What's the Relationship Between the Theory and Practice of Moral Responsibility?

Henry Argetsinger[*]
hargetsi@ucsd.edu

Manuel Vargas[*]
mrvargas@ucsd.edu

ABSTRACT

This article identifies a novel challenge to standard understandings of responsibility practices, animated by experimental studies of biases and heuristics. It goes on to argue that this challenge illustrates a general methodological challenge for theorizing about responsibility. That is, it is difficult for a theory to give us both guidance in real world contexts and an account of the metaphysical and normative foundations of responsibility without treating wide swaths of ordinary practice as defective. The general upshot is that theories must either hew more closely to actual practice than they appear to, or they must provide some normative foundation for responsibility that does not go through actual practice.

Suppose that, in a wide range of cases, ordinary people ascribed moral responsibility by relying on concepts, tracking properties, and engaging in judgments that, from the standpoint of our best theories of responsibility, were irrelevant, superfluous, or apparently erroneous. What, if anything, would follow? The answer is complicated. Depending on one's methodological commitments, there are several positions the responsibility theorist might reasonably take. However, each line of reply raises further puzzles. Given this, our ambition in this paper is two-fold: (1) to make the case for a novel challenge to standard understandings of responsibility practices, animated by experimental studies of biases and heuristics, and (2) to use this to illustrate a general methodological challenge for theorizing about responsibility. More specifically, we argue that it is difficult for a theory to give us both guidance in real world contexts and an account of the metaphysical and normative foundations of responsibility

[*] University of California San Diego, USA.

without treating wide swaths of ordinary practice as defective. This may seem unsurprising and unproblematic. Considered in itself, it might be. However, taking such a position increases the burdens on the account of the theory's *normative* foundation, for it requires something other than those very same practices as a basis for critiquing the practice. Until very recently, comparatively few accounts have explicitly done this latter sort of work.[1] In its absence, a methodological challenge persists: theories must either hew more closely to actual practice than they appear to, or they must provide some normative foundation for responsibility that does not go through actual practice. In the first two sections of this article, we canvas some background presumptions and identify some general features of contemporary theories of moral responsibility. In the third section, we offer evidence for thinking that ordinary responsibility ascriptions oftentimes rely on factors that don't figure in standard theories of responsibility, and that seem to pollute the quality of responsibility ascriptions. There, we draw from work in the heuristics and biases tradition in psychology, and especially, research on responsibility attributions. In the fourth section, we argue that the resultant situation is more complex than it might initially seem. There are a variety of responses available to a proponent of conventional theories of responsibility, but we argue that the appeal of each of these positions both depends on, and casts new light on, a range of submerged methodological considerations in the theory of moral responsibility. Last, we argue that whatever position the responsibility theorist adopts, there are distinctive costs that one has to take on board, and some of these consequences are not entirely compatible with the ways in which these theories have been understood.

1. The methodological tension

Below, we argue that there are empirical considerations that favor thinking that many ordinary attributions of responsibility are propelled by apparently irrelevant considerations that do not figure in ordinary theories of responsibility. One

[1] For example, speaking in broad brushstrokes, Strawson (1962), Frankfurt (1971) Fischer and Ravizza (1998) offer theories that are not centrally concerned with the normative foundations of responsibility, whereas, say, Wallace (1994), Vargas (2013), and Brink (2021) explicitly focus on the normative foundations of responsibility. Different theories can rightly have different aims, but ours is an argument about the urgency of an account of normative foundations if theories depart from practice, as we think nearly all must do.

might accept those findings without sharing our thought that it has methodological significance; conversely, one might think the methodological tension we describe is a chronic feature of theorizing without being concerned about the empirical literature we canvas. Even so, we think the empirical and methodological considerations can and do work together in theorizing about moral responsibility.

Consider some widely (although perhaps not universally) shared methodological presumptions that govern theory construction for normative notions. On the one hand, as theories about putatively real-world phenomena, philosophical accounts of normative notions must be extensionally accurate, or coherent with major features of the phenomena as we understand them. On the other hand, theories are often proposals about how we ought to understand the thing in question, and for theories about normative phenomena, it is typically a desideratum that the resultant theory provides practical guidance about what to do in concrete, everyday cases. Thus, normative theories, including theories of moral responsibility, face a dual burden of capturing ordinary usage, while simultaneously providing normative guidance about it.

The demands of extensional accuracy and normative guidance sometimes involve tradeoffs. If a theory simply adheres to characterizing our practices as we find them, without articulating a normative ground or without offering practical guidance, the resultant theory seems unsatisfying as a piece of normative theory. Yet, guidance that simply reiterates what we find in practice is only minimally helpful. This creates a challenge of its own: the more the theory claims to prescribe something distinct from our everyday practices, the greater the pressure to say what justifies the departure. Call this *the methodological tension* in normative theories.

There are notable efforts to grapple with different versions of this problem for theory construction (Rawls, 1971; Jackson, 1998; Pettit, 2018), but the proposed strategies are all controversial, and we are inclined to think the methodological tension persists. In the next section we argue that: (i) there is reason to think that attributions of moral responsibility are often driven by factors that do not figure in philosophical theories of responsibility and (ii) those factors seem practically and epistemically irrelevant to the question of responsibility. Even if we agree that such factors should be excluded from a normative theory of responsibility, the question is: why? Here, the methodological tension raises its head. To the extent to which accounts are beholden to pressures for descriptive accuracy, the normative account seems to depart from the facts of the actual

practice. However, to the extent to which the account is untethered from the actual facts of how responsibility is ascribed, there is greater pressure to show that the normative authority of the account is grounded in a satisfying theory of the normative basis of responsibility practices. Yet even if we forsake extensional accuracy, the puzzle remains: on pain of losing our grip on what we are talking about, whatever normative ground the theory appeals to will itself face at least some pressure to cohere with ordinary thought and talk.

Perhaps the tension can be eased. It is always open to a theorist to reject a certain range of folk practices as mistaken, inessential, or degenerate. In better regimenting our concepts, theorists might have to carve out some pieces of ordinary practice and leave them by the wayside. But each move away from ordinary thought and talk raises the stakes for justifying those departures. This issue is perhaps most trenchant for theories that take themselves to be especially tightly tied to an analysis of our existing practices (e.g., Strawson, 1962). Paralleling the case of normative ethics, though, the methodological tension is not restricted to accounts so strongly committed to the priority of the social phenomena of responsibility. To the extent to which any adequate theory is beholden to standards of both extensional accuracy and normative guidance, the methodological tension lurks. If we are right, many theories of responsibility either do less than we think, or they have more work to do before they can show what they hope to show.

Here at the outset, we acknowledge an important limitation in the scope of what follows: we are restricting our focus to "success theory" (as opposed to error-theory) approaches, and we focus on compatibilist options. There is a broad family of error-theoretic approaches to responsibility, including skepticism, eliminativism, fictionalism, and some varieties of revisionism (Strawson, 1994; Caruso, 2012; Pereboom, 2014). Such accounts have done a great deal to explore a particular way of resolving the methodological tension. On those accounts, the presumption that ordinary attributions of responsibility are sometimes veridical must be abandoned. However, unlike standard error-theoretic accounts that are typically committed to a broad and unified metaphysical thesis for why people aren't responsible, our interest is in the sometimes diverse psychological phenomena that seem to intermittently but frequently interact with everyday responsibility ascriptions. It is a further question, one we will not explore here, whether the considerations operative in our account generalize to error theories, and vice-versa. Our focus on compatibilist theories is less a matter of ambition than an artifact of the fact that compatibilists have

tended to provide the most systematic theories of moral responsibility. We take it as an open question the extent to which the points we make here generalize to incompatibilist accounts, and we do not take a stand on the independent debate about whether ordinary attributions of responsibility favor compatibilist or incompatibilist positions, although we think the point we ultimately emphasize, about the need for a normative foundation for theories of responsibility, apply to success theories regardless of how one comes out on the compatibility question.

2. Theories of moral responsibility

Although there is tremendous diversity among approaches to the theory of moral responsibility (Nelkin and Pereboom, 2022), there are also some broadly recognizable attractor positions in the literature. In what follows, we employ these to identify some features that exemplify wider theoretical approaches common to contemporary theorizing about moral responsibility.[2]

Whatever else contemporary theories of responsibility may aspire to do, most accounts purport to offer a story about the properties or features of agents and actions in virtue of which responsibility obtains. This might be coupled with a story about free will. Or it might include an account of the nature of blame and how it operates. Or it might include an account of the normative grounds of responsibility practices and/or its relationship to other normative practices. However, most theories of moral responsibility say something about the features of agents that are required for responsibility. Minimally most accounts presume that blameworthy agents are "normal" mature human agents, in the sense that they do not suffer from some clear mental impairment (delusions, severe cognitive deficits, atypical affective dispositions) or some incapacity in circumstance (such as duress or physical constraint).

The precise features of agency that are said to ground responsibility vary by account. One family of views—call them Self-Expression views—says that

[2] Following a common practice, we will use 'responsibility' and related terms to refer to moral responsibility in what is sometimes called the "accountability" sense, namely, a sense concerned with backward-looking culpability. We take no stand on whether the present account illuminates other varieties of responsibility. Although our account does not address questions about answerability (Shoemaker, 2015) and attributability (Watson, 1996), we hasten to add that a focus on responsibility attributions (the act) concerning accountability (the sense of responsibility) does not imply that one is focused on attributability responsibility (a non-accountability sense of responsibility). Our thanks to a reviewer for inviting us to clarify this.

the responsibility-enabling feature of agents is a particular psychological configuration, e.g., the presence of higher-order desires that some effective first-order desire be the will (Frankfurt, 1971), or the agent's values (Watson, 1975; Doris, 2015), or some other expression of a privileged or authoritative bit of the agent's psychology that can be said to speak for the agent or the agent's evaluative perspective (Smith, 2005), or a suite of abilities to engage with others in a variety of distinctively affect- and evaluatively-laden ways (McKenna, 2012; Shoemaker, 2015; Fricker, 2016). A different family of views—call them Reasons views—have focused on whether the agent is suitably related to what reasons there are. Sometimes this is cashed out in terms of rational abilities, rational capacities, or reasons-responsiveness (e.g., Wallace, 1994; Fischer and Ravizza, 1998; Nelkin, 2011; McKenna, 2013; Vargas, 2013; Brink, 2021). These views can identify some relation to reasons as a standalone ground of moral responsibility, or in conjunction with some self-expression condition (as in Wolf, 1990, and McKenna and Van Schoelandt, 2015), or a condition concerned with opportunities, context, or the "ecology" of action (McGeer, 2012; Vargas, 2013; Brink and Nelkin, 2013; Brink, 2013).

There are, of course, other approaches on which these sorts of features do not centrally figure in the account. However, every account of moral responsibility that seeks to meet some standard threshold of fit with ordinary thought and talk must explain why the behavior of ordinary adults is sometimes a candidate for moral blame in a way that the behavior of human infants, alpacas, and hurricanes is not. In keeping with the methodological tension we identified at the outset, substantial departures from ordinary thought and talk are taken to suggest a shortcoming of these theories, a kind of theoretical cost to be explained away. Theories that intentionally violate such a constraint are sometimes characterized as revisionist, but it is a central burden on such accounts to justify their departures from ordinary thought and talk.

The pressure for conforming to extensional features of thought and talk can interact with pressures for precision in the responsibility theorist's account. A theory that gestures at "ordinary abilities" does not help us decide whether a mild phobia or some instance of neuroatypicality is a sufficient departure from ordinary to constitute unsuitability or excuse in blameworthiness. In contrast, a theory that articulates detailed and perspicuous conditions – e.g., that the agent's endorsement of the intention to act is uncontested by any higher-order attitudes (Frankfurt, 1988), or that the agent is reasons-responsive in this specific fashion and has taken ownership of his or her deliberative mechanisms

(Fischer and Ravizza, 1998), is more promising for sorting hard cases. The upshot is that on the most detailed theories – some will say *the best* theories–the conditions on responsibility can be given with relative precision, appealing to things like the precise degree of reasons-responsiveness, or exact psychological structure, or the relevant collection of evaluative attitudes, and so on, that we rightly look to for grounding responsibility.

An instructive example of the pressure exerted by the methodological tension emerges in the literature on whether and why an agent's past matters for responsibility. Many parties to these debates allow that ordinary thought and talk seems to hold that agential history *does* matter. That is, we often modulate our blame and praise based on facts about an agent's past circumstances, mental states, actions, and so on. Yet, philosophical theories have disagreed about whether the best theory of moral responsibility must adhere to this aspect of ordinary thought. Some have thought it must (Fischer, 2000; Kane, 1999; Pereboom, 2014), others that it need not (Frankfurt, 1971; McKenna, 2012), and others that it should in some cases and not others (Mele, 2019; Vargas, 2013: 267-301). What almost all theorists accept is that theories that depart from ordinary thought and talk about the importance of history for responsibility must justify or otherwise account for the gap between their theories and that thought and talk.

Notice that ordinary thought and talk often underdetermines whether a given feature is epistemic or metaphysical. Again, the history debate is illustrative. One way to close the theory/practice gap in this domain is to argue that history matters epistemically but not metaphysically. For example, one might maintain that our interest in some agent's past is to be explained by the fact that it serves as *evidence* of some impairment to current function. On such a view, history doesn't matter for the metaphysics of responsibility, or for the fundamental grounding of responsibility. Its role is merely epistemic. Adopting such a position allows theorists who reject a history requirement to explain away our apparent interest in the dispute over the agent.

In making the foregoing points, we do not endeavor to take a stand on the merits of any position about the importance of agents having a particular history for moral responsibility. Rather, we gesture at the history of this issue to illustrate the idea that an important aspect of the demand for extensional accuracy (which along with normative authority constitutes the methodological tension for theories of normative notions) is that philosophers in this domain generally recognize that they must account for departures from ordinary thought

and talk. However, departures from ordinary thought and talk can be addressed in different ways.[3]

This fit between theory and practice is not under much pressure insofar as departures in theory aim at mitigating the fallibility of our everyday judgements.[4] In the context of responsibility, it is easy enough to admit that we often make mistakes in attributing responsibility. This kind of fallibility, however, is often trivial. That is, everyone admits we are fallible creatures: the fact that we make mistakes is no knock against any particular theory. We are also capable of doing epistemic and moral reasoning that corrects our mistakes, or that lets us avoid trivial mistakes to begin with. Indeed, we might hope that this is something the regimentation of concepts offered by theories of moral responsibility allows us to do.

But we can fail to appropriately render judgments of responsibility for reasons of fallibility that go beyond triviality. When I judge you more (or less) harshly because, for instance, I have an affective reaction to a complex set of facts involving the color of your skin, your social position, what I ate for breakfast, and a half-remembered story I'd heard about "people like you," something more than making trivial mistakes is occurring. Notice, first of all, that many parts of what I'm reacting to in rendering a responsibility ascription may be very cognitively opaque to me. Indeed, I might even deny them were someone to press me: "Aren't you just blaming him because of the color of his skin?" "Of course not!" This is not to say there is no possibility of correction or improvement. It is to say that the situation is not one where we can always easily recognize where we've gone astray.

The costs of revisionism become clearer as the difficulty of correction and the widespread nature of the fallibility increase. If our best theories of responsibility enjoin us to take seriously certain metaphysical properties or conceptual tools, and if there is a wide range of cases where ordinary responsibility practitioners do not do so, it becomes less clear what we ought to say. Should we jettison wide swaths of a practice as flawed, outside the "core" of responsibility, or as somehow non-genuine?

[3] Some error-theoretic approaches may happily jettison ordinary thought and talk itself in favor of relatively radical reformations or replacements of ordinary practice. Revision on that scale is both a potential strength of a theory and a potential cost to its plausibility. As we noted above, though, our focus is not on error-theoretic and related approaches.

[4] Thanks to an anonymous referee for pushing us on this point, and for suggesting the language of fallibility and triviality.

Take as an example a theory which employs a variant of the Self-Expression story – in this case, a "Deep Self" – to make sense of a pattern of folk responses. Now imagine that in a range of cases the folk do not seem to possess or make use of this concept. Suppose we find that in roughly 20% of cases people aren't actually concerned with the things that the Deep Self concept picks out. From the standpoint of the theory, those 20% of cases are erroneous. Still, we might not be troubled. But what percentage would tempt us to say that the costs of adopting the theory are too high because the theory is too disconnected from ordinary practice? This is the kind of tension we think most theories of responsibility face, once we look at the empirical data concerning ordinary responsibility attribution practices.

3. Everyday ascriptions of moral responsibility

We now turn to making a more specific case for the claim introduced above: our ordinary responsibility judgments may often be successful, but they are subject to a range of failures as well. In particular, from the standpoint of going theories of responsibility, there is a wide range of cases where our actual attributions of responsibility rely on properties and concepts not identified by philosophical theories of responsibility. Once the range (and kind) of cases are clear, we'll have more to say about how this puts pressure on contemporary theories of responsibility.

As an analogy, consider the aforementioned literature on situationist social psychology and moral responsibility (Doris, 2002; Nelkin, 2005; Brink, 2013; Vargas, 2013b; McKenna and Warmke, 2017). A number of philosophers have thought there is a prima facie threat here from experimental findings that show that normatively irrelevant and incidental features of the environment play a much larger role in explaining people's behavior in normatively-laden contexts than has been appreciated. Philosophical replies in this literature have been varied, with some holding the empirical evidence can be accommodated without needing to make concessions in our theorizing (Brink, 2013), others holding that reforms are required in our frequency of ascriptions (Vargas, 2013) or in the suitability of notions of individual evil (Murphy and Doris, 2022). We

think that the psychological literature on biases and heuristics poses an analogous threat to responsibility theorizing.[5]

Still, if biases and heuristics are frequently present in our responsibility attributions, there are a range of cases (call them "polluted" cases) which all (non-error-theoretic) theorists of responsibility will have reason to reject as defects in the practice. And, as we noted above, this rejection of a widespread feature of ordinary practice, however infelicitous, heightens the pressure to have a compelling case for the normative authority of responsibility practices. Yet, once we begin to concede aspects of extensional fit, and to assert that this is earned in light of the normative accuracy of the proposed theory of responsibility, we need to know something about why that normative theory is authoritative or binding in the way asserted by the theorist. To meet that burden, it is not enough to have identified a common structure in our practices (as Strawsonians are sometimes held to offer), or to have demonstrated the explanatory power of the proposal (as in accounts that foreground an agent's identification or rational capacities). As error-theorists are quick to note, such things are compatible with our practices being in widespread error or otherwise not normatively authoritative. What the success theorist requires is something that justifies the normative authority claimed for the theory.

If one's account of responsibility ascriptions drew exclusively from existing philosophical theories of responsibility, it would be tempting to think that ordinary attributions of moral responsibility work something like this: when there is a candidate instance of wrongdoing, we search for a proximal perpetrator and evaluate that perpetrator in light of the properties identified by a theory of moral responsibility as being those that ground moral culpability. If there is the right fit, we blame. If there is not, we excuse or exempt the candidate wrongdoer from blame.

[5] This project is pursued in the spirit of other efforts to explore the ways in which specific kinds of empirical considerations might be thought to bear on moral responsibility, as in debates about Libet-style research (Mele, 2009; Maoz et al., 2019), situationist social psychology (Doris 2002; Nelkin, 2005; Brink, 2013; Vargas, 2013b; McKenna & Warmke, 2017), and experimental philosophy efforts to ascertain the nature of folk commitments of responsibility (Sommers, 2010; Nichols, 2011; Cova & Kitano, 2014). Like those accounts, we take it that one can do philosophy grounded in and responsive to empirical findings. Building on a wider family of experimentally-informed work, we consider what experimental findings show about everyday responsibility ascriptions, and how and why this gives rise to an important version of the problem of methodological tension.

Different theories will fill out the details of what we are looking for in different ways. So, for example, we might understand a particular Self-Expression theory as holding that we assign responsibility by looking to see if, upon performing some wrongful act, the offending agent identifies with the action. Or, if we favor a reasons-responsiveness account, we should expect that ordinary responsibility attributions will involve deciding whether the action was done by an agent who had a suitably robust ability to recognize and respond to morally salient reasons.

Psychologists who study responsibility attribution have tended to suggest that this is not, in fact, how we assign responsibility – or at the very least, not all the time. Empirical findings suggest a very different picture of how ordinary ascriptions operate (for relevant surveys, see Alicke et al., 2015; Devine & Caughlin, 2014; Earp et al., 2021; Feigenson & Park, 2006; and Woolfolk et al., 2006). In what follows, we'll briefly sketch some of the main strands of contemporary psychology that bear on responsibility ascriptions. We draw from a standard picture in contemporary psychology according to which our minds are not entirely transparent to us and a good deal of what we believe and what we do is cued automatically or non-consciously, relying on affect and heuristics rather than conscious deliberation. Some of the details of this framework are matters of ongoing dispute (e.g., one can take a "dual systems" approach too literally (Christensen and Michael 2015)). In broad outlines, though, there is widespread acceptance of three key claims: First, when we form judgements, beliefs, and plans of action, those mental states and activities blend automatic (and, to some extent, non-conscious) mental processes with effortful thought and deliberation. Second, when our mental states are tinged, distorted, or, more forcefully, "captured," by affect, biases, and heuristics, we are not always consciously aware of this fact. And third, we do not have the cognitive processing power to always deliberate rationally about all aspects of the world – we rely on shortcuts to ease our cognitive load. Empirical work indicates that all three of these claims apply to ordinary ascriptions of moral responsibility, i.e., we often rely on stereotypes, first impressions, and learned schemas in attributing responsibility.

Take two kinds of important findings from psychology, in particular from the literature on attribution theory: (1) that our initial impressions and perceptions of people are enormously influential in the formation of our assessments of their (supposed) character traits, which, in turn, color the way we view their actions (Westra, 2018 and 2019 give a good overview), and (2) that our

assessments of the causal contributions of agents are often influenced by seemingly irrelevant up-stream and down-stream information (Cuddy, Fiske, and Glick, 2007, for instance).

Call (1) the Imperviousness of Initial Impressions: A body of research finds that stereotypes and judgements of warmth and competence strongly influence our initial perceptions of agents, and that these initial impressions very quickly allow us to infer and model more stable character traits (see Cuddy, 2007; Feigenson, 2016; Fiske, 2002; 2007; Nadler, 2012; Nadler and McDonnell, 2012; Rahimi, 2016; Westra, 2018, 2019). Indeed, "personal proximity" – that is, how close we are to an agent along various dimensions of in-vs-out group assessment – has much to do with how we assess them as trustworthy, likable, or virtuous (see Cuddy, 2007; Fiske, 2002; Malle, 2014; Nadler, 2012; Rai, 2011; Spaulding, 2018; Suedfeld, 1985; Willemsen, 2018; and Zell, 2009).

Most importantly, many of these initial impressions of warmth, competency, and character are "sticky." Once we have some stable character-based models of agents, it is unlikely that they will be significantly revised, even if they are updated with new information.[6] It is easier to model downhill than uphill, in other words: if we think that someone is a liar and a jerk, it is hard for us to overcome this assessment even when presented with evidence to the contrary. Perhaps, for example, the lying jerk is just being nice to set us up for a future con. And, if we get more evidence of jerkiness, the model can incorporate this with ease.

Call (2) Causal Confusion: our estimations of causal contribution and control are quite often influenced by irrelevant information. Whether we like someone, what our current mood is, how bad the consequences are, what someone's particular social role is: all these factors influence our estimations of the causal control an agent has in a responsibility relevant context. We are particularly sensitive, for instance, to the severity, or goodness/badness of an action's outcome (see Alicke, 2008; Alicke, 2015; Fishbein, 1973; Gerstenberg, 2012, 2014, 2018).

Finally, it's worth noting that both (1) and (2) are subject to (3) Framing Effects: All these effects are mediated, enhanced, and diminished by the informational process itself. That is, the way we find out about what an agent has

[6] This is the core of Alicke's (2000) arguments, but the relationship between emotion, affect, and blame is complex and also explored by, for instance, Feigenson (2016) and Weiner (2006).

done, or the kind of narrative we encounter or build in thinking about responsibility relevant actions, affects the outcome of the attributional process. Different kinds of informational presentation activate and ramp up the affective and emotional aspects of our moral psychology in different ways, as well as making salient certain streams of information and masking others. Different ways of narrating the events can lead us to different conclusions about agents and events, and therefore, to different attributions of responsibility.

We think this process is often subtle and frequently entangled with various social demands (Doris, 2015). Often, what we are responding to in attributing responsibility is not a direct re-construction of a person's relationship to a particular event, but an intra- and interpersonally constructed narrative that emerges among various strands of popular discourse. By the time we form the judgment, "*x* has behaved poorly," we may have some inchoate sense of the previous judgments of a large range of interlocutors. The opinions and judgements of others shape and modify our own and are often foundational for where we start when attributing responsibility to an agent.

There are several psychological accounts of the mechanics of responsibility attribution that take on board some of these effects (Alicke, 2015; Guglielmo and Malle, 2017; Shaver, 2011; Weiner, 2006). What emerges is a model where responsibility ascriptions arise from an attributional process with diverse starting places (a snap judgment, deliberation, testimony about character, and so on), and which is subject to conscious deliberation and non-conscious (or automatic) reactions, emotions, and mental processes. [7] It might well be the case that there are cases when we start the attributional process by searching for mesh-like identification or reasons-responsive mechanisms. It's clear from the attributional research that, at the very least, most reasoners *are* concerned with things like causal control and agential character. This is to be expected – after all, it would be surprising if the leading theories of responsibility had *no* connection to what everyday reasoners cared about. Still, we need not begin with or utilize the properties such theories specify, and indeed, we frequently do not proceed in this way. Sometimes, we simply find ourselves struck by the wrongness of an action, or a strong feeling of dislike for an agent who seems to be involved. We might then follow several paths that involve diverse

[7] Elsewhere, one of us has argued for a particular version of this model—the "ping-pong" model—that seeks to capture the organic complexity of everyday responsibility attributions and the diverse ways in which Imperviousness, Causal Confusion, and Framing Effects shape how that process unfolds (Argetsinger, 2022 and Argetsinger, *in progress*).

psychological processes. Each process produces a proto-judgment with a certain orientation on it which affects the next step of the attributive process until a final judgment is reached. So, for instance, we may judge you harshly upon hearing about something you've done, then see a picture of you and have an affective reaction of great warmth or pity which modifies our initial judgment, then recall things you've done in the past which further modifies that judgment, and so on.

Putting this model together with the concepts of Imperviousness of Initial Impressions, Causal Confusion, and Framing Effects explored above, gives us reason to think that our responsibility judgments exhibit distinctive tendencies for bias. Alicke (2000), for instance, argues that once a negative evaluation enters our attributional system, it is far more likely that the agent in question will be found blameworthy than they otherwise would be, whatever other judgements follow. As he says, "predisposing biases, which represent departures from normative responsibility models, are endemic to ordinary blame ascription. For this reason, the psychological processes manifested in cognitive and motivational biases are central rather than peripheral to the psychology of blame" (2000: 556). Once negative spontaneous evaluations enter our mental processing, we are primed to review, seek, and interpret evidence in ways that lead to blame judgments.

Put another way: negative affect has an outsized effect on the orientation or trajectory of our judgments. Once a negative evaluation attaches to an agent, it is likely to stick all the way through our judgmental process. Indeed, given enough negative affect, it's likely that a snap judgment will be arrived at without much processing or deliberation at all. If we begin with a blame judgment before considering other interpretational factors, the bare affect of unfavorableness is likely to stick. Our emotions will run hotter, we will be oriented towards confirming evidence of "badness," and we are likely to overestimate causal control and more easily find unfavorable aspects of people's characters.

The model we've been sketching in this section is, of course, somewhat idealized and there are a variety of ways in which the account calls for supplementation. For example, this kind of deliberation can be very quick or very drawn out: judgments may be arrived at entirely non-consciously, or entirely through effortful deliberation. Even so, the psychological literature bears out the kind of idealized model we're sketching here, and it is useful to have such a model to deploy to help us notice the differences between everyday ascriptions and those that theories of responsibility predict.

With the aim of making perspicuous how these phenomena interact, and to provide some specific examples from which to hang our reflections on the theoretical options, we offer a pair of fictional vignettes, starting with the following:

> **SNAP**: Tara is drinking her morning coffee and reading a local news site when she sees the following headline: "Repeat Felon Charged in Deadly Shooting of Convenience Store Clerk." Immediately, she has a strong negative reaction — she quickly (and subconsciously) connects this headline to other recent news stories she has read about a rise in violent crimes in the city. "Ugh," she says aloud, "I hope they lock him up and throw away the key."

We take it that cases like this are relatively common in that we are often willing and able to make ready inferences about responsibility from relatively limited data. In Tara's case, it is a split-second conclusion based on a headline. Whether Tara *ought* to have judged the felon so harshly on so little evidence, people undoubtedly can and do make these kinds of judgments. To form the judgment, Tara didn't need to search for evidence about the properties of the agent or action in question. Instead, she made a snap or "intuitive" judgment, one saturated and perhaps partly motivated by a brief emotional flare.

We note two things. First, this snap judgment may be the downstream result of highly tuned sensitivities to precisely the properties of agents that our best theories of responsibility discuss. It's possible that Tara is, in some sense, reacting to reasons, albeit automatically and perhaps non-consciously. We will say more about this possibility in a moment. Second, a snap judgment is just a first pass of her judgment. She could step back and consider whether she has enough evidence to warrant the conclusion, or she might ask what sorts of things would have to be true about the alleged offender to determine whether he was indeed culpable and deserving of criminal punishment. Or she might just turn the page to read an article about real estate prices.

Imagine that she does take a bit more care, as in the following case:

> **MULL**: As before, Tara reads the story and makes a snap, condemnatory judgment. This time, though, she continues to read the story. The felon, it turns out, was convicted of his past crimes based on a case brought by a notoriously racist district attorney. He has always maintained his innocence. The article goes on to relate that, he has claimed that the police and district attorney have a personal vendetta against him. He maintains that he was nowhere near the scene of the crime on the night of the shooting.

These new facts produce new emotional reactions in Tara and cause her to reflect. She recalls her conviction about the need for police reform and better oversight of law enforcement. She also remembers an unsavory rumor she'd heard about the district attorney in question. At this point, Tara looks up from the paper and asks her wife what *she* thinks about the case.

In MULL, the initial intuitive assessment of responsibility is reconsidered because of a range of contextual features and contingent associations available to Tara. There is the fact of how the article frames some of the features of the case. Some aspect of the story, or perhaps more sustained consideration of the details, elicits background associations that further shade her assessment as she continues to read. At each stage – initial snap judgment, modification by more information, further associations about potentially involved factors –Tara might have stopped with further deliberation. The result might overturn or reaffirm the initial judgment. The vignette ends with a familiar enough phenomenon that could result in a further transformation of Tara's judgment – checking with another's assessment to calibrate or compare with our own.

For our purposes, there are two important things to note about MULL. First, none of the proximal causes of any of these changes in assessment has any obvious and direct connection with the properties that theories of moral responsibility emphasize as conditions of responsibility. Indeed, the things Tara considers in MULL seem stubbornly disconnected from the kinds of things identified as relevant to responsibility: the issue of whether and how much to trust police information, for instance, is purely epistemic.[8] Second, one might wonder whether many of the considerations driving Tara's assessment are not just detached from the things that putatively matter for responsibility but grounded in considerations that are rationally orthogonal to judgments about it. Whether one has a personal commitment to police reform is independent of the truth of a particular case; unsavory rumors about the district attorney might make one dislike a person more, without it following that the district attorney acted inappropriately. In short, cases like MULL (and SNAP) point to everyday ways in which effects like Imperviousness of Initial Impressions, Causal Confusion, and Framing Effects can do much of the heavy lifting in forming assessments of culpability, precluding, trumping, or swamping the sorts of things that putatively figure in proper ascriptions of responsibility.

[8] Thanks to an anonymous reviewer for highlighting this point.

One might wonder whether even in SNAP and MULL there *is* a kind of tacit tracking of the properties (in the sense of an ascriber's judgments being reliably and non-accidentally sensitive to the properties) that figure in our best theories of responsibility. Suppose a person's assessment in snap-judgment cases, or even in slightly more informationally rich mulling-over cases may not be *directly* assessing the properties that our best theories of moral responsibility identify. Still, why not read SNAP and MULL-style cases as instances where those properties are *indirectly* being tracked? Perhaps the features identified by our metaphysical theories are playing important roles in the background of the various phases of her judgment, such that, it really is reasons-responsiveness – or alternately, a Deep Self, or what have you – that Tara is reacting to.

One could think that responsibility ascriptions may well be tacitly relying, in part, on a theory of responsibility in sorting what features to track when ascribing responsibility. But notice that, if the claim is that our everyday judgments are mechanisms meant to track and assess the metaphysical properties identified by theories of moral responsibility, they aren't necessarily very good ones. The evidence of agents' vulnerability to epistemic pollution and various defeaters briefly canvassed above should make us cautious about accepting the claim that there is a tight connection between the mechanisms by which we arrive at responsibility ascriptions and the metaphysical theories that ground the goodness or correctness of those judgments. At best, we are simply not as good at reliably tracking the right kinds of properties as these theories assume (see suppressed for review, 2022, for this kind of argument).

On the one hand, we are happy to admit that everyday reasoners may do a decent job of ascribing responsibility.[9] After all, the practices seem, generally, to work. But, on the other hand, we suspect most theorists would readily admit that everyday reasoners are not perfect, and often get things wrong. We suggest that one reason for the many failures and injustices endemic to responsibility practices is the kind of underdiscussed pollution we canvass above. Finally, we

[9] Some philosophers have argued that we are systematically unable to track responsibility-relevant features. Vargas (2007; 2009) has argued that ordinary ascriptions of responsibility cannot track the features that matter for libertarians. Byrd (2010: 412) explicitly adopts this position as an assumption. Sehon (2013) develops the view in detail, and extends it to Fischer and Ravizza's (1998) semi-compatibilism. There is a family of related but distinct views that argue for agnosticism about free will and/or moral responsibility on the basis of broader conceptual, logical, and methodological considerations, including Vilhauer (2009); Byrd (2010); Kearns (2015); Chevarie-Cossette (2021). As noted above, we take no stand on systematic skepticism about ascriptive success. Our concern is the more variegated story from empirical psychology.

resist the claim that, in all cases, agents like Tara are obviously intending to track and assess the kinds of properties metaphysical theories posit, even poorly and indirectly. If there's an argument for an indirect tracking view, it needs to be made, and it needs to be made against the backdrop of the evidence that this is not always what agents see themselves as doing, nor something they are reliably good at.

In this context, notice a further challenge for the idea of tacit tracking: the likelihood of its vindication will depend to a large extent on the metaphysical theory one goes in for. That is, it matters for the argument whether we are supposedly tracking features of reasons-responsiveness, identificationism, or whatever else. To read SNAP or MULL as a tacit assessment by Tara of the modal profile of a responsibility-relevant mechanism, for instance, a case would need to be made. So too, if one wanted to claim that we are tacitly tracking whether an agent's action fits a wider pattern of characterological cases.

Importantly, it might be that certain kinds of cases look easier to fit with Self-Expression theories, and others with reasons-responsive theories, and so on. Notice, though, that a problem lurks for anyone who concedes this and wants to insist that ordinary practices are indirectly tracking the relevant metaphysical features. *The actual phenomena of our everyday responsibility practices underdetermine which theory we ought to adopt.* If that is right, then it will be very hard to show that we are reliably (indirectly) tracking the precise features that a candidate theory identifies. Even if such features play important roles in the background of the various phases of our judgments, we think it is more accurate (and straightforward) to say that, as described, Tara is not trying to assess whether the alleged offender is reasons-responsive or acting from his Deep Self.

There is more to say about all of this, of course, but the principal upshots are simple: (1) in a characteristic range of cases, we form judgments of responsibility based on considerations that don't figure in philosophical theories of responsibility, and (2) in those cases we tend not to seek out or identify agential properties that figure in standard philosophical theories of moral responsibility.

Before we move on, let us say a bit more about how prevalent this range of cases might be. As we argued above, revisionism gets costlier as the percentage of to-be-rejected "non-core" cases increases. It is difficult to say in the abstract how common these phenomena are. Are 15% of cases polluted? 85%? We are unaware of any systematic assessments of the amount of pollution there is in responsibility assessments. We are inclined to think it is more rather than less,

but reasonable people can disagree and the empirical evidence isn't decisive either way. Still, what the argument we are making requires is only that there is a non-trivial frequency of such cases in ordinary responsibility ascriptions – and this claim, we are confident, the attributional research does support. The wider and more prevalent cases of attribution involving the Imperviousness of Initial Impressions, Causal Confusion, and Framing Effects are, the greater the cost to a theory which seeks to explain them as non-central or degenerate kinds of responsibility attribution.

4. Methodological options

If the foregoing is correct, there is a divergence between what philosophical theories identify as driving responsibility ascriptions, and what empirical findings seem to suggest we are sometimes doing. Given the methodological tension between demands for extensional accuracy and normative authority, contemporary philosophical theories are vulnerable to a challenge from normative authority when the theory proposes an account of responsibility that departs from widespread features of ordinary thought and talk. Given that the sorts of biases and heuristics we have identified are plausibly widespread features of ordinary thought and talk, theories can be pressed to justify their departures. The seriousness of this challenge, though, depends on what one thinks one is doing with one's theory of responsibility, and the plausibility of the normative foundations of responsibility on offer, if there are any. On different ways of conceiving a theory of moral responsibility, the role of ordinary practices might matter in different ways. So, a detour through the methodological landscape is in order before we consider ways theorists might want to respond to this divergence.

For all the ink spilled by theorists of responsibility, there has been comparatively little attention given to the question of the rules of the theoretical game for success theories concerning *how* we are to build a theory of moral responsibility and what the basis is for our deciding it is successful.[10] In what follows, we identify two broad strategies that are recognizable in recent work on responsibility: *concept-first* and *practice-first* methodologies.

[10] There are, of course, important exceptions both within and adjacent to the core of contemporary philosophical work on responsibility. For a sample of self-conscious efforts to articulate a methodological framework for theorizing about moral responsibility see Strawson (1962), Wallace (1994), Vargas (2013), Nichols (2015), Pettit (2018), and McCormick (2022); for efforts

On one way of constructing a theory of moral responsibility, we start from our representational devices, focusing on the meaning of words or the contents of concepts like FREEDOM, CULPABILITY, BLAMEWORTHINESS, and COERCION. We analyze these terms and concepts, as well as their relationship to relevant adjacent notions. We test our proposals about them with various thought experiments and arguments that work from proposals about how to understand terms and concepts. On this approach, our goal is to regiment our understanding of those things to produce a theory that is internally coherent – and, hopefully, plausible enough to count as describing the world. It is an approach exemplified by, for example, classical compatibilist accounts which have sought to analyze the meaning of 'can', as well as more recent efforts where the central arguments turn on imaginative thought experiments about manipulators, interveners, and the like. Call this approach *concept-first*.[11]

A second way of constructing a theory of moral responsibility begins by thinking about the way our responsibility practices operate. The starting point for theorizing in this mode are our everyday practices of holding one another responsible, the conditions under which we praise and blame, and the characteristic ways in which we do so. The primary theoretical aspiration is to accurately describe the practices as we find them, in the hopes that by correctly describing them, we will learn something about their conditions that can properly inform philosophical debates about the nature and limits of moral responsibility. This sort of work is exemplified by accounts that have taken their methodological cue from Peter Strawson's (1962) "Freedom and Resentment." Call this approach *practice-first*.

Most contemporary theories employ a mix of methodologies. The pure forms of either may not even be possible: a concept-first method will have to make use of concepts and thought experiments that are entangled in concrete practices. A practice-first approach will be unable to describe its targets in ways

concerning free will see Double (1996), Sommers (2011), Deery (2021a) and (2021b). Still, methodological concerns have tended to be peripheral, with theorists usually helping themselves to a tacit picture of the principles of theory construction in these domains. Our account of some of the main methodological options and their labels follows Vargas (2022).

[11] If one is averse to concept-talk, one can instead read this as "representational device-first." Nothing in the characterization of this approach depends on a traditional notion of a concept, but simply requires that there are ways of representing information, things, possible actions, and the like.

innocent of any theory-laden terminology and concepts. Moreover, most theories are supplemented by other methodological principles, techniques, and substantive commitments. These may include more and less explicit methods of reflective equilibrium; the ambition to justify to others the basis of our demands and expectations; a strong commitment to methodological or substantive naturalism; the aspiration of remaking practices in light of concerns for the amelioration of injustice, the goal of vindicating independent theological commitments; or aspirations for identifying transcultural, or alternately, culturally specific commitments.

Even so, we think the distinction between concept-first and practice-first methodologies captures a recognizable difference in methodological sensibilities – including starting points and in which commitments the theorist takes the theory to be distinctively beholden to – within existing approaches to philosophical theories of moral responsibility. This does not mean there are no other methodological options.

There is arguably a third methodological approach available here, one that proceeds by applying a set of substantive commitments from an adjacent literature, for example, foundational philosophy of action, criminal law, metaphysics, or empirical research concerning human psychology, neuroscience, and the like. On such an approach, it can appear that a theory of moral responsibility is simply an application of antecedently developed first- and second-order commitments about how a theory is to be generated. But here, too, the basic options re-emerge in the application of the exogamous theoretical commitments. One can either apply those commitments in a way that is principally responsive to our representational commitments concerning responsibility (i.e., one can proceed in a concept-first fashion) or in a way that is principally responsive to features of our responsibility practices (i.e., practice-first). So, going forward, we will help ourselves to the thought that a good deal of theorizing about moral responsibility is distinguished by concept-first and practice-first methodologies.

5. Some possible responses

To the extent to which all theorists must account for departures from extensional accuracy, we think there is a general challenge here, and that the challenge is especially trenchant for theories that do not have an account of the normative authority of responsibility. Still, one's particular package of methodological commitments can alter what kinds of positions are available. In what follows, we

show that several potentially appealing strategies for non-normative theories do not, despite their differences, escape the underlying methodological challenge. In what follows, we discuss three broad families of responses and their attendant positions: *The Authoritative Armchair*, *Indifference*, and *Ascriptivism*. We do not suggest that these are the only kinds of positions available in logical space. That said, we focus on these because they seem to us some of the most immediately appealing options for those who would rather not bother with the labor of constructing a theory of the normative ground of responsibility.

On one family of approaches, a theory of responsibility is typically offering an account of one or more metaphysical matters: an account of what responsibility is, how it fits into broader pictures about thought and talk, and what must obtain for ascriptions to be true. On this approach, the practices hold little metaphysical authority in themselves. Their role is often a matter of imperfect reflection of the metaphysical ideal. When they are well-ordered, our responsibility practices, and the everyday norms they involve, accurately reflect and codify some prior metaphysical fact. To the extent to which they do, we can rely on ordinary practices and norms of attribution as ways of sussing out whether particular real world cases satisfy the metaphysical conditions identified by the theory. On this view, practices are to be measured by the extent to which they track the true theory of responsibility, and the theorist of moral responsibility needn't be especially concerned by the discovery that ordinary practices operate on principles at some remove from the theorist's account. Call this approach *The Authoritative Armchair*.

This is not to say that the state of our practices is entirely irrelevant to the Authoritative Armchair theorist. If our everyday ascriptions of responsibility have loose or worse connections with the sorts of things identified in our metaphysical theories, one might suppose that something has gone wrong (Ciurria, 2020). Still, from the armchair, we should expect there to be some slack between what our ascriptions respond to and what properties are identified by a metaphysical theory. We don't have direct access to the psychology of those we blame, including the arrangement of their higher-order attitudes, values, and policies. Nor can we directly perceive the precise shape of an agent's rational powers. Instead, we do our best to make fallible determinations with the diverse tools at our disposal, because there is no direct access to the properties that matter metaphysically and normatively. Thus, to the extent to which our ascriptions suffice to keep us "latched on" to the metaphysically relevant properties, our epistemic practices are in good standing.

The Armchair theorist takes our metaphysical theories as privileged. But is it enough to sidestep the pressure for a theory of the normative foundations of responsibility? The answer depends on how we understand the methodology presumed by the Armchair theorist. On the one hand, if one embraces a concept-first approach, we might wonder why this or that concept and its ensuing metaphysical theory has authority over our practices (including our epistemic practices) as we find them. On the other hand, if our practices of holding one another responsible are explanatorily foundational (as one might think some practice-first theories like Peter Strawson's seem to imply), then it may seem especially urgent for the Armchair theorist to explain the authority of the metaphysical over and above the structure of our ascriptive practices as we find them.[12] In either case, there might be various ways to explain this away. (After all, one could go in for a story about the normative ground of responsibility, although recall that the issue at hand is whether there is any way to avoid that task.) Still, the point is that there is a burden here that must be discharged for the Armchair to have the authority it claims for itself.

We think there is a more immediate problem for the Armchair theorist. If our account in section 3 is in truth's ballpark, there is evidence that our ordinary epistemic practices for ascribing responsibility may often be ill-suited for helping us latch on to the properties that figure in metaphysical theories of responsibility. They are frequently too coarse-grained (appealing to good and bad characters, indifferent to quality of will or contextual rationality) and sensitive to defeaters (including the usual suite of difficulties for cognition that relies on biases, heuristics, and other pattern-sensitive assessments). If ordinary cognition—cognition susceptible to framing, in-group/out-group bias, and all the

[12] Recall that a practice-first account treats the practice as the explanandum for a theory of responsibility. One could have a practice-first view without thinking that our existing practices are foundational in explaining those practices. For example, according to the Armchair theorist, metaphysical facts and our imperfect epistemology of the relevant metaphysical features explain why our practices have the shape they do. The point here, though, is that any account that takes conceptual-metaphysical elements as authoritative (regardless of whether one thinks practices or concepts are the central explanatory target), will need to explain the mechanics or nature of that authority over the practices as we find them. If one is an antirealist about responsibility facts, for example, it is not obvious that one will share the Armchairer's enthusiasm for the authority of the metaphysical account over the practices as we find them. For recent discussions of the prospects of an antirealist account of responsibility, see Shoemaker (2017) and Wang (forthcoming).

rest—is our best hope for tracking the metaphysically subtle properties of responsibility, one might worry that the entire enterprise is so polluted as to be unreliable.

The foregoing considerations might lead one to a position of *Indifference*, which is usefully thought of as a special variant on the Authoritative Armchair. Both views can allow that departures from one's preferred theory in ordinary practice may very well constitute defects in the practice. If it turns out that everyday ascriptions of responsibility frequently fail, perhaps only rarely successfully tracking the sort of properties identified by the theorist, then the Indifference theorist will shrug and conclude *so much the worse for the practice*. Where the Armchair theorist tries to close the gap, taking on board the thought that ordinary epistemic practices have some connection to our metaphysics, the Indifference theorist is content to abandon that thought, instead counseling disinterest in the messiness of our ordinary ascriptions of responsibility.

Indifference might seem especially appealing if one prefers a concept-first methodology in the theory of moral responsibility. On this approach, the theorist's task is to explicate the conceptual contours of thinking about responsibility, and to adduce the metaphysics involved in the realization or obtaining of instances of responsibility. Confronted with the messiness of ordinary responsibility ascriptions and their disconnection from the sorts of considerations that loom large in the theorist's account, the Indifference theorist might insist that her account is about ideal theory, or about *true*, or *ultimate* responsibility, and that it is not beholden to the unruly details of our psychology and the messy contingency of our local practices. Because the Armchair Indifference theorist's conception of the role of responsibility theory is to provide an articulation of the truths of responsibility – something the practices should reflect – empirical failures to do so are merely a mark of how far we are from morally ideal responsibility practices.[13]

[13] What about an error-theory? One would have to do more work than we have done to show that ordinary biases and heuristics are sufficiently pervasive so as to fund an error-theory of responsibility. However, some error-theorists may be a particular kind of Indifference theorist, holding that metaphysics reveals we lack some feature that our practices could never track, and so much the worse for our practices. Perhaps Strawson (1994) has a view in this neighborhood. In contrast, Pereboom (2014)'s position claims to find unsustainable metaphysical commitments within ordinary practices (via his "four case" argument), which suggests his is not a version of error-theoretic Indifference.

The considerable appeal of Indifference is offset by some distinctive costs. In severing the link between concepts and found practices, the responsibility theorist seemingly abandons the hope of offering an account that is extensionally accurate. Instead of providing a theory that captures actual thought, talk, and practice, the theorist promises a refined or idealized version of thought, talk, and practice. This is, of course, not an unreasonable aspiration; arguably it is one with a distinguished history in philosophy. Still, adopting this strategy produces some distinctive burdens. In particular, the Indifference theorist owes some account of (1) why the theory is still a theory of responsibility (i.e., why this isn't a change of topic), and (2) why this idealization and not some other is the right regimentation away from ordinary practice. In particular, why think it is more likely that ordinary practice is radically defective rather than the philosopher's (radical) theory? In short, one's Indifference to the messiness of ordinary thought, talk, and practice risks producing a philosopher's construct that either does not bear on the practices and statuses that we initially hoped to understand, or whose radical implications become a basis for doubting the plausibility of the theory.

An Indifference theorist might have a variety of things to say about these matters. The theorist might hold that a retreat to idealization or revision away from ordinary thought and talk is worth the cost, or there might be a principled way to animate and defend the preservation of some contours of thought and talk without being beholden to seemingly unprincipled motley of psychological mechanisms that produce responsibility ascriptions. We think that the natural thing to appeal to (to explain a basis for why we can sever dependence on ordinary thought and talk) would be some independent account of the normative authority of responsibility practices. But that is, of course, just to concede the point we have been emphasizing: a metaphysical story without an account of the normative foundations of responsibility is hard-pressed to ease the methodological tension seemingly faced by all theories of responsibility.

Here, we put our cards on the table: the concept of responsibility is not particularly interesting for its own sake, so theories about it are genuinely informative only to the extent to which those theories promise us insight about actual thought, talk, and practice. An entirely novel concept of, for instance, deserved blame that has *no* contact or bearing on the kind of desert or blame we

actually engage in, threatens to be little more than a philosopher's toy.[14] The more one is indifferent about actual practice, and the more one leans in a broadly concept-first direction that attempts to adduce an armchair concept that satisfies decontextualized linguistic intuitions, the less obvious it is that one's theory is any longer connected to the phenomena of everyday responsibility. This is the challenge that lurks for Armchair and Indifference theorists who try to make do with only a metaphysics and no systematic story of normative foundations.

Are there notable alternatives to Authoritative Armchair and Indifference? We think there is at least one further possibility worth mentioning, something we call *Ascriptivism*. Ascriptivism, in the sense at issue here, is the thesis that the best theory of responsibility is one that treats our practices (as opposed to some armchair elucidation of our concept or metaphysics) as its central explanatory target.[15] The ascriptivist regards the function of theory as providing a philosophically satisfying regimentation of our practices, in whatever form we find them. That is, the Ascriptivist holds that any metaphysics of responsibility must ultimately be grounded in actual (if perhaps disappointingly pedestrian) features of everyday responsibility practices. To put it crudely, the Ascriptivist inverts the Armchair Authority's picture: where the latter holds that practice should reflect a metaphysics, the former insists that any metaphysics should reflect the practice.

In the Ascriptivist picture, the explanatory authority of a theory of responsibility is found in its fit with ordinary thought and talk, but most of all in its fit with our enacted practices as we find them. For the Ascriptivist, theory needs to take account of our messy everyday epistemology of responsibility, and indeed, it is beholden to it. The greater the departure from everyday ascriptions of responsibility (including the epistemology involved in such ascriptions), the more work needs to be done to justify that departure. In the absence of clear

[14] Dennett (2006) has written about this sort of problem, modeled on thinking about theories of chess. To put it in the terms of that essay, indifference theorists run the risk of failing to generate truths about chess, the actual game, fixating instead on truths about some nearby but fictional practice of something he dubs "chmess."

[15] We recognize that *ascriptivism* has other usages in philosophy, some of which fit nicely with the thesis as it is characterized here. However, all we mean to pick out with our usage is the idea specified in the text, namely, that proper ontology of responsibility is given by our practices, particularly in our ascriptions, which can include first-, second-, and third-personal ascriptions of responsibility.

justification from those departures, we should conclude that the theory is false, or at least, that there is significant cost to adopting it.

One might think of Ascriptivism as very much in the spirit of Strawson's effort to think about responsibility practices as something organized around our psychologies and the pressures of intra- and interpersonal psychological management. It nicely coheres with a practice-first methodology, emphasizing existing responsibility practices as the explanatory target. However, it adds a specific metaphysical commitment to that methodology, one that grounds facts about responsibility in facts about the practice. In so doing, the epistemology embedded in everyday ascriptions isn't something to be explained away or ignored, but something to be taken as a centerpiece for one's theorizing about responsibility.

As with the other options, we think there are various ways one might further develop the Ascriptivist position, for example, by leaning into the apparently interpretive aspect of ordinary responsibility attributions. On the approach we have in mind, the Ascriptivist might hold that responsibility practices are from top to bottom a matter of contested interpretation about the nature and meaning of one's act, and that there is no prior and independent metaphysical fact about these things, apart from when interpretive practices and conventions have reached (a potentially temporary) equilibrium.[16]

What about the normative authority of this approach? Can the ascriptivist explain why our current practices have normative authority? There is some reason to think the ascriptivist fares no better than anyone else in needing some story about the normative ground for her account. Indeed, this thought echoes a point made about Strawson long ago: a purely descriptive, proto-anthropology (or the in-fact psychology of our responsibility attributions) doesn't by itself answer the normative challenge of why those practices have authority over us, why they bind us if and when they do, and why these practices and not some others are the right ones for us to have (Vargas, 2004). The Ascriptivist, like the Indifference and Armchair Authority theorist, needs a theory of the normative ground of moral responsibility.

In noting these challenges for Armchair, Indifference, and Ascriptivism, we don't mean to imply that they cannot be met. Instead, our ambition

[16] See McKenna (2012a) and Fricker (2016) for theories that might be read as taking steps in this direction, although neither account is obviously committed to the radical ubiquity of interpretation suggested by the picture we have sketched.

has been to call greater attention to an underappreciated methodological challenge for most existing theories of responsibility, one that becomes particularly visible when we consider the role of heuristics and biases in responsibility attributions. The most obvious ways of responding to the divergence between theory and practice each raise non-trivial challenges for a metaphysical theory of moral responsibility that attempts to do without some account of the normative grounds of responsibility. If we are right, a satisfying theory of responsibility will give us both an account of the metaphysical and normative foundations of responsibility. Having both is the most promising way to address the tension generated by pressures for extensional accuracy and normative guidance. In sum, metaphysics is not enough.

ACKNOWLEDGMENTS

We thank Oisín Deery, two referees at *Humana.Mente*, as well as Will Albuquerque, Leo Mauro, and Dan Speak, for feedback that greatly improved the paper.

REFERENCES

Alicke, M. D. (2000, October). Culpable Control and the Psychology of Blame. *Psychological Bulletin*, *126*, 556-574. https://doi.org/10.1037//0033-2909.126.4.556

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin*, *34*(10), 1371-1381. https://doi.org/10.1177/0146167208321594

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal Conceptions in Social Explanation and Moral Evaluation: A Historical Tour. *Perspectives on Psychological Science*, *10*(6), 790-812. https://doi.org/10.1177/1745691615601888

Argetsinger, H. (2022). Blame for me and Not for Thee: Status sensitivity and moral responsibility. *Ethical Theory and Moral Practice*, 1-18. https://doi.org/10.1007/s10677-022-10274-z

Bertram, M. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147-186. https://doi.org/10.1080/1047840X.2014.877340

Brink, D. O. (2013). Situationism, Responsibility, and Fair Opportunity. *Social Philosophy and Policy*, *30*(1-2), 121-149. http://www.journals.cambridge.org/abstract_S026505251300006X

Brink, D.O. (2021). *Fair Opportunity and Responsibility*. Oxford University Press.

Brink, D. O., & Nelkin, D. K. (2013). Fairness and the Architecture of Responsibility. *Oxford Studies in Agency and Responsibility*, *1*, 284-314. https://doi.org/10.1093/acprof:oso/9780199694853.003.0013

Byrd, J. (2010). Agnosticism about moral responsibility. Canadian Journal of Philosophy, 40(3), 411-432. cjphil201040328

Caruso, G. (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lexington Books.

Chevarie-Cossette, S.-P. (2021). Knowing About Responsibility: A Trilemma. *American Philosophical Quarterly*, 58(3), 201-215. https://doi.org/10.2307/48616056

Christensen, W., & Michael, J. (2015). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, 40(A), 48–64 https://doi.org/10.1016/j.newideapsych.2015.01.003

Ciurria, M. (2020). The Mysterious Case of the Missing Perpetrators. Feminist Philosophy Quarterly, 6(2). 10.5206/fpq/2020.2.7322

Cova, F., & Kitano, Y. (2014). Experimental Philosophy and the Compatibility of Free Will and Determinism: A Survey. *Annals of the Japan Association for Philosophy of Science*, 22, 17-37.

Cuddy, A. J.C., Fiske, S. T., & Glick, P. (2007). The BIAS Map: Behaviors from Intergroup Affect and Stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631-648. https://doi.org/10.1037/0022-3514.92.4.631

Deery, O. (2021). Free actions as a natural kind. *Synthese*, *198*(1), 823-843. https://doi.org/10.1007/s11229-018-02068-7

Deery, O. (2021a). *Naturally Free Action*. Oxford: Oxford University Press.

Dennett, D. (2006). Higher-order truths about chmess. *Topoi*, *25*(1-2), 39-41. https://doi.org/10.1007/s11245-006-0005-2

Devine, D. J., & Caughlin, D. E. (2014). Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments. *Psychology, Public Policy, and Law*, *20*(2), 109-134. https://doi.org/10.1037/law0000006

Doris, J. (2002). *Lack of Character*. New York: Cambridge University Press.

Doris, J. M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Double, R. (1996). *Metaphilosophy and free will*. Oxford University Press.

Earp, B. D., McLoughlin, K. L., Monrad, J. T., Clark, M. S., & Crockett, M. J. (2021). How social relationships shape moral wrongness judgments. *Nature Communications*. https://doi.org10.1038/s41467-021-26067-4

Feigenson, N. (2016). Jurors' Emotions and Judgments of Legal Responsibility and Blame: What Does the Experimental Research Tell Us? *Emotion Review*, *8*(1), 26-31. https://doi.org/10.1177/1754073915601223

Feigenson, N., & Park, J. (2006). Emotions and attributions of legal responsibility and blame: A research review. *Law and Human Behavior*, *30*(2), 143-161. https://doi.org/10.1007/s10979-006-9026-z

Fischer, J. M. (2000). Responsibility, History, and Manipulation. *Journal of Ethics*, 4(4), 385-391. https://www.jstor.org/stable/25115787

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Fishbein, M., & Ajzen, I. (1973). Attribution of Responsibility: A Theoretical Note. *Journal of Experimental Social Psychology*, *9*(2), 148-153. https://doi.org/10.1016/0022-1031(73)90006-1

Fiske, S. T., Cuddy, A. J.C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Science*, *11*(2), 77-83. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J.C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878-902. https://doi.org/10.1037/0022-3514.82.6.878

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5-20. https://doi.org/10.2307/2024717

Frankfurt, H. G. (1988). *The Importance of What We Care About: Philosophical Essays*. Cambridge University Press.

Fricker, M. (2016). What's the Point of Blame? A Paradigm Based Explanation. *Noûs*, *50*(1), 165-183. https://doi.org/10.1111/nous.12067

Gerstenberg, T., & David, L. (2012). When contributions make a difference: Explaining order effects in responsibility attribution. *Psychonomic Bulletin and Review*, *19*(4), 729-736. https://doi.org/10.3758/s13423-012-0256-4

Guglielmo, S., & Malle, B. (2017). Information-Acquisition Processes in Moral Judgments of Blame. *Personality and Social Psychology Bulletin*, *43*(7), 957-971. https://doi.org/10.1177/0146167217702375

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford University Press.

Kane, R. (1999). Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *Journal of Philosophy*, 96 (5), 217-240. https://doi.org/10.2307/2564666

Kearns, S. (2015). Free Will Agnosticism. *Noûs*, 29(2), 235-252. https://doi.org/10.1111/nous.12032

Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2019). Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. eLife, 8, e39787.

McCormick, K. (2022). *The Problem of Blame: Making Sense of Moral Anger*. Cambridge University Press.

McGeer, V. (2012). Co-Reactive Attitudes and the Making of Moral Community. *Emotions, Imagination, and Moral Reasoning*, *1974*, 299-326.

McKenna, M. (2012a). *Conversation and Responsibility*. Oxford University Press.

McKenna, M. (2012b). Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism. *Journal of Ethics*, 46, 145-174. https://www.jstor.org/stable/41486955

McKenna, M. (2013). Reasons-Responsiveness, Agents, and Mechanisms. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility*, Vol. 1 (pp. 151-184). Oxford University Press.

McKenna, M., & Van Schoelandt, C. (2015). Crossing a Mesh Theory with a Reasons-Responsive Theory: Unholy Spawn of an Impending Apocalypse, or Love Child of a New Dawn? In *Agency, Freedom, and Responsibility* (pp. 44-64). Palgrave Macmillan.

McKenna, M., & Warmke, B. (2017). Does Situationism Threaten Free Will and Moral Responsibility? Journal of Moral Philosophy, 14(6), 1-36.

Mele, A. R. (2019). *Manipulated Agents: A Window to Moral Responsibility*. Oxford University Press.

Murphy, D., & Doris, J. (2022). Skepticism about Evil: Atrocity and the Limits of Responsibility. In D. Nelkin & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 697-726). Oxford University Press.

Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and Contemporary Problems*, *75*(2), 1-31. https://scholarlycommons.law.northwestern.edu/facultyworkingpapers/218

Nadler, J., & McDonnell, M. H. (2012). Moral character, motive, and the psychology of blame. *Cornell Law Review*, *97*(2), 255-304. https://scholarship.law.cornell.edu/clr/vol97/iss2/3

Nelkin, D. (2005). Freedom, Responsibility, and the Challenge of Situationism. Midwest Studies in Philosophy, 29(1), 181-206.

Nelkin, D. (2011). *Making Sense of Freedom and Responsibility*. Oxford University Press.

Nelkin, D. K., & Pereboom, D. (Eds.). (2022). *The Oxford Handbook of Moral Responsibility*. Oxford University Press.

Nichols, S. (2011). Experimental Philosophy and the Problem of Free Will. *Science*, 331(6023), 1401-1403.

Nichols, S. (2015). *Bound: Essays on Free Will and Responsibility*. Oxford University Press.

Pettit, P. (2018). *The Birth of Ethics*. Oxford University Press.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. OUP Oxford.

Rahimi, S., Hall, N. C., & Pychyl, T. A. (2016, August). Attributions of responsibility and blame for procrastination behavior. *Frontiers in Psychology*, *7*, 1-7. https://doi.org/10.3389/fpsyg.2016.01179

Rai, T. S., & Fiske, A. P. (2011). Moral Psychology Is Relationship Regulation: Moral Motives for Unity, Hierarchy, Equality, and Proportionality. *Psychological Review*, *118*(1), 57-75. https://doi.org/10.1037/a0021867

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Sehon, S. (2013). Epistemic Issues in the Free Will Debate: Can We Know When We Are Free? *Philosophical Studies*, 166, 363-380. https://www.jstor.org/stable/42920274

Shaver, K. (2011). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer New York.

Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press.

Shoemaker, D. (2017). Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to Blame. *Philosophical Review*, *126*(4), 481-527. https://doi.org/10.1215/00318108-4173422

Smith, A. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, *115*, 236-271. https://doi.org/10.1086/426957

Sommers, T. (2010). Experimental Philosophy and Free Will. Philosophy Compass, 5(2), 199-212.

Sommers, T. (2011). In Memoriam: the X Phi Debate. *The Philosopher's Magazine*, (52), 89-93. https://doi.org/10.5840/tpm20115218

Spaulding, S. (2018). *How We Understand Others: Philosophy and Social Cognition*. Routledge, Taylor & Francis Group.

Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187-211.

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Studies*, 75, 5-24.

Suedfeld, P., Hakstian, R. A., Rank, D. S., & Ballard, E. J. (1985). Ascription of Responsibility as a Personality Variable. *Journal of Applied Psychology*, *15*(3), 285-311. https://doi.org/10.1111/j.1559-1816.1985.tb00902.x

Tobias, G., & Lagnado, D. A. (2014). Attributing Responsibility: Actual and Counterfactual Worlds. In *Oxford Studies in Experimental Philosophy: Volume 1* (pp. 91-130). Oxford University Press.

Tobias, G., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2017, December). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122-141. https://doi.org/10.1016/j.cognition.2018.03.019

Vargas, M. (2013a). Building Better Beings: A Theory of Moral Responsibility. OUP Oxford.

Vargas, M. (2013b). Situationism and Moral Responsibility: Free Will in Fragments. In T. Vierkant, J. Kiverstein, & A. Clark (Eds.), Decomposing the Will (pp. 325-349). New York: Oxford University Press.

Vargas, M. (2022). Instrumentalist Theories of Moral Responsibility. In D. Nelkin & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 3-26).

Vilhauer, B. (2009). Free Will and Reasonable Doubt. *American Philosophical Quarterly*, 46(2), 131-140. https://www.jstor.org/stable/20464445

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.

Wang, S. (forthcoming). Response-Dependence in Moral Responsibility: A Granularity Challenge. *American Philosophical Quarterly*.

Watson, G. (1975). Free Agency. *Journal of Philosophy*, (72), 205-220. https://doi.org/10.2307/2024703

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics*, 24, 227-248.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press.

Weiner, B. (2006). *Social Motivation, Justice, and the Moral Emotions: An Attributional Approach*. Lawrence Erlbaum Associates.

Westra, E. (2018). Character and theory of mind: an integrative approach. *Philosophical Studies*, *175*(5), 1217-1241. https://doi.org/10.1007/s11098-017-0908-3

Westra, E. (2019). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, *196*(7), 2821-2846. https://doi.org/10.1007/s11229-017-1575-9

Willemsen, P., Newen, A., & Kaspar, K. (2018). A new look at the attribution of moral responsibility: The underestimated relevance of social roles. *Philosophical Psychology*, *31*(4), 595-608. https://psycnet.apa.org/doi/10.1080/09515089.2018.1429592

Wolf, S. (1993). *Freedom Within Reason*. Oxford University Press.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283-301. https://doi.org/10.1016/j.cognition.2005.05.002

Zell, E., & Alicke, M. D. (2009). Contextual neglect, self-evaluation, and the frog-pond effect. *Journal of Personality and Social Psychology*, *97*(3), 467-482. https://doi.org/10.1037/a0015453