


9-1-2010

Responsibility in a World of Causes

Manuel Vargas

University of San Francisco

Follow this and additional works at: http://digitalcommons.brockport.edu/phil_ex

 Part of the [Ethics and Political Philosophy Commons](#), [Metaphysics Commons](#), [Philosophy of Mind Commons](#), and the [Philosophy of Science Commons](#)

Recommended Citation

Vargas, Manuel (2010) "Responsibility in a World of Causes," *Philosophic Exchange*: Vol. 40: No. 1, Article 4.
Available at: http://digitalcommons.brockport.edu/phil_ex/vol40/iss1/4

This Article is brought to you for free and open access by Digital Commons @Brockport. It has been accepted for inclusion in Philosophic Exchange by an authorized administrator of Digital Commons @Brockport. For more information, please contact kmyers@brockport.edu.



The College at
BROCKPORT
STATE UNIVERSITY OF NEW YORK



Manuel Vargas

Responsibility In a World of Causes

Manuel Vargas

1. The issue

Our lives intertwine with praise and blame in ways both simple and complex. If you ask me to pick up your child after work, and I fail to do so even after promising that I would, you will very likely be angry at me. This is not the irritation we experience when the weather is chilly or when we don't win the lottery. This attitude is directed at a specific person, ordinarily a (somewhat) aware and responsive being. That is, you are presuming that I am a particular kind of entity, a responsive, choice-making agent. It is partly in virtue of this fact that your reaction of irritation has its distinctive flavor; unlike bad weather or unreliable lottery drawings, I can knowingly and willingly bind myself with commitments that we both take to license blaming when I fail to live up to them. In short, I am a special kind of entity—a responsible agent. In virtue of my conduct, I can be worthy of praise and blame.

These moralized reactions are not limited to interpersonal relationships. People spend years in prison, beyond what is plausibly useful for rehabilitation, and usually to the exclusion of victim restitution, out of an oftentimes inchoate or implicit conviction that criminals *deserve* punishment in light of their culpable failure to exercise their agency in the right ways. Indeed, it is difficult to make sense of the impulse to execute criminals without appeal to some notion of deservingness bound up in the idea that the criminal is morally responsible for his crime in some deep way. So, moral responsibility, the idea of praiseworthiness, blameworthiness, and associated notions of merit are all important parts of our shared lives.

This picture, however, is threatened by a very familiar chain of reasoning. The reasoning goes like this: if everything is caused, no one is genuinely free, and thus, no one can be genuinely morally responsible for anything. It is a very old argument.¹ Versions of it have been banging around in the Western intellectual tradition for millennia, and every age has its favorite formulation of it. Perhaps the most common contemporary incarnation of the argument frames things in terms of a threat from science. You don't have to look very hard to find people who will say, for example, that neuroscience or biology or scientific psychology has shown that we are not beings that act independent of the causal, physical order.² We are told that modern science has sufficiently mapped out the underlying psychological, biological, or chemical roots of our behavior so that we can say with confidence that free will does not exist. Consequently, all those notions we associate with free will—ideas of praiseworthiness, blameworthiness, and

merit—are in trouble as well. So we are told.

It is easy to overstate the conclusion of the familiar chain of reasoning. It does not claim that we make no difference to what happens. At least some of what happens does so precisely in virtue of our participation in the causal sequence. One's sordid extramarital affair does not happen without the causal chain working through one's body parts. The point of the argument is that one should not be blamed for what one does, because it is part of an inexorable causal chain extending back in time prior to the existence of any humans. This chain of reasoning also does not claim that we must let all the criminals go. We would still have practical reasons to separate incorrigible cases from the rest of us.³ However, our treatment of criminals might seem closer to quarantine and rehabilitation than punishment in any conventional sense.

Still, the ramifications of the familiar argument are significant. They are significant enough that accepting them may require a steely backbone, one might think. At any rate, this is surely part of the reason why the familiar chain of reasoning is oftentimes presented in something of a macho, gruff tone, claiming the authority of science.

The gruff tone can be important. It is, I think, intended to convey the conviction that only someone with a weakness for namby-pamby, feel-good myths about humanity's special place in the world would resist the conclusion of the familiar chain of argument.⁴ Do not let the gruff tone mislead about the state of scientific and philosophical inquiry into the subject matter. There is plenty of sincere, thoughtful disagreement about both empirical and philosophical matters. The conceptual and empirical issues tied to free will are very, very difficult, and we are unlikely to get informed consensus about them any time soon. Moreover, the basic issues extend beyond any particular scholar's domain of expertise: there is the scientific bit, sure, but there is also the moral bit and the matter of our concepts what we mean, what they are committed to, and whether they permit of transformation or rehabilitation in the face of an uncooperative world. Having decisive reasons to think something in any of these domains does little to settle the matter in the other domains. In sum, we do well to be suspicious of any sweeping claims that imply a special authority on the part of the claimant.

My aim here is to focus on only one part of the familiar chain of argument.⁵ I wish to focus on the nature of moral responsibility, bracketing larger issues concerning free will. On the matter of moral responsibility, I think we can make a kind of progress by asking what's at stake when we ascribe moral responsibility. What does keeping track of it allow us to do? Is there anything that might make sense of these practices, independent of whatever else turns out to be true of free will and the causal order? My idea is this: if we have a plausible account of the structure and nature of moral responsibility, we can describe powers and purposes that make sense of our practices, more or less as we find them. My

aim here is to sketch how that story might go, and thus, to offer some reason for thinking that the familiar chain of reasoning does not go through. To put it bluntly, there is good reason to think we are morally responsible even if we accept a broadly scientific worldview. The trick is to be clear about the kind of thing responsibility must be. Once we are clear about *this*, the threats from determinism, reductionism, and the like subside.

2. Terms and assumptions

Philosophy is hard enough without the murkiness that comes from failing to clarify one's terms and presuppositions. So, I'll start by defining a few key ideas. "Moral responsibility" is obviously an important term here, and one I'll be using in a somewhat specialized way. We sometimes use the word to pick out obligations, as in "you failed to meet your responsibilities." I'm not using it in the sense where it is a synonym for "obligations." Instead, I'm using it to pick out the property of, roughly, being morally praiseworthy or morally blameworthy.

(I say "roughly" because what is at stake is actually a fairly large set of characteristic practices, attitudes, and judgments, of which conclusions of worthiness of praise and blame are only a part; the fuller story would involve reactions including indignation and gratitude, outrage and envy, and behaviors from avoidance to congratulations. These latter things can include more than what we sometimes think of as praise and blame in the narrow sense. They are, I think, broader than the canonical "reactive attitudes" detailed by P.F. Strawson.⁶)

It is important to notice that the notion of responsibility I am interested in is distinct from, for example, legal or causal notions of responsibility. Indeed, a remarkable feature of moral responsibility is its seemingly person-centric focus. A hurricane might be causally responsible for flattening a trailer park, but it is not *morally* responsible for having done so. And, within the realm of interpersonal assessments of responsibility, moral responsibility need not entail legal responsibility. We might think a person is blameworthy for failing to provide a sympathetic ear to a distressed friend, but it does not follow that the bad friend has broken any laws.

One word I'll be throwing around a bit is *agent*. By "agent" I mean to pick out entities that are capable of acting, or undertaking courses of intentional action. These are beings that paradigmatically act on the basis of beliefs, desires, and intentions. So, you are an agent. I am an agent. Your cat is probably an agent. Hurricanes are not agents, and neither are subatomic particles or single-celled organisms.

Finally, I'll be helping myself to some assumptions that I won't try to defend here, but that will be operating in the background of what follows.

Continuity with nature: We are continuous with and a part of the larger natural world, and are ultimately material or physical beings. The presumption that we are part of the rest of the natural order is also a presumption against appeals to souls, ectoplasmic substances, radical independence from the causal order, or other spooky properties ... unless one musters a really, really good story.

The viability of moral discourse:

(A) Morality talk is not like talk of unicorns or phlogiston; that is, there is something in the world that we are usefully and rightly getting at when we talk of morality (what, exactly, that comes to I want to leave open).

(B) Moreover, for language or thought to be moral it need not rely on high-blown language. If you think I'm a jerk because I'm insensitive, that's a piece of moralized thinking—at least to the extent to which you think people ought not be jerks or ought not conduct themselves in jerk-like ways. Moral vocabulary shifts over time, and much of our interpersonal judgments are moralized. You and I tend to think there are better and worse ways to be, better and worse ways to act, and that one, all things considered, ought to act in the better ways relative to the available choices. When we condemn or praise people in light of those standards, we are typically expressing moralized judgments.

With the main terms and assumptions clarified, we can now start to push forward. I'll begin by saying a bit about what I mean when I talk of “the work” of our concept of moral responsibility. Then, I'll articulate some options available to us, and I'll go on to motivate the picture I favor.

3. What's the work of our concept of moral responsibility?

We use concepts to carve up and categorize parts of the world. Concepts do a kind of work for us: they may demarcate one thing from another. Relatedly, they identify a collection of (we suppose) interrelated inferences we can make about things. So, for example, the concept of ‘car’ does the work of capturing a subset of transportation-related thoughts we have: a car is likely to have a motor, it is likely to travel on wheels, it is likely to be used as transportation for small groups of people on paved surface streets, and so on. The work of the concept of ‘llama’ is to specify a cluster of inferences regarding a particular kind of mammal in the

world; the work of the concept of ‘felony’ is to specify a set of inferences regarding a legal status; the work of ‘touchdown’ is to specify a set of inferences regarding a scoring event internal to the game of American football. We can say that the *work of a concept* is defined by its primary, general inferential role.

Nothing in this picture requires that all concepts necessarily do an identifiable piece of work for us. Perhaps there are concepts that do not have any practical or inferential role in our thinking or practices. Nor does this picture presume that the work of a concept is univocal. Indeed, some concepts are very highly contested, subject to substantially different individual conceptions of what counts as the principal work of the concept. We might think of the concept of marriage as regulating, for example, (1) a legally sanctioned and privileged relationship, (2) a religious sacramental relationship, (3) a property relationship, (4) a privileged emotional relationship, or (5) conditions of socially sanctioned sexual intercourse. If the conceptual work of ‘marriage’ is different for you than it is for your neighbors, you are likely to disagree with them about the propriety of different ways of extending the concept. Much of the recent dispute concerning gay marriage in the U.S. is rooted in disagreements about the work of the concept of marriage. One’s conception of the work of the concept plays a big part in the ways in which one is willing to think a novel proposed usage is an apt one. So, disagreement about the work of a concept is both possible and sometimes actual.

Finally, the fact of a concept having a role does not guarantee that the concept as we have it or use it does a good job of fulfilling that role. The ancient concept of “blood purity” might have been like this. Presumably, the work of the concept as its users conceived of it was to demarcate real differences in human kinds, tied principally to lines of inheritance. The concept failed to do its work in two ways. First, there was nothing in the world that neatly corresponded to blood purity as it was ordinarily conceived. Second, what work the concept did in practice fell considerably short of the role that it was generally understood to have: rather than tracking real, essential purity of a blood line, it tracked various contingent social and class differences. So, the work of a particular concept might not be well executed by the concepts we have: the concept could be defective and the world might not cooperate.⁷

Despite the inevitably abstract talk of conceptual work, these ideas allow us to make sense of the notion that it can be useful to ask about the conceptual role of moral responsibility. As a first pass, I’d say that the work of the concept of moral responsibility has to do with marking a set of inferences about differential moral praiseworthiness and blameworthiness. People who knowingly and intentionally do the wrong thing deserve condemnation. People who do the right thing deserve approval. Keeping track of when people deserve praise and blame is the work of the concept of moral responsibility. It is, in some sense, why we rightly have a concept of moral responsibility.

(To clarify: this is *not* a historical claim. It is a claim about why, whatever its history, we now are right to have such a concept.)

Given this picture, one where the work of the concept is tied to differential assessments of praise- and blameworthiness, the challenge for us is to consider whether there are any good reasons to suppose that people can be worthy of praise and blame. At this point, the temptation may be to return to familiar issues of free will. But let us postpone those concerns for a moment. Instead, let us ask why we should care about praise and blame. My thought is this: if we know what work the concept does (and I think we do), and we have some grasp on why we should care about the work that it does, this gives us a vantage point from which to pick and choose among various candidate accounts of the conditions of moral responsibility. To do this, it helps to know some more about the thing in the world we are trying to track. What is this blameworthiness stuff, anyway?

4. First pass at responsibility-as-blameworthiness

I have suggested that one way to understand the relevant sense of moral responsibility is in terms of blameworthiness and praiseworthiness. To get some sense of what we are after, it may be helpful to begin with a philosophically reviled account of blameworthiness: the classical consequentialist account of moral responsibility. An idea fundamental to the classical consequentialist account was that praise gets us to do good things and blame gets us to avoid doing bad things. So, we have reason to care about moral responsibility inasmuch as we care about getting people to behave in the right ways and getting them to avoid behaving in the wrong ways. On this picture, then, to be blameworthy is just for it to be the case that, given that you did something wrong, were I to blame you, it would lead you to behave in the right ways.⁸

In its classical form, there are a number of important problems for this account. For example, it seems to do a bad job of explaining lots of ordinary cases where we are making judgments of praiseworthiness and blameworthiness. In many cases, we are making judgments that look like they have no hope of influencing anyone, both as a matter of what I might believe and as a matter of what is actually possible, apart from my beliefs. If I admiringly praise a dead relative for her dedication to the fight against the now-extinct disease of kuru⁹, I need not presume to be influencing that relative. Influencing the dead is, I suspect, substantially beyond my limited powers of persuasion.¹⁰ It is also implausible to suppose that my judgment is an attempt to influence, for example, myself, or my acquaintances, to fight kuru. Kuru seems to have become extinct a few decades back. I suppose we *could* imagine that I might be motivated to praise my dead relative for her fight against kuru as part of an elaborate attempt to get me or people I know to fight diseases in general, or even to just dedicate some part of

our lives to a worthwhile cause. But this certainly doesn't seem necessary for me to make the judgment that my relative was praiseworthy for her endeavors. Yet, the classical consequentialist story seems to require that my particular praising or blaming have these effects *in this particular case* for us to rightly care about praiseworthiness and blameworthiness. So, whatever it is that we are tracking with our concept of moral responsibility, it looks like the classical consequentialist story misses the mark.

A more promising account holds that what we are looking for isn't susceptibility to praise and blame, but rather, a kind of relationship between agents and what moral reasons there are. Roughly speaking, if you are doing a good job responding to moral reasons, you are praiseworthy. If you are doing a bad job of responding to them, you are blameworthy.

That's all a bit compressed. I've said that praiseworthiness and blameworthiness are a kind of relationship to reasons. As we will see, though, that relationship is deceptively complex. To see how, we need to unpack some ideas implicit in our judgments of blameworthiness, for they specify some aspects of that relationship that, while simple enough to state, pick out complex features of the world. (I'll focus on the blameworthiness case, but I think much of the basic picture can be extended to praiseworthiness).

Thoughts like "Joe is blameworthy" are comparatively minimal in their explicit conceptual commitments. There are essentially two ideas at work in this sort of thought: that the person we are responding to is a being of the right sort to regard in responsibility-assessing ways (what I will call a *responsible agent*) and that he or she has met, exceeded, or violated some norm that we regard as justified. For each of these subsidiary judgments to be *true* or *justified*, well, that's the part that can be convoluted. But notice this basic situation is not unique to our thoughts about moral responsibility. There are plenty of judgments we make in everyday life that have the structure of being on the one hand cognitively minimal, (in terms of what you must suppose to coherently entertain the thought), and on the other hand, pretty robust in terms of what features the world must have for the thought to turn out to be true. The thought that I married the right person is pretty conceptually thin—something like "this particular person to whom I am married to is a sufficiently good fit for me in all the ways that matter." (We could nitpick about whether "the right person" indicates that there can be only one person or whether it instead is an implicit notion of sufficiency of rightness, but let's leave this to the side.) What would make this thought that "I've married the right person" true, though, is presumably a pretty complex set of facts about the people involved. These facts might include such things as compatibility of personality, values, aspirations, and the like, but also a means of communication, and facts about the existence of marriage practices and the permissibility of such practices.¹¹

Let's return to judgments of blameworthiness. Such judgments seem to be

frequently backward-looking but cognitively thin. Saying what needs to hold for the judgment to be true thus requires theories of two things: (1) the nature of responsible agency, and (2) the nature of the responsibility norms. I'll take them in order.

5. Reasons responsiveness

Before sketching a theory of responsible agency, I want to say a bit about reasons and what responsiveness to them can consist in.¹²

A reason is, roughly, a consideration that counts in favor of something. And, I take it, a core feature of our self-conception is that we do things for reasons. We don't *always* act for good reasons. And, surely, there are often reasons for us to do things of which we are blissfully unaware. And, sometimes, we end up acting for reasons that we cannot articulate, or of which we may be systematically unaware. Nevertheless, in a wide range of cases we are concerned to act for reasons.

I don't want to suppose an overly intellectualistic conception of reasons. Reasons might well depend to a very large degree on our emotions, preferences, or desires. Nothing about this picture is meant to preclude the possibility that reasons are ultimately dependent on aims or affect. For present purposes, I just want to help myself to the idea that we can usefully talk of reasons, and that the utility of talking this way emerges ubiquitously, for even comparatively rudimentary systems.

So, for example, the presence of a nut in the waning days of autumn usually generates a reason for a squirrel to figure out a way to get that nut safely to his or her den. All that is required is the idea that it makes sense to speak of a squirrel having aims and that there are features of the world that are relevant to the attainment of those aims. Indeed, these very features seem quite plausibly present in entirely simulated worlds. We might imagine a video game in which there are computer-controlled characters that have reasons to respond to their artificial environments in various ways. A computer-controlled character that is, for example, hunting for dangerous aliens will have a reason to proceed cautiously when there is evidence of aliens in that artificial environment. The lesson is pretty simple: to believe that there are reasons, all we need to think is that there can be things relevant to an agent's aims.

Note, though, that agents can have variable sensitivity to reasons. We might imagine that some computer-controlled characters are more and less able to recognize evidence in their artificial world (if you don't play video games, just trust me on this). And, similarly, we might imagine that some squirrels are better than others at recognizing that there are acorns out there.

But the success of our simulated agent or our imagined squirrels doesn't just depend on recognitional capacities. Success also depends on the agent acting on

that information in the right way, which is something we might call a volitional capacity, or a capacity for self-governance. A squirrel that is an excellent acorn-detector but acorn-phobic will do badly at the business of acorn collection. So will a squirrel that is excellent at detecting acorns but completely apathetic about pursuing them. So, the ability to recognize reasons for action is of limited utility by itself—it is absolutely crucial that it be connected to a further ability to act on the detected information in the right way. At least from the philosophical armchair, there is no reason to suppose that excellence in reason detection is necessarily coupled with excellence in being appropriately moved. I know I have good reason to go for a run later today. I also know I won't do it.

This is a relatively simple picture, relying on two key ideas: recognition of reasons and implementing a suitable course of action in light of the significance that reason has for the agent. Things are, of course, much more complicated in the real world. For example, agents find themselves infested with desires, values, and interests that can operate at cross-purposes in the near and long terms. Acorn hoarding is less important if a predator is around. And, indeed, the environment interacts with our complex recognitional, deliberative, and volitional systems in a multitude of ways.

For example, I might normally be substantially resistant to having a third shot of tequila, viewing it as a bad idea—at least until I've had something to eat. When my brother comes into the room, though, the difficulty of resisting that temptation goes up. (Perhaps he makes drinking even more fun; perhaps his enthusiasm for tequila is contagious; perhaps I simply need to be liquored up to get along with family—imagine any scenario you like.) Suddenly, having that third shot, maybe even a fourth shot, doesn't seem so unreasonable. And note what is happening—my ability to resist the temptation can change in light of two different forces, both originating from that single change in the situation. First, any brute desire to have a shot can go up, or alternately, where there was no desire for a shot there now is such a desire. So, the configuration of my desires might change, making it harder to resist temptation. But second, and perhaps more nefariously, the change in environment can hijack my evaluation of what counts as a good idea. That is, not only might I want the shot more, my sense of what counts as a good idea, including my ability to attend to reasons of tequila moderation, might have changed. Note further that this change can be independent or dependent on changes in my desires. That is, my sense of what constitutes a good idea might change precisely because the force of new, pro-tequila desires overwhelms my evaluations. But, even if those desires stay stable, my evaluations might nevertheless change in light of the change in situation.

Now in the case I've just described, I'm pretty conscious of the phenomenon of sibling-triggered enthusiasm for tequila. But it isn't hard to imagine parallel cases in which we are not conscious of the ways in which the environment changes

our volitional powers, making it harder (or easier) to translate recognition of some reason into the right sort of behavior. Notice what this means, though: our volitional capacity, the capacity to move ourselves to act in accord with a reason we recognize, in some sense doesn't *just* depend on us. It also depends on the context or circumstances of action. There is a relatively straightforward way in which it makes sense to think that at least some of our ordinarily understood agential capacities depend on the world for their powers.¹³

This is, I think, an important point that has been for too long underappreciated by philosophers who think about agency. We have tended to focus almost exclusively in understanding agents in relative isolation from the environments in which we act, to commit ourselves to an implausible form of methodological individualism. By this I mean that the powers of agents are almost always presented and discussed in a way that makes it seem as though if we wish to understand what capacities any given agent has, all we need to know are facts about the agent. The circumstances of action only matter as inputs on fixed capacities. In contrast, the picture I'm suggesting is one where our capacities themselves are somewhat malleable, subject to situational pressures that we would do well to understand and incorporate in our accounts of agency.

What we have, then, are three ideas: the capacity to recognize reasons, a corresponding capacity for self-governance in light of those reasons, and the idea of context-dependence in at least some of our capacities. I now want to add a fourth idea to the mix, the idea that we can differentiate between varieties of reasons. There are, for example, prudential reasons. These are reasons whose significance depends solely on what is of benefit to the agent. So, for example, it would be prudent for me to not bore you, and it would be prudent for me to end on time, or even to end early. It would be wildly imprudent to fail to give this talk at all. So, there are reasons of prudence. But there are also reasons of other varieties. There are presumably legal reasons, that is, considerations grounded in various aspects of the law. And, there are presumably aesthetic reasons, or considerations that count in favor of one or another artistic choice. These reasons would be indexed to whatever it is that gives rise to aesthetic properties, which might include things such as visual, tonal, or linguistic properties, but also artistic traditions, personal preferences, or the collective sensibilities about what is new, worthwhile, or recognizably norm-breaking. It is plausible that there are also moral reasons. For people of a particular generation, this may sound odd. We sometimes regard explicit talk about morality with skepticism or the sort of raised eyebrows we save for public expressions of wild-eyed religiosity. But all I mean by talk of moral reasons are those reasons whose significance depends on morality, whatever that comes to. So, for example, if morality is purely conventional, then moral reasons will depend on conventions. If, on the other hand, morality depends on, say, the will of God, or the compatibility of an agent's intentions with the categorical

imperative, or the issuances of an ideal observer, then moral reasons will depend on that. The details are not, for present purposes, important.

What is important, though, is that there can be moral reasons, and that agents can vary in their abilities to recognize such reasons and to respond to them accordingly. The variation operates along several dimensions, including recognitional and volitional sensitivity, but also in terms of how these things operate across contexts. And, of course, these variations hold across particular agents. This should not be surprising. We all know people who are incredibly responsive to the suffering of others, highly polished at avoiding chagrin-inducing social interactions, or particularly skilled at providing good cheer to those in need. And of course, we also know people who seem perhaps pathologically insensitive to the needs of others, blind to what gives offense, and immune to suggestions that they better regulate their behavior around others. We call these latter people professors. The point, though, is that people have varied capacities for recognizing and responding to moral reasons, and that the capacities vary across individuals and circumstances.

So, we've got four ideas on the table connected to the role of reasons and agents: (1) a capacity to recognize reasons, (2) a capacity to respond in appropriate ways to reasons, (3) the situation-dependence of those capacities, and (4) the possibility of moral reasons and the idea that people can be better and worse at responding to such reasons. I now want to sketch a way to build a promising account of responsibility out of these four ideas.

Putting all these pieces together, we can say the capacities of responsiveness to moral reasons are what matter for moral responsibility. It is their presence or absence that—in conjunction with some other things¹⁴—underwrites the propriety of moralized praising and blaming. These capacities do not operate in isolation from the world. They are subject to enhancement and degradation in light of a wide range of factors, including the situations in which agents act and the cultural scripts or narratives that structure how agents understand and respond to the situations they are in.

6. Norms, and how it all fits together

As I noted above, the comparatively thin content of our judgments of blameworthiness appeal to some ideas (i.e., that there is a special kind of agent involved, that there was a violation of some particular kind of norm) that require a theory of what these things could be. I've given the outlines of an account about the relevant form of agency, but we still need a story about the norms of blame. What gives content to these norms? What makes these norms—whatever they are—the right ones? To answer *this* question I need to introduce two ideas that can initially seem a little bizarre, but that do some important work in the

business of theory building. The first is the notion of an ideal observer.

An ideal observer is just that— an observer that is, in some relevant respect, ideal. The ideal observer is a theoretical construction that helps us wrap our heads around what properties or characteristics are of relevant interest. So, for example, suppose you are given the task of saying what constitutes an excellent philosophy talk for a particularly diverse audience of laypersons and scholars. Given the range of expertise in the audience, it might be very difficult or seemingly impossible to say how that talk should go, beyond platitudes about being engaging and informative and accessible. But we could also say that the best talk to give would be one that an ideal observer, aware of everyone's interests and tolerances would recommend to the speaker, given the speaker's aims. Now this reply isn't going to directly settle the matter of which talk you should give. But it might help you characterize what sorts of things would make a good talk because it gives you a standpoint from which to think about which features of a talk matter or don't. And, it might illuminate how some of those things could fund the truth or falsity of claims about what would be a good or bad talk to give.

Now consider the case of norms of praising and blaming. What determines which norms are the ones that we properly bring to bear on responsible agents? There are, after all, lots of ways we can describe how norms of when and how to praise and blame might go. As a first pass, we can say that the right norms are the norms that an ideal observer would select for responsible agents. As stated, this is clearly inadequate. We need to know what sorts of things would drive the selection of norms picked out by our ideal observer. But here our story of responsible agency can help us out.

Recall that a distinctive feature of responsibility ascriptions is that they target certain kinds of agents and not others. Now, we can say exactly what kind of agent it is that we are targeting and why: we're interested in that special class of agents that can recognize and respond to moral reasons precisely because that is what moral responsibility is about. That is, moral responsibility is about our relationship to how well or poorly we are tracking what moral reasons there are. So, when we think about what sorts of praising and blaming norms an ideal observer would be selecting for creatures like us, the most promising answer seems to be norms that are, in some central way, tied to the fact of our being creatures that can recognize and respond to moral reasons. So, what we need is some way of tying the kind of agency at stake in responsibility claims to a set of norms that are relevant.

We have now arrived at the second idea of what generates the particular content of the responsibility norms: indirect consequentialist justification. The idea is that what the norms of blameworthiness will be is that collection of norms that, if internalized by agents of the relevant sort, would be the ones that in fact, over time, do the job of getting agents to better recognize and respond to

what moral reasons there are. Or, to put the point a bit differently, what justify our web of responsibility-characteristic practices, attitudes, and judgments are, roughly, the effects of such a system on creatures like us, over time. That is, these practices foster in agents like us those remarkable capacities for recognizing and appropriately responding to moral reasons. Consequently, the correctness of judgments of responsibility are settled by those facts, i.e., the facts about the package of norms an ideal observer would select for us, given various facts about our internalizing the norms and acting on them, and given the observer's aim of selecting a package of norms that effectively cultivates in us moral reasons-responsive agency.

We don't want to fall into the bad old trap of supposing that the blameworthiness norms require that we always attempt to influence each other in some special way, or that they require that we suppose that someone only counts as blameworthy if they are susceptible to blame. The indirect part of the consequentialist justification is important because it buys us considerable flexibility in what the content of the responsibility norms can look like. Particular first-order norms (e.g.: praise people for selflessness, blame people for duplicitous infidelity, etc.) need not make any appeal to consequences at all, whether in the specific or general case. Indeed, it is quite plausible to think that the best and most effective set of norms will include many that are exclusively backward looking. This is compatible with the account; we only appeal to consequences at the level of the effects of having a diverse set of internalized norms, many of which make no appeal to the consequences.

So, to sum up the picture, to judge that some person X is responsible for some action A is to judge that X is an agent of the right sort—a responsible agent—and thus subject to a distinctive set of norms concerning responses to X's violation, meeting, or exceeding some moral norms. The structure of the blaming norms is connected to what is distinctive about responsible agency, namely, the capacity to respond appropriately to moral reasons. And, we can think of the particular details of those norms as being settled by the ideal observer with full information about how implementation of various sets of possible blaming norms might go. That is, the norms are those such an observer would select for the collection of agents, given the aim of fostering reasons-responsive agency, and the aim of enhancing that sensitivity across contexts of action. Or, *we might just say that moral responsibility is about building better beings.*

There is, of course, a lot more to say about each of these pieces and about all the details. Nailing down philosophical details is a painstaking task, and some of the details I've glossed over are particularly painful. Still, we're now in a position to see how some pieces might hang together. Earlier, I claimed that moral responsibility is really about our relationship to how well or poorly we are tracking what moral reasons there are. Now we can see what that comes to,

for we have some account of what responsibility consists in, and why it should matter. I began by suggesting that the work of the concept of moral responsibility is to mark differential assessments of praiseworthiness and blameworthiness. A theory that explains how we might do *that* (i.e., how we might rightly mark these distinctions in the world and on what basis) has claim on being a philosophical (normative, prescriptive) theory of moral responsibility. Different accounts of this, however, will come with better and worse resources for explaining the point of keeping track of moral responsibility, or the point of retaining the concept in the face of calls to jettison it. On the account I have offered, the point of recognizing and responding to distinctions in praiseworthiness and blameworthiness is the cultivation of a special form of agency, one that is sensitive to moral considerations. To the extent to which we care about morality and our relationship to it, we have an investment in our being better agents of the sort that our practices, attitudes, and judgments of responsibility are properly organized around cultivating. So, it seems, we can explain the importance and basic structure of moral responsibility in terms that do not appeal to spooky powers or features of agency we otherwise have reason to doubt.

7. Breaking the chain

What does all of this mean for the familiar chain of reasoning with which we started?

Notice that the account of responsibility I have offered is fully compatible with any standard story about the details of the causal order. We can be fully caused, even deterministically caused, and still morally responsible. I do not claim to have captured our ordinary conception of what is required for moral responsibility. Perhaps it is true that we typically suppose that moral responsibility requires powers that are incompatible with a broadly scientific conception of human beings. Nevertheless, what I have offered is an account of the work of the concept, and how that work can be done without appeal to anything other than features of agency that we plausibly have. My claim is that regardless of whether the world is deterministic or even just fully causally ordered, we can make sense of the underpinnings of moral responsibility and the kinds of capacities it requires.¹⁵

What this means for the rest of the familiar chain of reasoning is not clear. If you thought that free will was something like the self-governance or control condition on moral responsibility, then it looks like this account shows how you can have that regardless of whether or not larger-than-nano-sized objects are mostly deterministic in their operations. Second, the account also seems to block some standard worries about the idea that human agency is reducible to lower level properties, whether they be brain states or something else. On this account, there might well be a fully adequate reduction of all interesting features

of agency. Nevertheless, what the account points to are those features of agents that, reducible or not, properly drive responsibility ascriptions. So, again, the account shows how the business of judging and holding one another responsible is insulated from worries that this part of our self-image will be undone by a broadly materialistic or physicalist account of persons. My hope is that by showing that responsibility can stand on its own feet, the familiar chain of reasoning will seem a good deal less threatening. At the very least, we will have made some trouble for those who would gruffly conclude that a consequence of modern science is that things like blameworthiness are illusory, or at best, useful fictions.¹⁶

8. Extending the picture

Before concluding, I'd like to briefly draw out some of the implications of this account of responsibility. In particular, I'll say a bit about what this account tells us about children, psychopaths, and responsibility under various forms of mental illness.

Children raise some interesting questions for a theory of moral responsibility. Intuitively, young children are paradigm instances of agents that cannot be morally responsible. Somehow, though, they come to be fully responsible agents by the time they are adults. A theory of responsibility should be able to say something about that change. It is notable how much of childrearing we direct at socializing children to be sensitive to moral norms. But for all that, much of it is feigned, at least at first. That is, we might praise and blame children to help inculcate in them the various responsibility norms, but in doing so we need not really think that children are ubiquitously responsible. Still, at least sometimes, we really do seem to hold our children responsible. I remember being caught by surprise at how much anger and disappointment I felt when one of my older kids was particularly cruel to her comparatively defenseless and guileless baby sister; and, at least in talking with other parents, this sense of outrage is not as rare as we would like. What this points to, though, is an interesting feature of how responsible agency can, in some important sense, grow over time. That is, we can never rightly think of very young agents—infants—as sensitive to moral considerations. It takes time before they are able to recognize what we think of as moral considerations. And, even when it does happen, it is usually in a piecemeal fashion, limited to particular contexts. But, in those contexts we have a genuinely responsible agent.

So, we can explain why—in some comparatively limited cases—it makes sense to hold children responsible, and why, in other cases, it does not. The capacity for moral responsibility is not some unified, global, cross-situationally stable capacity that is either had or not. Over time, though, the pattern of feigned praising and blaming ordinarily helps such agents expand their sensitivity to moral considerations, or, at least, those considerations regarded as moral in that

society.

Psychopaths are another intriguing limit case for a theory of moral responsibility. Unfortunately, it is beyond the scope of this paper to discuss all the interesting empirical and conceptual aspects of this category, including some very interesting issues about what constitutes psychopathy—for example, the category is not in the DSM, and among its earliest predecessors was a category with the suggestive label of “moral idiocy.” But, on one plausible construal of the category, two of the most salient features of psychopaths include: (1) the inability to distinguish between what psychologists call conventional and moral harms, and relatedly, (2) diminished emotional responses to witnessing harm. There is some evidence that a robust range of moral concepts cannot be acquired without our typical reactions to witnessing harm and injury. Without the resulting moral concepts, it looks like there is no way to satisfy the detection condition on responsible agency. If all of this is right—although matters are complicated—then in domains where psychopaths are constitutionally incapable of perceiving moral considerations of the relevant sort, we cannot rightly regard them as responsible agents, and thus, as properly subject to judgments that they deserve moral praise and blame.

There are several complicating factors here, though. First, it is not clear that the psychopath’s inability to recognize a wide range of moral considerations constitutes a uniform excuse from all aspects of responsibility. For example, it might be possible for the psychopath to recognize moral considerations of some restricted sort, or to recognize considerations that overlap with, for example, considerations of prudence or self-interest. If so, then the matter of whether or not a psychopath is properly evaluated in terms of the responsibility norms becomes, as in the case of children, a matter of partial or intermittent suitability. Second, it is unclear whether there are non-standard ways to bootstrap psychopaths up into something like conventional moral cognition. That is, even if the underlying features that give rise to psychopathy preclude the ordinary route to acquiring moral concepts, it is unclear whether there are non-standard routes to acquiring moral concepts, or reasonable analogs of them.¹⁷

One notable implication of the theory is that agents can have relatively isolated deficiencies that undermine responsibility in some contexts but not others. There are two take home points that follow from this last bit.

First, when we think about the possibility of responsible agency under conditions of severe forms of mental illness, we shouldn’t suppose that what is at stake is some sweeping exculpation or sweeping judgment of responsibility. The cognitive and affective impairments of various pathologies can intersect with the demands of moral responsibility in different and variable ways. So, schizophrenia will, perhaps sometimes, incapacitate one to a degree sufficient to undermine responsibility. Even when delusional, however, the schizophrenic is not necessarily immune to the possibility of recognizing and responding to

moral considerations. This is not to deny that in plenty of cases, the symptoms of mental illnesses of various sorts will make sufficient havoc of the machinery of responsible agency. But the when and the how of the impacts of mental illness are not uniform across cases or across circumstances of action.

The second take home lesson about the context specificity of deficiencies in responsibility is that there remain deep and difficult questions surrounding the extent to which we can shape our environments to facilitate the building of better beings. One idea that is at least as old as Aristotle is the idea of a degree of path-dependence in moral cognition. That is, one acquires the ability to recognize and apply particular moral concepts only if one had particular experiences. There is a growing body of research in social psychology that suggests something like this for a variety of concepts, including honor and respect. If some moral and quasi-moral notions work this way, then we should wonder what sorts of environments foster sensitivity to what moral considerations, and whether there is anything we can do to shape the paths upon which our moral notions depend. However, even if there is no interesting story to be told about the path-dependence of moral cognition, there are numerous, well documented ways that situational forces enhance the frequency with which we engage in helping behavior, suppress our prejudices, and come to successfully act on the reasons we consciously recognize. But this also means that there are situational forces of which we are usually unaware that degrade our agency.

Navigating these waters is the next step, and deciding whether and how to shape our social world will be difficult. But self-creation has never been an easy task.¹⁸

University of San Francisco

Notes

¹ Helpful discussions of the early roots of this chain of reasoning can be found in Susanne Bobzien, “The Inadvertent Conception and Late Birth of the Free-Will Problem,” *Phronesis* 43 (1988): 132-75; Richard Sorabji, “The Concept of the Will From Plato to Maximus the Confessor,” in *The Will*, ed. Thomas Pink, and Martin Stone (London: Routledge, 2003).

² John A. Bargh, “Free Will is Un-Natural,” in *Are We Free? Psychology and Free Will*, ed. John Baer et al. (New York: Oxford University Press, 2008); P. Read Montague, “Free Will,” *Current Biology* 18, no. 14 (2008): R584-R585; Susan Pockett, “The Concept of Free Will: Philosophy Neuroscience, and the Law,” *Behavioral Sciences and the Law* 25 (2007): 281-93; Daniel M. Wegner, *The Illusion of Conscious Will* (Cambridge, MA: MIT Press, 2002).

³ Derk Pereboom, *Living Without Free Will* (Cambridge: Cambridge, 2001).

⁴ I also suspect there is a kind of mind attracted to hard, outlier conclusions. The urge to stand apart (to be rebellious, to be intellectually “tough”) is no less a posture in the professor than in the teenager.

⁵ In fact, I reject almost every aspect of the chain of argument, as it is presented. Here, though, my focus is restricted to blocking its implications for moral responsibility.

⁶ See P. F. Strawson, “Freedom and Resentment,” *Proceedings of the British Academy* XLVIII (1962): 1-25.

⁷ A further difficulty: the conceptual role itself could be ill-advised or incoherent. Something like this seems to have often been the case with the history of scientific concepts. A category arises to explain a variety of different phenomena, and it turns out that this explanatory role for which the concept was generated does not exist in that fashion.

⁸ Let us assume we have an account of right and wrong action. For the consequentialists of the time, this was frequently understood to mean action that increased or decreased utility.

⁹ Kuru was a disease afflicting the brain that came to international attention in the 1950s because of an outbreak in Papua New Guinea; one acquired it by eating dead humans, especially their brain and nervous system. It became extinct within a generation of Australia’s 1957 ban of cannibalism.

¹⁰ One could have a view on which, given the existence of an afterlife, it is a somewhat common and ordinary thing to influence the dead with, say, prayers of petition. But I put this possibility to the side precisely because it does not seem plausible to think that the ordinary case of making a responsibility judgment about the dead *requires* such a possibility. Even less interestingly, one might think that one can influence the dead by changing relational properties that include the dead as one of the relata. My grandparents can be made more genetically successful by my procreative endeavors, and in this way, I might be said to exert some influence on them, even in death. But this sort of influence does not seem relevant or even required in the case of ordinary judgments of praise and blame.

¹¹ The point here is that we can have a set of conceptual or connotative content that is distinct from the oftentimes trickier matter of identifying the property or constituent properties referred to in uses of the concept.

¹² There are a number of philosophers who have done important work regarding what reasons responsiveness comes to. In what follows, I will not attempt to note each of my intellectual debts and the various points on which my account departs from those who have influenced me. Instead, I will simply note that in thinking about reasons responsiveness, I take myself to have learned a good deal from the work of John Martin Fischer, R. Jay Wallace, and Nomy Arpaly, among others.

¹³ Here I'm cheating a bit. There is a lively philosophical disputation concerning how we should understand capacity talk, and in particular, the capacities of agents in the context of praise and blame. In the text, I am helping myself to the idea that there is a notion of capacity in ordinary language that entails that even if our capacities are stable in lots of ordinary contexts, that same capacity changes under specific conditions. Unsurprisingly, I think that there is a philosophically respectable way of cashing out this basic idea, although in the text above I take myself to just be relying on the intuitiveness of natural language use. The substantive philosophical account I prefer depends on the idea that the capacities we are properly interested in when thinking about responsibility will be somewhat general or coarse-grained, not appealing to the precise circumstances of actual action given the actual past and actual laws of nature. Since this species of capacity talk is, on my view, dependent on our practical interests, it permits this phenomenon (above) where the general powers we focus on do not neatly track our motivational variations. One could think, instead, that our capacities are best understood in a fine-grained way, tracking actual motivational variation (so, for example, we just have "capacity

to resist temptation when my brother is in the room” and “capacity to resist temptation when my brother is not in the room”). I think there are reasons to disfavor this fine-grained approach to capacity talk, which hinge on the complexity of making the presumably vast number of micro-capacities salient in ordinary deliberation and collective social organization of practices. However, characterizing the generality I think is appropriate, motivating its viability as an account of capacity, and explaining how it maps on to ordinary practices of deliberation about responsibility are some the challenges that arise for the more coarse-grained account I favor. See my “Situationism and Moral Responsibility: Will in Fragments” (forthcoming).

¹⁴ What other things? Well, whether the agent did something good or bad. And, for example, what the norms of blameworthiness say we ought to do in light of the agent having done that good or bad thing in that context. More on this latter idea in a moment.

¹⁵ I don’t pretend to have here offered an account of what the relevant capacities come to. Call a particularly demanding notion of capacity a “Garden of Forking Paths” picture of capacity—one where we hold fixed facts about the actual past and actual laws of nature, and ask what is possible holding fixed those starting conditions. I do not think this is the sense of capacity required for moral responsibility. Instead, I think much looser conditions will hold, akin to those that hold when we truly say that someone has the capacity to speak Spanish (even though he or she may be speaking in English at the moment). On this picture, our capacities are fixed partly by our practical purposes in ascribing those capacities. So, on the present approach, the relevant notion of capacity will be partly given by the notion of capacity that would be required to effectively cultivate moral reasons responsiveness. For more on this matter, see note 13.

¹⁶ This is not to suggest that I take my account to be immune to empirical disconfirmation. On the contrary, I think that (for example) were someone to show that even the best system of praising and blaming over time corroded moral reasons responsive agency or gradually constricted the range of contexts in which we had such agency, then this would be grounds for rejecting the theory. Thanks to Gordon Barnes for raising this issue.

¹⁷ There is some evidence that in very rare cases people with antisocial personality disorder (a category that imperfectly overlaps with the psychopathy diagnosis) can “snap out of it” or begin to live more conventional lives, usually in mid-life. Also, we might wonder whether and how differences in social context and social networks can affect the ability of at least some psychopaths

to compensate for the characteristic emotional impairments. Finally, there is the matter of whether some of the cognitive strategies adopted by autistics for navigating moral norms might be transferred to the psychopathic case. My own sense, though, is that we are unlikely to find that there are many—if any—cases where psychopaths are non-accidentally sensitive to moral considerations.

¹⁸ Thanks to the material support and hospitable environs of the Radcliffe Institute for Advanced Study and the Stanford Center for Ethics in Society, where I worked on this paper. Thanks too, to Kristin Drake, Heather Fox, and audiences at the Radcliffe Institute for Advance Study, the College at Brockport, and Sacramento State University, for feedback on earlier incarnations of this paper.

Bibliography

Bargh, John A. "Free Will is Un-Natural." In *Are We Free? Psychology and Free Will*, edited by John Baer, James C. Kaufman, and Roy F. Baumeister, 128-54. New York: Oxford University Press, 2008.

Bobzien, Susanne. "The Inadvertent Conception and Late Birth of the Free-Will Problem." *Phronesis* 43 (1988): 132-75.

Montague, P. Read. "Free Will." *Current Biology* 18 (14), no. 14 (2008): R584-R585.

Pereboom, Derk. *Living Without Free Will*. Cambridge: Cambridge, 2001.

Pockett, Susan. "The Concept of Free Will: Philosophy Neuroscience, and the Law." *Behavioral Sciences and the Law* 25 (2007): 281-93.

Sorabji, Richard. "The Concept of the Will From Plato to Maximus the Confessor." In *The Will*, edited by Thomas Pink, and Martin Stone, 6-28. London: Routledge, 2003.

Strawson, P. F. "Freedom and Resentment." *Proceedings of the British Academy* XLVIII (1962): 1-25.

"Situationism and Moral Responsibility: Free Will in Fragments." in *Decomposing the Will*, edited by Till Vierkant, Julian Kiverstein, and Andy Clark, New York: Oxford University Press, forthcoming.

Wegner, Daniel M. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press, 2002.