common causes* and *agency-first* models. χ^2 , RMSEA, and BIC are measure badness-of-fit, for which higher values indicate poorer fit (more unexplained covariance). The comparative fit index (CFI) measures goodness-of-fit, with higher values indicating better fit between the model and the data. This measure reflects the marginal improvement in prediction each model provides over a model assuming a random distribution of responses. The SRMR provides a measure of the mean absolute value of the unexplained covariance between each possible pair of variables, and the residual covariance measures the remaining association between those two variables for which no path is specified on the model

* $p < .001$

employ either psychological construct cannot be successful without sufficiently distinguishing the two.

Second, judgments of free will across moral contexts appear to be driven primarily by evaluations of moral responsibility. No aspect of the acts with which participants were confronted, other than those acts' tendencies to elicit attributions of greater or lesser moral responsibility, significantly affected beliefs about the freedom of choice exercised by their perpetrator. Importantly, this means that attributions of free will and moral responsibility came apart in precisely the way that an optimist about the ordinary ability to recognize and to distinguish between moral responsibility and free will might hope they would.

When modeled according to these two principles, ordinary attributions of free will did not appear to be significantly affected by any form of concrete or affective bias. This discovery stands in stark contrast to previous characterizations of these attributions as noisy and irrational. On the basis of these findings, I will now sketch a series of responses to the questions I set forth earlier. I expect that my account will be incomplete and not entirely precise. The discussion that follows is intended primarily as a framework within which to develop future research programs bridging conceptual analysis with the study of social cognition.

6.1 From Where Do We Derive the Apperception of Free Choice?

There appear to be at least two ways in which people become inclined to view others as possessed of the freedom of choice. The first route to attributions of agency may at first seem ineluctable: unless there is strong evidence to the contrary (and often *despite* strong evidence to the contrary), the free will of others may often be taken for granted. I will come back to this route in a moment. The second positive route is through the ascription or perception of moral responsibility. Other well-trod explanatory roads to agentic attribution, including *ACT*, may only be indirect tributaries to ascriptions of agency, contributing only as a downstream effect of their more immediate connections to moral responsibility.

To be clear, those who have wondered why *ACT* bears on *FW* have not been mistaken about the existence of such a relationship *per se*, but have failed to recognize the chain of influence through which this indirect effect is transmitted. A framework that recognizes this chain of influence is uniquely capable of supporting a psychologically and philosophically satisfying explanation of these conditional effects. *ACT* does not influence *MR* and *FW* so much as it influences *FW* *because and only because* it influences *MR*. This chain of influence may be difficult to see because it contradicts the deeply-entrenched *agency-first paradigm*. The *agency-last paradigm* remedies this error, and the reinterpretation of the data it makes possible should be seen as a victory by all those who believe that genuine philosophical concepts and their logical employ are accessible to most people. But how might we account for the tendency to default to the *agentic assumption*, even in non-moral contexts?

The *agency-last paradigm*, when more fully developed, may actually allow us to understand and study the seemingly ineluctable route to free will beliefs as well. A major reason that attributions of free will seem so inescapable is that countless trivial cases seem to invoke free will in the absence of moral responsibility. One possibility is that the invocation of free will in these cases truly does occur in the absence of moral

responsibility judgments, with such free will attributions being vestigial but ultimately derivative of moralistic concepts. This is consistent with the conception of free will as moral ether, as countless heuristics may once have had value to humans but have outlived their usefulness.

A second, more direct possibility is that in many cases where free will seems to be attributed absent moralization, there really is moralization. Consider a common example in the philosophical literature, concerning the free will to choose between fruit or chocolate cake for dessert. First, one can ask whether the perceived gluttony at work in choosing Devil's food cake over, say, grapes, really lacks moralization. But even if the choice is between, say, grapes and strawberries—a decision that may lack deontic connotations—hedonic and, ultimately, normative considerations seem to be at work. Why consider what you *could* eat, if not to make decisions about what you *should* eat? Given the opportunity to do two things, one of which will bring you more pleasure than the other, is it not true that (all other things being equal) you *should* do the one that brings you more pleasure?

To take another example, one might wonder about the extent to which freedom of choice plays a central role in cultural identity, and to which cultural identity informs beliefs about freedom of choice. In recent work, Prinz ([in preparation](#)) has noted that the tendency to automatically associate personhood with agency is very likely learned. If so, and if the *agency-last* hypothesis is correct, then it might be expected that socialized differences in moral responsibility norms should produce corresponding differences in free will belief. On the other hand, if moral responsibility presupposes free will, then fatalists should presumably partake of less retributivist forms of punishment, since only non-moral, practical concerns should guide their punitive decisions. To test this hypothesis, Prinz looked at the penal systems of the ten most fatalistic countries in the world (according to the World Values Survey; Minkov [2012](#)). Although less than one-third of all countries have retained capital punishment in law and practice, eight of the ten most fatalistic countries in the world today are retentionist states. While these findings border on anecdotal for the moment, they are suggestive of the kind of research impact that might be achieved in philosophy and psychology by flipping Kant on his head in the way suggested by the *agency-last* paradigm.

6.2 What are the Common Conceptual Schemata that Allow Us to Coherently Debate and Discuss Such Opaque Phenomena as Liberty, Willfulness, and Agency?

It follows from the non-identity of the psychological constructs of *MR* and *FW* that the philosophical concepts of moral responsibility and free will, which those constructs are meant to subserve, are also distinct. This must be true whether or not error theories are appropriate for characterizing the variance of *MR* and *FW* across moral contexts. If we are to sacrifice error theories, as I have suggested we should, then we plainly are dealing with two different concepts. But even if the effects of *ACT* on *MR* and *FW* might properly be accounted for by some error theory, then the full set of processing errors to which each of these constructs is prone would have to differ. It follows from this that the core competencies underwriting each must also be different. Because core competencies by definition involve the 'right' use of concepts, deflationists and normative theorists alike are forced to acknowledge that the philosophical concepts

of free will and moral responsibility, and the psychological constructs that underwrite their attribution are, at their core, distinct.

An objector might reply that the principles of deduction permit the inference of causes from their effects—or more precisely, the inference of necessary preconditions from the observation of phenomena that require the satisfaction of those conditions. One might therefore be tempted to think that the inference of *FW* from *MR* tells us nothing about the metaphysical relationship between moral responsibility and free will, or even about the priority of *MR* over *FW* in individuals' conceptual schema. But the inference of a cause from its ostensible effects requires external evidence that the proposed cause exists at all. Without such evidence, theoretic causes only play a nominal role in arguments that ultimately beg the questions they set out to answer.

For example, the existence of our atmosphere can be validly deduced from the continued observation of lightning on Earth; but so can the existence of Thor, if we assume *a priori* that Thor exists. And not only the equations of Special Relativity, but also the supposed existence of a luminiferous ether, was “deduced from the phenomena of light” (Maxwell 1878). But the latter, unlike the former, was unjustifiably premised on the apparently false (and inherently etheric) assumption that light always “must be somewhere, and supported by some material agency” (Poincaré 1902). Likewise, the inference of free will from moral responsibility requires the further, unwarranted assumption that moral responsibility must be mediated by some metaphysical agency—that is, that that free will exists in the first place.

Against the presupposition that there exists a material ether that mediates the transmission of light between distant bodies, John Stuart Mill once pointed out that “cases may be cited, even in our imperfect acquaintance with nature, where agencies that we have good reason to consider as radically distinct, produce their effects, or some of their effects, according to laws which are identical” (1868). This same point can be applied to the presupposition that free will exists. Just as the revolution in physics at the turn of the twentieth century eventually enabled a shift away from ethereal theories of luminescence, recent advances in neuropsychology, epigenetics, and other human sciences at the turn of the twenty-first century have begun to obviate the postulation of free will.

6.3 What Do People Really Believe About the Compatibility of Free Will and Causal Determinism, and are Their Beliefs Self-Contradictory?

There is nothing it could mean to discover that most people are “natural” (as opposed to “unnatural”) compatibilists or incompatibilists. This is not just due to the fact that most people have never thought about compatibilism, or that they (like Strawson) “do not know what the thesis of determinism is” (1963), or that they (like I) do not know what free will is supposed to be. It is because any claim about whether or not most people are “natural compatibilists” will rest on an error theory, and this will in turn rest on theoretic assumptions about the philosophical significance of *ACT* details. Thus, any reading of the data that purports to *discover* what most people “naturally” believe can only *telegraph* what a particular researcher already believes. As such, any claim about what people really believe will rest on assumptions about the philosophical significance of certain facts. But the fitness of these facts to the principles of normative ethics and metaethical truths, and their relevance to the metaphysics of causation, is beyond the

scope of behavioral science, which can only account for their bearing on moral cognition in practice. Therefore, the dichotomy between a purely descriptive, ‘psychological’ error theory of moral-agentive attribution, and an inherently evaluative, ‘philosophical’ error theory, is false.

This insight is crucial in defusing the questions raised by philosophers who have previously asserted that “Abstract + Concrete = Paradox” (Sinnott-Armstrong 2008). The variance in ordinary attributions of moral agency across contexts is not indicative of any inherent contradiction or paradox in ordinary beliefs about moral agency and determinism. In the experiment discussed here, there was no direct effect of *ACT* on *FW*. This revelation implies that there is nothing special or mysterious about “abstract framing” or “affective vignettes” that causes different behaviors undertaken in the same deterministic circumstances to be more or less frequently viewed as freely chosen.

6.4 Why Do We Tend to View Others as Autonomous Agents in Proportion to the Extent to Which They Have ‘Done Wrong’?

Ordinary judgments of free will vary across moral contexts differ precisely because the philosophical concept of free will is derived from that of moral responsibility. Admittedly, the skeptic might be justified in taking the fickleness of ordinary moral-agentive attributions to lack philosophical significance. The problem with such a flippantly deflationist interpretation of the data presented here, however, is that these data do not speak to any such unreliability. To the contrary, the experiment discussed herein revealed attributions of free will to be highly reliable, and extremely predictable as an effect of moral responsibility attribution. If deflationists believe ordinary philosophical judgments to be so wildly unpredictable, so uninformed, and so unfounded, then they should be absolutely baffled by the regularity with which I have found ordinary people to use moral-agentive concepts. When statistically modeled as a moral construct arising from the notion of moral responsibility, there was nothing significantly unreliable about ordinary views regarding free will and its compatibility with determinism. The consistency with which moral responsibility guides judgments of free will across contexts is striking, and the implications of this for analytic philosophy should not be overlooked and cannot be dismissed.

Philosophers may claim that *what* people believe about moral agency is none of their business, as philosophers. Nonetheless, *how* people come to have these philosophical beliefs (moral epistemology), *whether* they should act upon these beliefs (normative ethics), and *why* these beliefs turn out to be true or false (moral metaphysics) constitute three of the core problems of ethics and metaethics. Identifying these fault lines in the bedrock of philosophical conviction, and studying the theoretical rifts and argumentative forces that arise from these conceptual divisions, are at the heart of analytic philosophy.

6.5 Why Do Philosophers Care About Free Will?

Most philosophers seem to care about free will because of “the internal conception of agency and its special connection with the moral attitudes as opposed to other types of value” (Nagel 1979). But postulating entities solely on the basis of their conformity to our internal conception of phenomena in the external world is not especially instructive

or practical. Furthermore, philosophy stands to gain little by characterizing ordinary philosophical judgments as paradoxical, errant, unnatural, unreliable, or irrelevant. I therefore think that we should shift our efforts away from such causes, and focus instead on answering two more basic questions.

First, we might ask what guides our conceptual behavior in the moral-agentive domain. The past decade of research in moral psychology and experimental philosophy has already begun answering this question, and has yielded a wealth of knowledge to which I am indebted for many of the ideas I have discussed here. Second, we might ask what the psychological structure of moral-agentive constructs can reveal to us about debates in philosophy. The *agency-last paradigm* equips us for both lines of inquiry, providing a propitious opportunity for philosophical prospecting. On this approach, ostensibly competing ordinary attributions of free will might be brought into harmony. Might competing philosophical theories, which have heretofore been considered by many to be equally motley, be seen to have much more in common with one another when viewed through the lens of the *agency-last paradigm*? In the next (and final) section, I offer an affirmative answer to this question.

There are several directions in which one might take this line of inquiry. A historian of philosophy might ask, as Nietzsche did, where “the thought that ‘the criminal deserves punishment, because he could have acted otherwise’” first arose within “the psychology of mankind in its early stages” (Nietzsche 1887/1967). A cognitive psychologist might seek to study the component processes that lead to the ascription of certain moral-agentive attributes under various conditions. I think both of these approaches may be well-suited for innovative and fruitful projects. But in the space that remains, I only want to explain how the discovery of the psychological tendency to infer free will from moral responsibility may help us enrich our understanding of some of the most important insights in twentieth century analytic philosophy. And so, after disowning the idea that it is worth pursuing any error theory at all, and after rejecting both the normative and deflationist stances, I expand in the final section on the deeper philosophical implications of the *agency-last paradigm*.

7 Agency-Last Perspectives on Classic Metaethical Puzzles

It may seem as though I have impugned some of the most fundamental beliefs of philosophers and non-philosophers alike, but this is not the case at all. I do not think that my beliefs are fundamentally all that different from those of most other philosophers. I am of the deepest conviction that in a literal sense, many of the most prominent figures in 20th- and twenty-first-century philosophy have tacitly operated within the *agency-last paradigm*. The views of many others can also be much better understood if we adopt this perspective. Just as the logically deduced taxonomies worked out by philosophers like Fischer (1987) have had major implications for the empirical study of moral cognition, discoveries in moral psychology—including those presented here—might help us better understand the philosophy of moral agency. The *agency-last paradigm* provides a theoretical and methodological instrument that not only helps us understand ordinary moral-agentive cognition and attribution, but which also yields a plausible philosophical framework that reduces free will to little more than a morally motivated accusation.

7.1 Contempt, Coercion, and Control

The *agency-last paradigm* can help us better understand Strawson's claims in *Freedom and Resentment*, where he suggests that free will is only an account of others' "attitudes towards us of goodwill, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other" (Strawson 1963). But attitudes like goodwill and contempt are inherently moralistic, and as Strawson himself points out, we can only hold these attitudes toward those we view as persons responsible for their own behavior. We could not, for instance, be contemptuous of a chalkboard for being the teacher's pet. All of this suggests that asking whether an agent who is thought to be morally responsible does or does not have free will might only serve to beg the question, and that Strawson does not push his view far enough. Whereas Strawson suggests that free will may be fundamentally ineluctable, I see no reason why this should be the case. I think that questions about moral responsibility are primary to, and cannot be furthered by, inquiries into free will. The attitudes for which Strawson takes free will to be a kind of shorthand may or may not be held by those to whom we attribute them. But the automaticity with which we deride wrongdoers for their calculated willfulness seems to unveil 'free will' as nothing more than a moral epithet.

Like Strawson, Frankfurt recognizes the immediacy of moral responsibility from (transgressive) action. In *Alternate Possibilities and Moral Responsibility*, he describes a man who, in accord with his own motivations, performs a certain action that another man would have forced him to perform, had he himself lacked the initiative to do so autonomously. In unpacking this case, Frankfurt points out that in at least some instances, the moral responsibility of an agent is decidable even if it remains unknown whether that agent could have freely chosen from among alternative possible courses of action:

The fact that a person was coerced to act as he did may entail both that he could not have done otherwise and that he bears no moral responsibility for his action. But his lack of moral responsibility is not entailed by his having been unable to do otherwise (1969).

But in trying to leverage free will into the space that Frankfurt's moral argument opens up, his interpreters¹¹ end up sapping free will of its primary value—subservience to the notion of moral responsibility.

The problems identified by Nagel in *Moral Luck* also largely resolve themselves if we reject the *agency-first* assumption, and instead take an *agency-last* approach. Nagel points out the (fairly obvious) fact that moral responsibility for what one has done depends on what one has actually done, and therefore also the circumstances in which

¹¹ I do not think it would be fair to attribute this interpretation to Frankfurt himself. My reading, though it seems to be an unpopular one, is that Frankfurt makes a conspicuous effort to establish moral responsibility without reference to the spooky notion of free will:

The two main concepts employed in the principle of alternate possibilities are "morally responsible" and "could have done otherwise." To discuss the principle without analyzing either of these may well seem like an attempt at piracy. The reader should take notice that my Jolly Roger is now unfurled (1969).

one has done it and the outcomes to which it has led. He recognizes that “from the point of view which makes responsibility dependent on control, all this seems absurd” (1979), but he also recognizes that the dependence of moral judgment on circumstance and outcome is clearly not absurd. In fact, its *independence* from these factors would be absurd, potentially leading to consequences like the jailing of all people who, while using cell phones at the wheels of their cars, had the good fortune *not* to inadvertently strike and kill children. Nagel is deeply perplexed by the apparent tension between the principle that we are not responsible for outcomes outside of our control, and the realization that in real-life cases people are never or almost never in control of the outcomes of their behavior. This tension perplexes him so much that he describes it as paradoxical. But for whatever reason, he fails to recognize that the rejection of this paradoxical ‘Principle of Control’ is not the only way to dissolve the absurdity he discusses. Instead, one might simply reject the *agency-first paradigm*, which is at the root of the usual “point of view which makes responsibility dependent on control” (Nagel 1979).

7.2 The Puzzle of Moral-Agentive Cognition

Perhaps the most ironic place in which a potential *agency-last* solution to—or rather, dissolution of—the “abstract/concrete paradox” in moral psychology is approached but never quite captured is the very work that first brought that puzzle into the spotlight. In a footnote in that article, Nichols and Knobe acknowledge that “one might maintain that determinism is compatible with moral responsibility but not with free will” (2007). This rather deft observation—which they credit to Fischer (1987), but whose philosophical roots, it can now be seen, run much deeper than that—is tantamount to an admission that factors which impinge upon free will do not necessarily impinge upon moral responsibility. And this, in turn, implies that free will is downstream of moral responsibility, as postulated by the *agency-last* paradigm.

At the same time, Nichols seems to recognize that there is widespread acceptance that facts about *MR* necessarily entail ones about *FW*. When *Science* published his repackaging of this same dataset from Nichols and Knobe (2007) as “The Experimental Philosophy of Free Will,” nobody blinked an eye at the fact that the data only reflected judgments of whether agents were “fully morally responsible” (2011) for their actions. But it should be clear to the reader by now that using ‘free will’ as a stand-in for ‘moral responsibility’ in this way is deeply problematic. The presumable justification for taking such liberties is that one need not hold any particular position about free will in order to hold a given position toward moral responsibility, but that views on moral responsibility necessarily inform free will attitudes. But this, of course, is an *agency-last* justification. Thus, it still fails to answer the question of how research into free will matters, even if a deeper understanding of moral responsibility does. This, in turn, raises the question of what to make of the concept of free will itself. I conclude with my views on this subject.

7.3 Free Will as Moral Ether

Freedom of choice, unlike violations of moral principles, is not something that others *commit* and that we *observe*. Freedom of choice is only something that other people are

said to *possess*. In this way, the concept of free will is by its very design opaque. Whereas the observation of moral transgressions and the assignment of responsibility for those transgressions are both significant because of the causal roles they play in individual and social behavior and regulation, free will occupies, at best, a latecomer's role in social cognition and explanation. But the *agency-last* structure of moral-agentive cognition calls even this minimal value into question.

The notion of free will may have been central to a framework that people once had to assume in order to most successfully theorize about the connections between moral-agentive relata. In this sense, there may have been a time when appeal to free will could be justified as a (metatheoretically desirable) form of inference to the best explanation. Over time, the centrality of that notion has been codified to the point of dogma, but the pre-scientific concept of free will has become so disconnected from the rest of our contemporary explanatory framework as to be merely nominal.

From an ontological perspective, the proposed existence of free will is not so much false as it is vacuous. Moral responsibility without free will is only inconceivable in the trivial sense in which "space without ether is unthinkable" (Einstein 2010/1920). In this sense, free will is only a kind of moral ether:

With our pre-scientific concepts we are very much in the position of our archaeologist in regard to the ontological problem. We have, so to speak, forgotten what features in the world of experience caused us to frame those concepts, and we have great difficulty in calling to mind the world of experience without the spectacles of the old-established conceptual interpretation. There is the further difficulty that our language is compelled to work with words which are inseparably connected with those primitive concepts (Einstein 1954).

Acknowledgements I am grateful to Joshua Knobe, Jesse Prinz, and Hagop Sarkissian for helpful comments on previous drafts of this manuscript.

References

- Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy*, 108(12), 670.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.
- Baumeister, R. F. (2008). Free will in scientific psychology. *Perspectives on Psychological Science*, 3(1), 14–19.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, 154, 136–136.
- Cappelen, H. (2012). *Philosophy without intuitions*. New York: Oxford University Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Sussex: Psychology Press.
- Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge: The MIT Press.
- Einstein, A. (1954). The problem of space, ether, and the field in physics. In C. Seelig (Ed.) & S. Bargmann (Trans.), *Ideas and opinions*. New York: Wings Books. (Original work published 1934).
- Einstein, A. (2010). Ether and relativity. In G. G. Jeffrey & W. Perrett (Trans.), *Sidelights on relativity*. Mineola, NY: Dover Publications. (Original work published 1920).

- Feltz, A., and Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*, 18 (1), 342–350.
- Feltz, A. (2013). Pereboom and premises: Asking the right questions in the experimental philosophy of free will. *Consciousness and Cognition*, 22(1), 53–63.
- Fischer, J. M. (1982). Responsibility and control. *Journal of Philosophy*, 79(1), 24–40.
- Fischer, J. M. (1987). Responsiveness and Moral Responsibility (1987). In Fischer, J. M. (2006). *My way: Essays on moral responsibility*. New York: Oxford University Press.
- Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), 829–839.
- Gramm, D. (2009, September 7). Trial by fire: Did Texas execute an innocent man? *The New Yorker*.
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4), 408–420.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York: Guilford Press.
- Holtzman, G. S. (2013). Do personality effects mean philosophy is intrinsically subjective? *Journal of Consciousness Studies*, 20(5–6), 27–42.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- James, W. (1884). What is an emotion? *Mind*, 34, 188–205.
- Kane, R. (1999). Responsibility, luck, and chance: reflections on free will and indeterminism. *The Journal of Philosophy*, 217–240.
- Kant, I. (1785/1998). *Groundwork for the metaphysics of morals*. (M. J. Gregor, Trans.). Cambridge: Cambridge University Press.
- Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations*, 10(2), 95–118.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: Guilford Press.
- Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, 64(282), 181–187.
- Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17(5), 421–427.
- MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, 17(2), 144–158.
- Mandelbaum, E., & Ripley, D. (2012). Explaining the abstract/concrete paradoxes in moral psychology: The NBAR hypothesis. *Review of Philosophy and Psychology*, 3(3), 351–368.
- Maxwell, J. C. (1878). Ether. In W. R. Smith (Ed.), *Encyclopedia Britannica* (9th ed.).
- Minkov, M. (2012). World values survey. *The Wiley-Blackwell encyclopedia of globalization*. G. Ritzer (Ed.). <https://doi.org/10.1002/9780470670590.wbeog840>.
- Nagel, T. (1979). *Mortal questions*. New York: Cambridge University Press.
- Nahmias, E. (2006). Folk fears about freedom and responsibility: Determinism vs. reductionism. *Journal of Cognition and Culture*, 6(1–2), 215–237.
- Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom : Folk intuitions about moral responsibility and free will. *Philosophical Psychology*, 18(5), 561–584.
- Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, 331(6023), 1401–1403.
- Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, 41(4), 663–685.
- Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, 22, 346–365.
- Nietzsche, F. W. (1887/1967). *On the genealogy of morals*. (W. Kauffman, Trans.). New York: Random House.
- Paradigm (June, 2005). In *Oxford English Dictionary Online* (3rd ed.; 2014). Retrieved from <http://www.oed.com/>.
- Paulhus, D. L., & Carey, J. M. (2011). The FAD-Plus: measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, 93(1), 96–104.
- Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.
- Poincaré, H. (1902). *Science and hypothesis*. W. J. G. (Trans.). Toronto: University of Toronto Press.
- Prinz, J. J. (2007). *The emotional construction of morals*. New York: Oxford University Press.
- Prinz, J. J. (2014). Metaethical naturalism. Manuscript in preparation.
- Prinz, J. J. (2014). Moral Bondage. Manuscript in preparation.

- Rose, D., & Nichols, S. (2013). The lesson of bypassing. *Review of Philosophy and Psychology*, 4(4), 599–619.
- Sinnott-Armstrong, W. (2008). Abstract + Concrete = Paradox. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy*. New York: Oxford University Press.
- Slote, M. A. (1969). Free will, determinism, and the theory of important criteria. *Inquiry*, 12(1–4), 317–338.
- Strawson, P. (1963). Freedom and resentment. *Proceedings of the British Academy*, 48: 1–25.
- Sunstein, C. R. (2005). Moral heuristics. *The Behavioral and Brain Sciences*, 28(4), 531–542.
- Van Inwagen, P. (1983). *An essay on free will*. Oxford: Clarendon Press.
- Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, 19(1), 49–54.