Constitutive Instrumentalism and the Fragility of Responsibility

Manuel Vargas, UC San Diego

Forthcoming in *The Monist* vol. 4.4 (2021)

ABSTRACT: Constitutive instrumentalism is the view that responsibility practices arise from and are justified by our being pro-social creatures who need responsibility practices to secure specific kinds of social goods. In particular, responsibility practices shape agency in ways that disposes adherence to norms that enable goods of shared cooperative life. The mechanics of everyday responsibility practices operate, in part, via costly signaling about the suitability of agents for coordination and cooperation under conditions of shared cooperative life. So, there are a range of identifiable conditions where the ordinary operation of responsibility practices—and thus, the usual normative force of the practices—is disrupted. Even so, these conditions are not so widespread as to favor a more thoroughgoing abandonment of responsibility practices.

Constitutive Instrumentalism and the Fragility of Responsibility

Manuel Vargas, UC San Diego

There is a family of views about moral responsibility that is sometimes characterized as instrumentalist. What unifies these accounts is that they hold that some or another central aspect of moral responsibility is rightly understood in terms of its producing particular effects. The instrumentalist element varies by account. On some, the instrumentalism is in judgments of responsibility, where these are construed as seeking to alter behavior in pro-social ways. On other accounts, the instrumentalism is in the way we hold others responsible, in expressing blame where it contributes to some good. On still others, the instrumentalism is systematic, located in the justification of responsibility practices as something that enables a further good or set of goods. This article offers a theory that is in the spirit of existing instrumentalist accounts. It focuses on moral responsibility as a distinctive kind of normative practice. The novelty of the present account is the idea that responsibility practices are a solution to a particular pair of problems that arise from the nature and conditions of our sociality.

The exposition proceeds in three main pieces: first, a sketch of *constitutive instrumentalism*, the aforementioned novel form of instrumentalism about moral responsibility; second, a notable departure from a recently proposed account of blame in terms of costly signaling, with a brief aside on some of the mechanics of that account; and third, a discussion of conditions that undermine the proper efficacy and normative authority of moral responsibility practices, given the foregoing. Crudely, what follows are answers to the following questions: Why have responsibility practices in the first place? How do those practices tend to function, and why do they function in that way? What kinds of things disrupt the functioning of those practices? The overarching account is one according to which our responsibility practices are both a product of social natures and an effort at regimenting those natures.

Instrumental thoughts

Let's begin with methodological matters and their motivation.

One venerable approach to constructing a theory of moral responsibility begins with our ideas about accountability.[1] On this approach, we start with our concepts of RESPONSIBILITY, FREEDOM, and so on, or else, from their apparent meanings. From there, we isolate and extract the commitments required for a philosophically adequate theory of moral responsibility. Metaphysical analysis—which may include reflective equilibrium, inference to the best explanation, linguistic intuitions, and thought experiments—delivers us the materials required for a theory of moral responsibility. This analysis, then, is what settles the key questions, for example, about whether responsibility is compatible with determinism. With a theory of the concept in hand, we can ask whether or not anything in the world corresponds to or realizes the requirements on a theory of responsibility. Call this methodology *conceptualist*.

A different place to start is not with our ideas of moral responsibility, considered in the abstract, but instead in our existing social practices of holding one another responsible. On this approach, our theories are beholden to the concrete phenomena of our social practices and their attendant attitudes (Strawson 1962). Capturing and

[1] Accountability responsibility takes as its target a kind of moralized blame according to which the agent is not merely defective or broadly normatively undesirable, but in some or another way culpable or at fault in a way that tends to license some negative response.

explaining those things (things a social scientist might point to as distinctive of our responsibility-characteristic attitudes and practices) is the primary aim for a theory of responsibility. Call this methodology *phenomenalist* or *practice-first*. It is phenomenalist in that its point of departure is not our ideas about responsibility, but concrete phenomena in the world—praise, blame, and the way attendant practices of these things function. It is practice-first, in that it is the practice of holding one another responsible rather than our ideas about agency or freedom or responsibility, as such, that provides the most important information for our theorizing.

These are methodological idealizations. Particular theories often employ elements of both. Methodology matters, though. Conceptualism requires a contested view about meaning, reference, and topic continuity: it relies on the idea that our present convictions about a concept settle questions about what the property or thing comes to (Hurley 2000; Vargas 2004; Deery forthcoming). In contrast, practice-first or phenomenalist methodologies can comparatively readily allow that the world or our interests can come apart from the things we think about that thing (Vargas 2004; forthcoming). By anchoring our theorizing in concrete phenomena—e.g., specific practices and attitudes—phenomenalist approaches reduce the risk that we inadvertently find ourselves developing an account of something like what Dennett (2006) called the "truths of chmess," which is to say truths about a thing akin to chess, but a version of the game that is not at stake in actual practice.

Challenges lurk for the phenomenalist, too. Her proto-anthropology of our social practices will be unsatisfying if it cannot explain how responsibility attributions can be in error, or why we typically have reason to care about or engage in such practices. Phenomenalist and conceptualists alike must say something about these phenomena. Ideally, their accounts would give us an explanation of why we should not forfeit the practice altogether, as responsibility denialists sometimes urge (Pereboom, this volume).

Recently resurgent instrumentalist theories are particularly well-positioned to perspicuously address puzzles about responsibility's normativity in a broadly phenomenalist fashion. Instrumentalist approaches to responsibility hold that the normative force of the practice of responsibility is located in the effects of the practice, for example, in the way the practice develops our moral or rational capacities (Arneson 2003; Vargas 2008; McGeer 2012; Fricker 2016; Jefferson 2019). Details differ.

On one version (e.g., McGeer 2015), instrumentalism is located in the justification for token instances of blame. An instance of blaming is appropriate when that blaming can expand the capacity of the blamed to recognize and respond to moral considerations. On a different version (e.g., Vargas 2013), the instrumentalism applies at the level of the practice as a whole, but not at the level of particular norms of responsibility or in particular judgments of responsibility. The practice qua practice needs to have the right effects, but particular first-order judgments and norms may be most effective if they are not themselves instrumental in character. This two-tiered structure mimics other social practices, such as the law and sports, where forward-looking justifications can license entirely backward-looking, desert-entailing, first-order judgments. For example, whether something is a foul in a sport is a matter of whether the rules were violated. The justification of those rules may be entirely instrumental, concerned with securing the safety of the players and the orderly progression of play.

In what follows, I take the viability of instrumentalism about responsibility for granted. Elsewhere, I've attempted to address standard concerns about instrumentalism, including worries about scapegoating, the wrong kind of reasons objection, the role of desert, objections about self-effacement, and the in-principle

independence of such accounts from consequentialist normative ethics (Vargas 2008; 2013; forthcoming). Here, my ambition is to offer an account of constitutive instrumentalism, where this is understood as a theory of specifically *moral* responsibility. Whether the account generalizes to other forms of responsibility is an open question.

Constitutive instrumentalism

We are norm-sensitive creatures. We attend to norms. Sometimes, we internalize and enforce them. Which norms we detect, acquire, retain, and enforce varies. Norms that are simple to learn and internalize, or that are readily backed by affect, tend to be easier to acquire (Nichols 2015). Norms endorsed by people whose respect we care about, or even simply norms valorized by my community, will come to shape my deliberative horizon more readily than norms that lack these features (Henrich 2016). In principle, any of these norms might be subject to reevaluation. In practice, some are more and less readily re-evaluable.

Moral norms present themselves as concerned with what we have special or most reason to do. We therefore have a corresponding interest in our competence at moral norms, or more fundamentally, in recognizing and responding to the moral reasons. This concern cuts in two directions. First, we typically want others to be competent at navigating social space in light of moral norms and considerations. People incompetent at recognizing and suitably responding to moral considerations are a cause for concern. They are unpredictable. We can't count on them to respect our interests. They are dubious partners for social cooperation, for coordination, and for peaceable collective life. Second, we tend to want moral competence for ourselves. Being perceived as competent at navigating moral considerations is a prerequisite for most ordinary interactions among mature adults.[2] And, ordinarily, the most reliable way to be perceived as competent across a wide range of social contexts is to indeed have such competence.

Our interest in being viewed as morally competent explains the otherwise curious fact of our real (if often reluctant) willingness to be held responsible for at least some of our wrongdoing. In taking responsibility, we are insisting that the episode in which we failed to respond appropriately was a local error, and not evidence of a systematic defect (Raz 2011; Vargas 2014). That putative competence is made plausible by our recognizing the propriety of blaming us for our wrongdoing in that case. It may also be true that in so accepting the blame we thereby enhance our capacity to rightly recognize and respond to the relevant reasons going forward (McGeer 2015). The value we place on being morally competent, and being recognized for it, is not one we readily forfeit for other goods.

So, we value our being morally competent, and other people value it in us. But how do we *achieve* the relevant competence? We aren't just born with it. Crucially, the problem isn't simply an individual one, to be solved on an individual and ad hoc basis. From the standpoint of human communities, the pressures are collective, involving temporally extended communities of agents that have to solve problems that affect both individual and collective access to goods of living in communities.

[2] Depending on the particular account, the stakes may be *actual* moral reasons, or the things we *regard* as moral reasons. There is a prudential interest in tracking what things are regarded as moral reasons around here. That's compatible with a concern for what reasons there are. In what follows, I will assume the stakes are actual moral reasons, although fallibilism about local convictions is in order, regardless.

It is a familiar fact that groups of individuals gain benefits from being able to cooperate and coordinate. Collectives can do more than individuals, and the standard view in the social sciences is that social organization, which is to say, groups of agents organized by conventions and norms, is a crucial element of the human toolkit (Boehm 2012; Henrich 2016; Bicchieri 2017; Bratman forthcoming). It is relatively easy to see why: cooperation and coordination that is on the fly or otherwise purely one-off cannot leverage the benefits of longer-term planning and coordination. Constant renegotiation is costly, and there are clear advantages to default solutions—conventions—that reduce or eliminate those costs. However, conventions are not always enough. Some cooperative goods can only be secured if group members adhere to the cooperative scheme even when it is costly to do so. The solution to the problem comes in two parts: individual emotions that function as *commitment devices* securing dispositions to act even when it is costly to do so, and norms that exploit these attitudes to enjoin the propriety of enforcement (Frank 1988, 4-7; Cushman 2015).

Responsibility practices exploit both of these features. Some emotions—e.g., the Strawsonian reactive attitudes—function as commitment devices, and our norms of holding people responsible build on and exploit those attitudes and correlative interests in enforcing social norms that foster stable, predictable social environments conducive to iterated cooperation and coordination. Perhaps it is possible to secure the goods of cooperation and coordination without responsibility norms. Nevertheless, responsibility practices are a widely distributed and common strategy for enabling cooperation and coordination within human societies, and it is unclear whether strategies that do without its main elements (blame, fault, and perhaps desert) are as stable, or whether they can readily scale up as responsibility practices do (Nichols 2015, 119-140). Responsibility practices are a reliable solution to the diachronic problem of how to get creatures with our particular affective and cognitive architecture to be reliable social cooperators. Agents acculturated by a responsibility practice will tend to weigh practical options in a way prospectively colored by estimations of praiseworthiness and blameworthiness, and with awareness that others in that community expect competence and some threshold of concern for cooperative norms. Internalizing those assessments and the related norms valued by that community disposes us in ways that facilitate cooperation and coordination (Henrich 2016; Shoemaker and Vargas, forthcoming).

Moral responsibility practices—those focused on culpable blameworthiness—are particularly important in the practical life of agents and collectives, but perhaps especially demanding to acquire. One can't readily see the world like a morally competent 11th century Buddhist, a 15th century Mexica, or a 21th century Malawian without extensive feedback from people who are already proficient within the local moral culture. Moral development is a communal practice. The attunement of the affects, and the proper dispositioning of our wills, requires sustained and nuanced feedback. In the ordinary case, the involvement of other people is central to the training up and calibrating of our relationship to a moral world.

The manner of attunement varies by our perception of the agent's development. If a normatively incompetent agent is young, vulnerable, or still maturing, then we more readily tend to think of our engagements as pedagogical. Sometimes this means feigning indignation or outrage. Other times it means more attention-directing—inviting the agent to reflect on how they would feel were it done to them, or asking them to reflect on what would happen if everyone was indifferent to those considerations. The patchiness of our rational capacities, and the stuttering development of them, makes the attunement process difficult. A relatively young child might do a good job of rightly recognizing and responding to considerations of physical harm. That competence need not carry over to her competence at recognizing that someone's feelings might be hurt by

recoiling at their less-than-pleasing visage, or in the child's recognizing that it is, all things considered, better to share one's toy with the new kid. That one can be competent with all the relevant moral considerations in one context and few if any in others, and the fact of our sometimes-fluid sense of which considerations matter, can make it difficult to discern whether atypical and developing normative agents are competent enough at moral reasons to be held responsible.

None of this entails that there is no difference between moral education and earnest blame. Earnest blame requires that the blamer thinks the blamed crosses some threshold of normative competence to count as a proper participant in blame practices. It is the fact of our having an arguably independent conception of *responsible agency*—apt participanthood—that animates efforts at moral education (although see McGeer 2015 for a rejection of the independence idea). In light of that conception, we employ a range of techniques that, with success, will make it the case that non-responsible agents eventually come to be responsible agents. However, it is precisely that—the absence of a conviction that the target is the right kind of agent—that distinguishes moral education from blaming.

Reflecting on the diachronic pressure for collectives to manage the conditions of sociality, and considering how a particular target conception of agency shapes our practices, these two thoughts suggest a distinctive picture of the normative foundations of responsibility. Rather than supposing the instrumental good of responsibility practices is located in some prior or antecedent notion (of desert, rational capaciousness, or welfare), we might look to the practices themselves.

Here's an initial conjecture: our having practices of holding and being held responsible, whatever one's local version of it comes to, is ordinarily how we become the kinds of agents we need to be, given that we live in communities that have their present shape and demands. Perhaps there are alternative methods of moral formation available to us. In the ordinary course of things, at least for some not trivial sense of "us," it is through responsibility practices that we shape our moralized ways of seeing—our sense of the practical value of different choices—in ways responsive to diachronic pressures for cooperation and coordination that beset people living in communities that over time have a collective interest in reliable and flexible solutions to the problems of this kind of communal life (see Bratman (2010, forthcoming) on diachronic rational pressures for individuals and collectives).

We can refine this basic idea by introducing a notion of mutual or reciprocal constitution. Responsibility practices—including a suite of affective dispositions, conceptual entailments about blame and wrongdoing, and normative commitments—shape agency. However, such practices also depend on the shaped agency to enable cooperation and coordination that enable that agency to achieve further goods. The mutuality arises because responsibility practices are a way of simultaneously solving a diachronic problem for collectives while at the same time serving as a vehicle for the solution of several more proximal challenges for the individual agent, including the formation of a moral sensibility and the securing of competences crucial for reliably accessing social goods. The relevant rational pressures for individuals and human collectives potentially arise somewhat independently, but they find a mutually supporting solution in responsibility practices.

On this picture, then, the normative authority of responsibility practices is grounded, at least in part, in the efficacy of such practices at provide an interlocking set of solutions to a web of fundamental social problem: (1) they secure the conditions for stable cooperation and coordination in collectives, (2) by shaping our moral

sensibilities in pro-social ways deeply anchored in our psychologies, (3) by enabling and securing our concerns for the social recognition of our competence at navigating moral demands and (4) by providing fast, flexible, and effective responses to local problems arising among all of the foregoing.

The normative authority of responsibility practices isn't just some good external to our sociality and social practices. Rather, the normative authority is at least partly grounded in the mutually constituting effects of our socially dependent agency on norm-using collectives. This is one way to explain a thought expressed by McKenna (2012, 53) that there is no metaphysical priority between being and holding responsible. The primary normative and explanatory function of responsibility practices is bound to the mutual entanglement of our practices and our nature.[3]

What makes this picture *constitutivist* is that responsibility is a normative practice arising from and justified by our being pro-social creatures who need certain kinds of practices to secure specific kinds of social goods. Responsibility practices that aim at cultivating our moral considerations-sensitive agency via standards of blameworthiness just are the way we (pro-social creatures with a particular range of psychologies) become morally competent agents. Absent such practices, it is unclear that we can become the kinds of agents we have reason to be, and it is unclear that we can reliably form social communities capable of reliably securing the distinctive goods of shared cooperative life. Responsibility practices constitute our collective solution to having agents capable of sustained cooperation and coordination. Participation in such practices, and especially, being genuinely competent at those norms ordinarily constitutes the individual's best solution to the problem of convincingly demonstrating to others one's suitability to shared cooperative life.

That a practice has a constitutivist structure does not entail that the particular contents of the practice are always justified. Judgments that are correct within the framework of a given responsibility practice can still be more and less justifiable relative to some external standard. This is particularly plausible for any practice concerned with morality, and it is an important constraint on the extent to which this picture entails relativism about responsibility. We can see this by considering a society that settled on responsibility practices at some remove from moral reasons. Suppose that society's members typically blamed agents (in fact) acting permissibly, and that its members failed to blame (in fact) culpable actions. That society would have a responsibility practice, and it might secure some of the goods that explain and justify having such a practice. That society would nevertheless have a practice less justified than one where the agency-shaping features better enabled agents to track the (in fact) moral considerations.

Practices can solve practical and normative problems in better and worse ways. Part of the normative authority of a practice derives from its efficacy as a kind of solution to goods that matter. Even in the limit case of a society only very loosely tracking moral considerations, there is presumably some good to be secured from enabling individuals to live cooperatively and to that society having stable norms that enable such agency. Still, if the rational sensitivities cultivated by responsibility practices operates only over reasons without moral authority, then the normative authority of the practice will be correspondingly limited. Where there is greater convergence between moral reasons and the rational capacities fostered by a given responsibility system (i.e., the responsibility practices, norms, attitudes, and judgments), there will be correspondingly greater normative authority in the practice. It is an interesting and further question—one I will not attempt to pursue here—

[3] The present proposal is silent on the question of whether there might be other, potentially cross-cutting, normative and explanatory functions to responsibility practices. For pluralism about functions of responsibility, see Wang (in preparation).

whether the bare fact of possessing responsibility practices is itself sufficient to provide an Archimedean point for individuals or societies to gradually identify more moral reasons. For all that has been said, it may be that communities can find themselves on epistemic islands unable to connect to either what moral reasons there are or even to the reasons of others, imperfect and variably connected to morality (whatever that comes to) in their own local ways (Taylor 2017).

Constitutive instrumentalism is compatible with many forms of instrumentalism, but a theory may be instrumentalist without a commitment to specifically constitutive instrumentalism. Holding that responsibility's normative foundations are tied to both our nature as social individuals and the suitability of particular practices for solving diachronic pressures that arise from collective life can buttress the explanatory and normative power of instrumentalist accounts.[4]

With a provisional sketch of how social pressures and the nature of our agency jointly produce the practical and normative pressures that favor responsibility practices (the answer to the question of *why* we have responsibility practices), we can turn to the *how* question, that is, how the aforementioned interest in indicating normative competence shapes everyday blaming practices.

Costly signaling

The way we blame reveals some important features and limits to a responsibility system. In prior work David Shoemaker and I argued that the difficulty philosophers have had in accounting for blame is a function of the complexity we find in things that seem to count as blame (Shoemaker and Vargas, forthcoming). In some contexts, blame is paradigmatically affective, as when one experiences hot indignation at having been slighted. In other contexts, though, that affect is largely or completely absent, or its precise affective nature can shift from hot to cold, from strong to absent.

One way a theory of blame might account for this complexity is to maintain that blame has multiple faces: (a) judgments of blameworthiness, and (b) blaming reactions (Vargas 2013, 117-8). One can dispassionately blame a philanthropist for wasting her millions on another educational program that benefits ten people who already enjoy extraordinary educational opportunities (Shoemaker and Vargas, forthcoming). In doing so the blamer may judge that the blamed is blameworthy, without feeling the flush of affect that is commonly characteristic of blame. But suppose I discover that Annie's proposal for studying the life cycle of bed bugs got funded, but not my project on Mexican existentialism. I might acknowledge that her project is indeed more meritorious, and that the matter was reasonably decided, while simultaneously being bugged that she won the funding from a committee chaired by a friend of hers. If so, I might be resentful of her receiving the award, even while grudgingly admitting that she deserved it. Blaming reactions can come apart from judgments of blameworthiness.

[4] The present proposal is both continuous and discontinuous with aspects of my prior defenses of instrumentalism. Previously (e.g., Vargas 2008), I have tended to suggest an "end state" instrumentalism where the justification of responsibility practices was in their contribution to attaining a form of agency that appropriately responded to moral considerations. At the same time, the idea of constitutive instrumentalism is very much in the spirit of the agency cultivation model's emphasis on practices that shape moral sensibilities in ways tied to pro-sociality (Vargas 2018). What is new here, beyond the making explicit an idea latent in earlier work, is the way these processes are a particular response to a general problem about how groups can shape individuals in ways that fit individuals to collectives, but that in doing so, typically benefit the individual as well.

The appeal of a disjunctive approach, where blame has distinctive cognitive and affective forms that can occur independently, is that it captures some of the complexity we find in real world blame. The problem, though, is that disjunctivism doesn't, by itself, explain the relationship between the judgment and which blaming reaction is taken up. Nor does it tell us why these reactions and attitudes count as blame.

Shoemaker and I have argued that a more promising account explains the diversity of blaming practices not in terms of the content of blame, but in terms of its signaling function. A tacit commitment of that view, shared by the view proposed here, is type-level functionalism. Individual instances of blame might be degenerate cases lacking the signaling function while still being produced in the characteristic way, but blame—as a phenomenon or type—should be understood as a costly signal.

Here, I recast the costly signaling story as a story about the nature and role that *expressing* blame attitudes and judgments perform in our social economy. In contrast to Shoemaker and Vargas (forthcoming), the present proposal grants a content-based story about blame but deploys the idea of signaling to offer a more complex story of blam*ing*, or the roles or aims in the service of which blame (whatever that comes to) is deployed. With that more complex story in hand, we can then see why both judgment-like and affective reactions are candidates for blame: both serve as anchors for a signaling practice that is particularly responsive to the social pressures that arise in collective moral life.

First, a posit: for someone to be blameworthy, that person must be (i) a responsible agent, i.e., one capable of recognizing and responding to moral considerations, and (ii) she must have culpably performed some wrong (Brink 2012; Brink and Nelkin 2013; Vargas 2013).[5] We can and do make mistakes about both elements. However, blame only rightly applies to certain kinds of agents—players, as it were, in the responsibility practice. Those agents only deserve blame when their behavior is both wrongful, and in a culpability-generating relationship with that behavior (i.e., they lack an excuse).

That blame *worthiness* tracks culpable wrongdoing by responsible agents is an important norm for blaming, but it doesn't yet explain the diversity we see in forms of blaming. Explaining this requires an account of the social economy of blame, that is, its significance in our social practices. What follows is an adaptation of Shoemaker and Vargas (forthcoming), recast as a theory of blam*ing* and not blame.

An initial observation sets the stage. Blaming imposes substantial costs on blamers, in both social investment and ongoing emotional and social burdens. First, as noted above, acquiring normative competence isn't automatic. Instead, it's an emotionally, attentionally, and time-intensive process of acculturation and socially mediated attunement of one's moral sensibility. Second, even with that competence, instances of blaming often require further costs. Some costs are psychological: the anger, the resentment, and the generally negative appraisals one directs at the blamed are unpleasant for the blamer. There are various substantive costs in terms of time, friendships, and the risk of retaliation.

So why do we do it? We do it because blaming is a costly signal. The core idea of costly signaling is of a hard-to-fake signal that tells us something about individuals that can be both valuable to the individual and to observers, but where the signaler might gain advantages from merely mimicking what is signaled and where the

[5] Earlier versions of this distinction can be found in Strawson (1962), Wallace (1994), and Fischer and Ravizza (1996).

observer's interest is in the truth or reliability of what is signaled. As we've seen, the possibility of being a target of blame alters one's deliberative landscape. This typically requires that one (a) cares about expressions of blame, and (b) thinks the blaming is genuine; it can also require that one (c) recognizes that others may engage in costly enforcement of that blame.

The reactive attitudes play an important role. Recall the idea of emotions as commitment devices, binding us to courses of action, even when the short-term benefits of doing so are unclear. Moral anger and the other blame-related emotions motivate us to stand up for our interests, and the interests of those we care about, even at great cost to ourselves. This costly self-binding has a benefit. When other agents recognize that we are so committed—that we are prepared to over-invest in the enforcement of these interests (that is, to expend disproportionate amounts of time and energy relative to the value of the good that has been lost at that time) we alter the context that shapes the deliberative possibilities of others. Our backward-looking reactive attitudes earn their keep in forward-looking ways (Frank 1988; McGeer 2013; Vargas 2013; Cushman 2015).

Here, then, is the social heart of blaming. In blaming, I (reliably, and typically inadvertently) signal an otherwise opaque trinity of information about myself: my competence at moral norms, that these norms matter to me, and that I support their enforcement on suitable targets. My emotional expressions are a sometimes involuntary signal to others, but they are not the only way to signal information about myself. The truth or reliability of the signal can be underwritten both by attitudes often difficult to convincingly feign (e.g., indignation and resentment) and by my willingness to enforce (when I blame another) or accept norms of enforcement (when I accept blame or accept the propriety of blaming a friend) even when it comes at a cost to my interests. What unifies the variegated things we recognize as blaming behavior is not some specific attitude or effort in response to a norm violation—e.g., the hot expression of a reactive attitude, a protestive reaction, the alteration of one's relationship, or even some effort at communication. Instead, blaming is a costly (frequently inadvertent) signal about oneself and one's moral understanding and commitments. Sometimes the audience for that signal is others. Sometimes it is oneself, as in cases of private blame. The gradual internalization of blaming norms, and the concomitant reflexiveness and spontaneity of the impulse to blame predictably produces cases where there is no audience at all. Such cases are a habituated tokening of a type whose function is costly signaling.

Reflecting on childhood can make clear how blaming is bound up in one's suitability for cooperative life. When children say that they want to be grown-ups, it isn't the age they want. They want the authority to make independent decisions that cannot be readily contravened by others. That authority rests, at least in part, on competence with moral norms. Being able to recognize and respond to those norms, and being prepared to enforce them, is a hard-to-fake signal about one's moral competence, and thus, one's standing in the moral community as an equal. Accepting blame is a way of signaling our competence at normative demands, and our commitment to them. Being able to both blame and accept blame is a signal that one is the kind of agent our social world relies upon for cooperation, coordination, and the ongoing transmission of the sensibilities required to sustain the goods of sociality. So, signaling expresses where one stands, but it also conveys the kind of agent one is—that is, one suitable for shared cooperative life.

If this account is correct, moral blame may be of a piece with blame more generally. That is, the psychological machinery of moral blame is the same psychological machinery involved in various non-moral cases of blame. When you condemn me for violating the norms of a good philosophy talk, you are signaling your

commitment to particular norms, and enjoining me to signal my commitment, or else, my incompetence at those norms. Whether the domain is art, sport, or any other norm-structured practice, the mechanics seem mostly the same. The most salient difference seems to be that in the case of morality, unlike signals about one's nuanced taste in corridos or savvy gameplay at croquet, we don't take competence with (and enforcement of) moral norms to be a substantially elective matter.

An aside on faking and signal shaping

A natural set of questions to press about the foregoing account concerns efforts to intentionally manipulate the blaming signal. In what follows, I canvas a few details about the mechanics of blaming, before returning to the main thread about how our sociality shapes the nature of responsibility.

Paradigmatic blaming emotions can, of course, be faked. So too with remorse, and other emotions characteristic of taking responsibility. However, sustained, effective deception is difficult. Even professional actors can fail to convince, and this is partly because the importance of the signal and the conditions of social life tend to train up our alertness for strategic deception. It is also one reason why we invest so much time cultivating the dispositions of children: internalization and ongoing calibration of these norms is a far better bet for displaying normative competence than on-the-fly attempts to model local demands and to suitably configure one's disposition toward them. Without the underlying convictions, continuously enacting appropriate responses is also difficult. It is challenging to sustain feigned grudge-holding, retractions of interpersonal warmth, and emotion-saturated denunciations; it is likewise difficult to sustain a pretense that one acknowledges that one is deserving of these things, that one can and should do better, without the attendant belief and commitment.

Signals can be complex in several ways. Suppose that during a question period after an academic talk you just gave, I angrily condemn your persistent and ongoing complicity in the colonial, racialized, Eurocentric erasure of the thought of non-white peoples. This sends a signal to you about what I care about, and that I'm prepared to make a case despite some amount of eye-rolling and dismissive smirking. At the same time, my condemnation may signal solidarity to the questioner whose reasonable question invoking Fanon was politely ignored. It may also serve to let a silent grad student know that the nature of the discipline and its canon can be contested. Depending on what signals I intend to convey, and my estimation of their perceived import, I may try to shape the signal—making my remarks with a wry smile but a steely gaze, or instead, uttering them with the dispassion of an Aristotle scholar pronouncing on a new interpretation of *Metaphysics* Book Zeta. The reliability of the signal, though, is found in the fact that these signals involve difficult-to-manage combinations of affect and socially salient content.

That signal-shaping is difficult, perhaps especially when one is in the grips of a "hot" conviction, does not mean it is impossible. We can, for example, cynically engage in moral grandstanding, to signal our interest in being regarded as a member of this or that identity group, or to create the appearance of particular moral or political commitments (Tosi and Warmke 2020). We can also try to shape our signals by dog-whistling—that is, by expressing normative commitments in ways designed pass unnoticed by some, but to be clearly heard by others (Haney-López 2014).

Grandstanding and dog whistling are easier to do in environments where there is a relatively narrow band of information that can be signaled—e.g., in social media, where one's affect, comportment, and sustained demonstration of costly commitment to those norms is not ordinarily available. To the extent to which our

social world is populated by these opportunities to control the signal, we should expect more attempts to exploit this possibility for signal control. We should also expect some attendant growth of skepticism about those signals. In environments where this sort of signal control is pervasive, we have less reason for taking blame as the reliable interpersonal signal it otherwise is. Even so, staring at a screen is unlikely to entirely wipe away our hard-won moral psychology. Blame may still sting even when we know it does not signal all the usual things.

Responsibility in the wild

I've sketched a picture of constitutive instrumentalism as an account of the justification for having responsibility practices at all. Two ideas were especially important in that account: (1) agents come to be competent at navigating a social world structured by moral norms via our collectively having responsibility practices, and (2) for a society, its having members that are capable of recognizing and responding to moral considerations is an enabling condition on securing the goods of sophisticated forms of cooperation and coordination. I then argued that the *how* of everyday responsibility practices is explained, to a large degree, in terms of costly signaling. At the center of the costly signaling account is a signal about the same form of agency at stake in the constitutive instrumentalist story: the ability to recognize and suitably respond to moral considerations—and a concomitant dedication to enforce the moral norms. So, we have an elegant fit between the normative structure of constitutive instrumentalism and the thing being signaled in blaming.

Difficulties lurk. First, as noted above, the justification given by constitutive instrumentalism is only partial. A responsibility system is more and less justified to the extent to which its particulars reflect what moral reasons there are. Even if we assume there is some non-trivial convergence between what moral considerations there are and a given community's convictions about such matters, a second problem looms. The instrumental efficacy of a given responsibility system depends, in part, on the extent to which blaming-as-signaling practices function as *reliable* signals. The reliability of a signal depends on a variety of contextual features. Given that the functioning of responsibility practices—and the goods secured by such practices—depends on the reliability of the signal, some further remarks are in order.

There are at least three cases where we should expect that constitutive instrumentalist justifications are weakened: conditions under which there is low or no social trust; conditions under which there are competing packages of norms; and conditions under which there are transformations in how signals are sent and received.

The first class of cases arise when hostility or a lack of trust leads people to disregard the ordinary signals sent by blaming. Where moral convictions aren't shared, where one supposes the other parting isn't acting impartially or with some minimal good will, or where one has reason to think the other party is feigning blame, observers will tend to regard the signal as noise. Writ large, this suggests that it is important for social institutions that operate in blame-saturated contexts—the criminal law, for example—to be regarded as impartial and uncorrupted by bias or purely strategic deployment. In essentially adversarial contexts, blame's signal and its presumptive justification may wither without a ready presumption of impartiality and legitimacy.

A second class of cases where the efficacy—and thus the presumptive justification of blaming practices—is undermined concerns instances of deep and significant moral disagreement. In such cases, it is perhaps less likely that blaming practices will perform their ordinary signaling function, in part because moral perception and our sense of what matters morally is a product of socially-scaffolded attunement (Rudy, forthcoming). In

contexts of substantial disagreement, moral competence in one community's terms may look like incompetence by the lights of another community. If those agents operate without a sense of convergence about which norms are authoritative (or at least, which decision procedures are authoritative for resolving disagreements), the responsibility practices may be less effective at producing the goods that justify the practice.

Cases of trenchant disagreement might arise in the case of blame across communities with different moral standards. It might also arise in a case of bifurcation in the normative commitments within a single community, or alternately, it might occur where one is a member of two communities with conflicting normative commitments. One can imagine that an agent might have a high degree of trust in diverse normative packages that are ultimately at odds with one another. Such cases need not be instances of low social trust, although they can be. The point is that we should be cautious in assuming that an instrumentalist justification, especially one that operates via a costly signaling practice, unproblematically extends across distinct moral ecologies.

A third case where blame's typical justification-producing function can degrade arises when there is a significant transformation in the mechanisms or technology of social signaling. Consider the case of social media-based blaming, in which the numbers of blamers blaming an instance of wrongdoing is subject to rapid scaling up on social media platforms. It is not just a matter of the numbers and ease of blaming. Such technologies seem to have enabled a ready intensification of the blame itself. An offense that might have been met with a raised eyebrow or the brief reproach of a friend or two can now become the target of attention, invective, and the unrestrained hostility of millions. To be sure, this is an artifact of mass media in general, and contemporary social media is merely a democratization of the basic phenomenon of scaled interconnectedness. However, the scaling up of access to blame afforded by this technology operates without any of the feedback mechanisms typically in place in traditional instances of directed interpersonal blaming—e.g., seeing the effects of the blame on the blamed party, a person with whom one stands in some interpersonal relationship, and being subject to condemnation by trusted peers when one is excessive in one's blaming. To the extent to which our moral psychologies and inherited moral practices presume relatively local and small-scale communities of blamers, increased exposure to a wider community of potential blamers can wreak havoc on the economy of blame.

These three kinds of cases suggest that we must be cautious in thinking that the basic signaling and justificatory structure of blame readily extends to all social contexts. At the same time, there is no reason to think that these conditions of atypicality are so pervasive as to recommend the total abandonment of practices blaming and holding responsible. If the foregoing is right, though, identifying the real-world conditions under which the presumptions that make sense of and justify our having responsibility practices, may be a central task for theorists of moral responsibility.

Conclusion

Instrumentalists hold that we can account for the normative authority of our practices in terms of some good achieved by participation in those practices. Constitutive instrumentalism holds that responsibility practices are a special solution to a matrix of pressures arising from our sociality, including both the interests of individuals to be the kinds of agents that can enjoy the benefits of sociality, and the pressures for diachronic stability in norms of cooperation and coordination among groups of such agents. Blaming practices leverage judgment-like attitudes and affective reactions in a system of costly signaling that better enables responsibility practices to simultaneously shape agents in ways that serve the typical interests of agents while at the same time enabling

the conditions necessary for enjoying the fruits of shared cooperative life. That the system is relatively stable and effective at shaping our moral lives should not distract us from the relative fragility of the background conditions that make it presumptively justified.[6]

[6] This paper, and a distant predecessor, benefitted from the feedback from philosophers at the U.C. San Diego, U.C. Berkeley, the University of Vermont, Texas Christian University, the University of Zurich, and the Princeton Center for Human Values. Special thanks to Lucy Allais, David Brink, Anneli Jefferson, Miranda Fricker, Chris Kutz, Tori McGeer, Stephen Morse, Dana Nelkin, Philip Robichaud, Katrina Sifferd, Dan Speak, David Shoemaker, Shawn Wang, and Monique Wonderly for helpful thoughts. Thanks, too, to the critics of my earlier work with David Shoemaker on the signaling theory.

References

Arneson, Richard J. 2003. "The Smart Theory of Moral Responsibility and Desert." In *Desert and Justice*, edited by Serena Olsaretti, 233–58. Oxford: Oxford.

Bicchieri, Cristina. 2017. *Norms in the Wild*. New York: Oxford University Press.

Boehm, Christopher. 2012. *Moral Origins*. New York: Basic Books.

Bratman, Michael. 2010. "Agency, Time, and Sociality." *Proceedings and Addresses of the American Philosophical Association* 84 (2): 7–26.

Bratman, Michael. Forthcoming. "Shared Intention, Organized Institutions." *Oxford Studies in Agency and Responsibility* 7

Brink, David O., and Dana Nelkin. 2013. "Fairness and the Architecture of Responsibility." *Oxford Studies in Agency and Responsibility* 1: 284–314.

Cushman, Fiery. 2015. "Punishment in Humans: From Intuitions to Institutions." *Philosophy Compass* 10 (2): 117–33.

Deery, Oisín. forthcoming. *Naturally Free Action*. Oxford, U.K.: Oxford University Press.

Dennett, Daniel. 2006. "Higher-Order Truths About Chmess." *Topoi* 1-2: 39–41.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Frank, Robert H. 1988. *Passions Within Reason*. New York: Norton.

Fricker, Miranda. 2016. "What's the Point of Blame? A Paradigm Based Explanation." *Nous* 50 (1): 165–83.

Henrich, Joseph. 2016. *The Secret of Our Success*. Princeton, NJ: Princeton University Press.

Jefferson, Anneli. 2019. "Instrumentalism About Responsibility Revisited." *The Philosophical Quarterly* 69 (276): 555–73.

McGeer, Victoria. 2013. "Civilizing Blame." In *Blame: Its Nature and Norms*, edited by Justin D. Coates, and Neal A. Tognazzini, 162–88. Oxford: Oxford University Press.

——. 2015. "Building a Better Theory of Responsibility." *Philosophical Studies* 172 (10): 2635–49.

McKenna, Michael. 2012. *Conversation and Responsibility*. New York: Oxford University Press.

Nichols, Shaun. 2015. *Bound.* New York: Oxford University Press.

Raz, Joseph. 2011. *From Normativity to Responsibility*. Oxford: Oxford University Press.

Rudy-Hiller, Fernando. Forthcoming. "Moral Ignorance and the Social Nature of Responsible Agency." *Inquiry.*

Shoemaker, David, and Manuel Vargas. Forthcoming. "Moral Torch Fishing: A Signaling Theory of Blame." *Nous.*

Strawson, P. F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* XLVIII 1–25.

Taylor, Kenneth. 2017. "Charting the Landscape of Reason." *Proceedings and Addresses of the American Philosophical Association* 91 43–64.

Tosi, Justin, and Brandon Warmke. 2020. *Grandstanding*. New York: Oxford University Press.

Vargas, Manuel. 2004. "Responsibility and the Aims of Theory: Strawson and Revisionism." *Pacific Philosophical Quarterly* 85 (2): 218–41

——. 2008. "Moral Influence, Moral Responsibility." In *Essays on Free Will and Moral Responsibility*, edited by Nick Trakakis, and Daniel Cohen, 90–122. Newcastle, UK: Cambridge Scholars Press.

——. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford, U.K.: Oxford University Press.

——. 2014. "Razian Responsibility." *Jurisprudence* 5 (1): 161-72.

——. 2016. "Responsibility and the Limits of Conversation." *Criminal Law and Philosophy* 10 (2): 221–40.

——. 2018. "The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology." In *The Social Dimensions of Responsibility*, edited by Marina Oshana, Katrina Hutchinson, and Catriona Mackenzie, 110–36. New York: Oxford University Press.

——. Forthcoming. "Instrumentalist Theories of Moral Responsibility." In *The Oxford Handbook of Moral Responsibility*, edited by Dana Nelkin, and Derk Pereboom.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Wang, Shawn Tinghao. In preparation. "Blame: A Pluralist Function-Based Approach." Unpublished manuscript.