DOI: 10.1017/psa.2023.60

This is a manuscript accepted for publication in *Philosophy of Science*.

This version may be subject to change during the production process.

The Evolutionary Roots of Moral Responsibility

Marcelo Fischborn

Farroupilha Federal Institute of Education, Science, and Technology.

Rua Monteiro Lobato, 4442, 97503-748, Uruguaiana, RS, Brazil.

Email: marcelofischborn@gmail.com

Abstract: Judging a person as morally responsible involves believing that certain responses (such as punishment, reward, or expressions of blame or praise) can be justifiably directed at the person. This paper develops an account of the evolution of moral responsibility judgment that adopts Michael Tomasello's two-step theory of the evolution of morality and borrows also from Christopher Boehm's work. The main hypothesis defended is that moral responsibility judgment originally evolved as an adaptation that enabled groups of cooperative individuals to hold free riders responsible more safely by acting in a coordinated way.

Keywords: moral responsibility; moral responsibility judgment; moral psychology; evolutionary ethics; evolution; punishment

Acknowledgments: For comments on earlier versions of this paper, I thank Rogério P. Severo, Gilberto Gomes, Beatriz Sorrentino Marques, Letícia Palazzo, Gabriel Maruchi, Alex Bastos, Natália Rigue, and especially the anonymous reviewers for this and previous journals.

1. Introduction

Moral responsibility judgment—the conviction that someone is morally responsible for some action—is a central component of human moral psychology. Judging that a person is morally responsible for an action involves believing that it is appropriate to respond to the person with such things as an expression of praise or blame, reward or punishment. Philosophy has addressed moral responsibility at least since Aristotle (2004 [Nicomachean Ethics], Book III). More recently, the topic has also been empirically studied in fields such as social psychology and experimental philosophy. This paper examines moral responsibility judgment from yet another angle, namely that of evolutionary ethics. Even though certain components of the practice of holding people responsible, such as punishment, have been extensively examined in the context of evolutionary theory, the same did not happen to moral responsibility judgment.

The main question posed here is whether and how moral responsibility judgment originally evolved. In order to address it, section 2 makes some terminological options explicit and characterizes contemporary moral responsibility judgment. Section 3 briefly contextualizes the evolutionary project to be developed and examines in more detail an account of the evolution of moral responsibility judgment developed by Matteo Mameli (2013). I argue that Mameli's account suffers from two significant problems. Section 4 presents and defends an alternative account, mainly by reference to Michael Tomasello's (2016) and Christopher Boehm's (2012) broader work on the evolution of human morality. In particular, I adopt Tomasello's two-step framework and discuss how moral responsibility judgment may have originally emerged in the first step and continued to develop by the second. The main hypothesis advanced is that one of the senses of justification involved in moral responsibility judgment was naturally selected because it enabled cooperative individuals to address free riders more safely through a coordination mechanism. Section 5 discusses the main limitations of the account and its potential to explain why moral responsibility judgment is robust and negatively biased.

2. Contemporary Moral Responsibility Judgment

If moral responsibility judgment evolved, then there is a path that goes from its beginning to its current form. This section offers a characterization of contemporary moral responsibility judgment and describes some of its key components.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

Moral responsibility judgment is part of the broader practice of holding people responsible. That practice involves a number of social and psychological elements, including values and beliefs, as well as emotional and behavioral patterns that lie at the core of what we commonly refer to as human morality. I use the term "responsibility episodes" to refer to those events in which someone is held responsible for having done something that is taken to be either good or bad according to some normative standard. It is hard to characterize, in general terms, what makes something a responsibility episode (Zimmerman 2015, 52–54), but there are some standard examples. Some responsibility episodes are negative, such as expressions of blame or punishments, and others are positive, such as praise and reward. The person who holds someone responsible is the author of the responsibility episode, and the one who is held responsible is its target. The author of a responsibility episode is usually different from its target, but there are self-directed responsibility episodes as well, such as self-blame or guilt.[1] Also, the target of a responsibility episode is usually held responsible for having *acted* in some way, but other categories of human phenomena can also motivate responsibility episodes, such as omissions, attitudes, emotional responses, and character traits. For simplicity, I focus on actions. A central role moral responsibility judgment plays is to regulate responsibility episodes.

What is a moral responsibility judgment? At its core, saying that a person is morally responsible for an action means, at the very least, that the person can be the target of some responsibility episode that can be described as deserved, appropriate, or simply *justified* (I will shortly distinguish between two layers of the justification under consideration). Virtually all philosophical accounts of moral responsibility mention that general aspect. I look for further details on empirically informed accounts, since they provide a more accurate view of how moral responsibility judgment figures in present-day human moral psychology. Two such accounts are Malle, Guglielmo, and Monroe's (2014) path model of blame and Hoffman and Krueger's (2017) model of the neural bases of third-party punishment.

According to Malle et al., blame judgments (which they call "cognitive blame") arise in situations where a norm violation is detected, is taken to have been caused by an agent, and the agent has violated the norm either intentionally (but for no good reason) or unintentionally (but had the obligation and capacity to prevent it). Even though the authors

1 Even though self-directed responsibility episodes may be central to human morality, they are not the focus here.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

avoid using the word "responsibility", they take overt expressions of blame (which they call "social blame") to require a justification or warrant (148). Moreover, according to them, the justification of a blame episode coincides to a great extent with the basis for cognitive blame (149).

Hoffman and Krueger's model, in turn, postulates that blame judgment and third-party punishment rely on three neural networks: the salience network, the default mode network, and the central executive network. The salience network "is involved with detecting and responding to norm violations or threats of norm violations" (215). The default mode network "integrates the assessment of the wrongdoer's mental state with the assessment of the harm from the SN [salience network]" and, as a result, produces a "blame signal" (215). And the central executive network is involved in decisions about punishment; here "the blame signal from the DMN [default mode network] is converted into a punishment decision after integration with a wide variety of context-dependent circumstances" (216). In sum, Hoffman and Krueger's neural model says that a punishment episode starts with the detection of a norm violation, proceeds to an assessment of the psychological involvement of the target with the norm violation, and then, under the influence of multiple and context-dependent considerations, culminates in a decision about whether and how to deliver the punishment. I take Hoffman and Krueger's description of the salience and the default mode networks to be largely consistent with Malle et al.'s model of cognitive blame.

The models just considered allow for an understanding of the psychological components of contemporary moral responsibility judgment, at least as it figures in negative responsibility episodes. I take those components to involve at least the following:[2]

1. a capacity to detect norm violations,
2. a motivation to respond in some way to a target who has violated a norm,
3. a capacity to assess the mental involvement of the target with the violation in a way that can affect the motivation to respond, and
4. a capacity to assess contextual factors that are relevant to whether and how to respond to the target in face of the relevant motivation to do so.

2 The next section raises a question about the completeness of the characterization of moral responsibility judgment offered here. I should stress that the components on my list are necessary, but potentially insufficient.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

In light of the those four components, I distinguish between two senses of justification that are present in contemporary moral responsibility judgment. First, a responsibility episode can be said to be justified in a *demand* sense, meaning that there is a positive motivation or demand for realizing the episode (component 2). That sense of justification is emphasized, for example, when we worry about impunity. I believe Malle et al.'s description of cognitive blame and Hoffman and Krueger's characterization of a blame signal are accurate descriptions of the demand sense of justification involved in a negative judgment of moral responsibility (I leave it open whether or not the same characterization works for negative episodes).

In a second sense, a *permission* sense, one can regard a responsibility episode as justified because it is permitted for someone to realize it. The second sense is emphasized when we worry about getting a responsibility episode wrong, something that is captured, for example, in the presumption of innocence. I take the "warrant" Malle et al. (2014, 148) describe as involving the permission sense. And I take the transition from Hoffman and Krueger's blame signal to a punishment episode through the central executive network to leave room for a justification in the permission sense to be part of the contextual circumstances considered (component 4).

In sum, contemporary moral responsibility judgment essentially involves considering whether an actual or potential responsibility episode is justified in two related senses. And at least four psychological components underlie a moral responsibility judgment, two of which are directly linked to the two senses of justification. The next section briefly reviews some of the literature on the evolution of the practice of holding people responsible (especially punishment) and then focuses on an account of the evolution of moral responsibility judgment developed by Matteo Mameli (2013).

3. Mameli on the Evolution of Moral Judgment

Among the components of the practice of holding people responsible, punishment has got most of the attention. In contrast, moral responsibility judgment has seldom been the subject of evolutionary theorization. This section reviews evolutionary perspectives on the practice of holding responsible and punishment, and then focuses on Mameli's (2013) work, which targets the evolution of moral responsibility judgment in a more direct and detailed way. I

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

argue that Mameli's account suffers from two problems and take that as the main motivation for an alternative elaborated in section 4.

Responsibility episodes are ubiquitous in human relations. Most current societies have formal institutions in charge of punishment, which themselves have a long history (see, e.g., Morris and Rothman 1995). And human relations, from family interactions (Laforest 2002) to interactions in the workplace (Aquino, Tripp, and Bies 2001) to interactions among strangers (Svennevig 2012) can involve forms of moral appraisal characteristic of moral responsibility, such as thankfulness, resentment, indignation, acknowledgments, among many others (Strawson 1962; Malle, Guglielmo, and Monroe 2014). There is evidence that children as young as 8 months have a primitive understanding of patterns associated with moral responsibility (Hamlin et al. 2011). And even other species, such as chimpanzees, can sometimes deliver punishments and rewards (Apicella and Silk 2019, R449).

The ubiquity of responsibility episodes, and especially the fact that even other species seem to have rudiments of them, has led many to hypothesize an evolutionary origin. Available theories describe responsibility episodes as part of the mechanisms that allowed for the evolution of the peculiar capacities for cooperation, altruism, and morality found among humans (see, e.g., Boehm 2012; Tomasello 2016; Boyd and Richerson 1992). Responsibility episodes—punishment, in particular—usually figure in those theories as a potential solution to the free rider problem. Because cooperation and altruism can be costly to their authors, and given that non-cooperators can benefit from the cooperation of others without incurring any costs, it may seem puzzling that cooperation ended up being naturally selected. Punishment promises to offer a solution to the puzzle by eliminating the advantage free riders might otherwise obtain.

In contrast to responsibility episodes, judgments of moral responsibility were less often the subject of evolutionary theory. To be precise, Malle et al. (2014), and Hoffman and Krueger, do make connections between their models and evolutionary considerations. Hoffman and Krueger (2017, 211, 217), in particular, hypothesize that third-party punishment was an adaptation to more complex forms of social life and an alternative to the faster and more automatic second-party punishment. Tomasello (2016, 33–34, 61, 67–70) also touches on themes such as resentment, desert, and protest, which are central to moral responsibility judgment. But those connections fall short of providing a more systematic account of the evolution of moral responsibility judgment.

Mameli (2013) addresses the evolution of moral responsibility judgment more directly, in the context of a more general account of the evolution of moral judgment. He

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

understands moral judgment as consisting in a set of emotional dispositions. Simply put, his expressivist account says that a judgment that A (a type of action) is morally required involves four emotional dispositions (see Mameli 2013, 905):

D1: a disposition to feel anger at those who act in ways that violate A;
D2: a disposition to feel guilty about having oneself acted in ways that violate A;
D3: a disposition to feel anger at those who do not have dispositions D1 and D2; and
D4: a disposition to feel guilty about not having dispositions D1 and D2 oneself.

For the present purposes, it is important to note that Mameli sees an understanding of moral responsibility judgment as embedded in the dispositions for meta-anger (D3) and meta-guilt (D4). Those dispositions give rise to what he calls "meriting":

> D3 and D4 may not be manifested very often and may as a result not be particularly salient to people when they casually reflect about morality. But they play an important role. They account for what in the literature is known as *meriting*. One central feature of judging an action to be morally required seems to be that, in addition to being disposed to react in certain ways to violations, we also regard such reactions as *deserved* or *merited* and we regard the lack of such reactions as *inappropriate*… (907)

According to this view, responsibility episodes (at least the negative ones) are ultimately based on anger directed at those who violated some expected behavioral standard. And a judgment of moral responsibility, which is a component of moral judgment according to

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

Mameli, is taken to consist in dispositions for meta-emotions (meta-anger and meta-guilt) directed at those who do not feel anger at first-order violators.[3]

Before assessing the plausibility of Mameli's understanding of moral responsibility judgment and his hypotheses about its evolution, it is worth noting that his account and the one I offer later on have slightly different goals. Mameli starts with an account of a fully-developed *moral* judgment and explores its evolutionary origins (i.e. he is concerned with the evolution of what he takes to be distinctively moral capacities in humans). My account, in contrast, is more centrally concerned with the early evolution of a moral *responsibility* judgment (i.e., with a judgment that guides certain types of responses, especially in the sphere of morality) without addressing its distinctively moral character. Given these differences, some components may be essential for Mameli's account but unnecessary for mine. In other words, it is possible that what I characterize as an original moral responsibility judgment would not count as fully moral by Mameli's standards. I nonetheless expect my account to capture something that eventually became part of a fully developed morality, whatever that may be.

Mameli (2013, 921) situates his hypotheses about the evolution of moral judgment in a broader view according to which human cooperation and altruism evolved in the context of large-game hunting (Boehm 2012). According to that view, around 500,000 years ago our ancestors became collaborative large-game hunters and were thus able to access new sources of food that were unavailable for solitary initiatives. The evolutionary pressure for that change could have been a much earlier scarcity of our ancestors' preferred type of food (plants) due to climate change and competition with other species (Tomasello 2016, 44). The products of large-game hunting, however, opened the door for free riding: uncooperative individuals might take advantage of the collaborative efforts of others. Mameli follows Boehm in counting bullies and cheaters as examples of free riders. In this context, according

3 Outside the context of Mameli's view on moral judgment, moral judgment and moral responsibility judgment are sometimes easier and sometimes harder to separate. A moral responsibility judgment can assert, for example, that an agent deserves punishment for having acted in certain way, while a moral judgment could simply say that what the agent did was wrong. Things get more mixed when we say, for example, that what the agent did is blameworthy.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

to Mameli (2013, 921–22), D1 dispositions emerged as a mechanism that motivated the delivery of punishments, which in turn favored the selection of both cooperators and individuals with D2 dispositions that could work as a form of moral conscience or self-control.

Delivering punishments, however, is another example of a costly activity that opens the door for exploitation, now in the form of a second-order free rider problem: individuals can be better off by benefiting from punishment implemented by others, and hence each individual has an incentive not to punish despite the collective benefits of someone acting otherwise (see, e.g., Henrich and Boyd 2001, 80; Boyd and Richerson 1992). It is in the context of the second-order free rider problem that Mameli places judgments of moral responsibility. He hypothesizes that meta-emotional dispositions D3 and D4 would be able to ensure that second-order free riders are punished just like first-order ones.

Mameli's account is a valuable attempt to develop a fine-grained description of the components of morality within the broader context of evolutionary theories. That virtue notwithstanding, the account suffers from two main problems that are crucial for an account of the evolution of an early moral responsibility judgment. One is that the evolutionary hypothesis about the selection of second-order dispositions is weakly supported. And another is that analyzing meriting in terms of second-order emotional dispositions is conceptually implausible.

On the first problem, Mameli relies on Boehm's work to a great extent, but he disagrees about how exactly punishment practices evolved. According to Boehm, the solution to the first-order free rider problem that evolved in our species was already able to prevent the emergence of a second-order problem. Boehm relies on ethnographies of what he calls "Late Pleistocene Appropriate" (LPA) groups, which are a selection of present-day foraging societies taken to represent a close approximation of how humans lived in Africa around 45,000 years ago (Boehm 2012, 79). As the ethnographies attest, LPA groups can control actual and potential cheaters and bullies through a collective punitive enterprise. By acting as a group, greater punishment power is available and retaliation is discouraged. Also, Boehm notes that the ethnographies have no mention of second-order punishment: "so far in my survey of group punishment among fifty LPA hunter-gatherers […], the punishment of nonpunishers is never mentioned in the hundreds of ethnographies even though punishment does take place so regularly—and even though there are plenty of abstentions" (206). If first-order dispositions evolved and made first-order punishment so common, why are manifestations of second-order dispositions so uncommon?

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

In response, Mameli says that

> The fact that ethnographies do not report the punishment of those who are not involved in specific occasions in carrying out specific punitive acts is unsurprising and does not in any way show that most people in the band are second-order free riders, *as Boehm seems to suggest* [1]. It would make no sense to punish those who support (with their vigilance and disapproval) those who have been given the task to carry out the group-mandated punitive acts […]. In LPA bands, it is not just that everyone disapproves of the deviant, but *everyone is expected to disapprove of the deviant* [2]. If you do not show some disapproval of the deviant, or are unwilling—if mandated by the band—to participate in carrying out some punitive acts or to support and protect those who have been assigned the task to carry out the punitive acts, then you will be disapproved of, and such disapproval will have negative consequences on your reputation (2013, 927, emphasis added).

I think the passages emphasized above include a misinterpretation of what Boehm says (1) and an unsupported hypothesis (2). Contrary to the first part, Boehm does not seem to suggest that most, or even some, people are second-order free riders just because they do not get involved in some particular punishment episode, despite supporting its execution. More than that, Boehm (2012, 206) explicitly acknowledges that eventual "abstentions need have no relation to free-rider genes". Thus, his view seems to be that, even though some experimental studies suggest that second-order punishment can stabilize cooperation, observations in settings with greater ecological validity suggest that it is not required.

In the second passage, Mameli claims that everyone is expected to disapprove of a deviant and that violations of that expectation would be met with negative consequences. But he fails to provide any evidence (ethnographic or otherwise) in support of that claim, which is, in addition, somewhat at odds with some cases discussed by Boehm. In an instance of collective punishment, for example, members of the group that lived close to the one being targeted "were obviously staying to one side and appeared to be neutral" (Boehm, 2012, 206); the explanation offered is that "close relatives or associates of a deviant may choose to stand back and let others deal with him harshly". Boehm suggests that other group members could reasonably accept that type of situation as part of social expectations related to certain roles, including family relations. Abstentions like that suggest that second-order dispositions were

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

not needed to stabilize first-order punishment or, at the very least, that some evidence is needed to establish that they were.

Again, Mameli's account purports to capture an essential feature of contemporary moral judgment and it is possible that second-order dispositions are some of those features. Also, I do not dispute that second-order dispositions could, in principle, help to stabilize first-order punishment. Even so, the reasons Mameli offers to support the claim that those dispositions were part of the evolutionary history of moral responsibility judgment and punishment episodes seem uncompelling.[4]

A second difficulty for Mameli's account, insofar as it addresses moral responsibility judgment, has to do with his characterization of meriting. As he rightly says, meriting consists in viewing responsibility episodes "as *deserved* or *merited*" (907). But it is conceptually implausible to identify a judgment that a response to some behavior is deserved with a disposition to feel anger at some (potentially different) individual who is not disposed to feel anger at that behavior. For example, Mameli says that "regarding guilt and anger as merited reactions toward a violation *just is* having dispositional anger at meta-violators" (907). If meta-dispositions are necessarily involved in full-blown moral judgment, they would also be present in judgments saying that a response to a genuine moral violation is deserved. The problem is that judgments of moral responsibility are first and foremost about first-order violators themselves and only secondarily about how others should react to a violation. For example, the models by Malle et al. and Hoffman and Krueger, which aim to provide an empirically accurate portrait of moral responsibility judgment in humans (and the sense of "warrant" it involves), do not mention an assessment of individuals other than first-order violators. In the same way, when we say a person deserves a response, we are primarily saying the response is correct for that exact person.

In short, Mameli's account ultimately describes moral responsibility judgment as a disposition for second-order anger (a disposition to feel anger at those who are not disposed to feel anger at first-order violators) whose evolutionary function was to prevent second-order free riding. More concretely, a sense of meriting consists in an expectation that everyone disapproves of deviants, and is willing to execute an act of punishment, if requested by the group, and to support those who execute punishment against retaliation. I have questioned, on two related grounds, how credible this specific part of Mameli's account is. The next section

4 Section 5 briefly suggests an alternative place for Mameli's second-order dispositions.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

offers an alternative which I take to rest on a more accurate description of moral responsibility judgment and to be better aligned with available works in evolutionary ethics.

4. A Two-Step Account of the Evolution of Moral Responsibility Judgment

This section offers an account of how the two senses of justification involved in contemporary moral responsibility judgment may have evolved. The account shares some assumptions with Mameli's account, although there are important differences in the details. I agree with Mameli that moral responsibility judgment was part of the evolutionary mechanism that stabilized responsibility episodes. I also agree that part of the reason why moral responsibility judgment (in my view, in the demand sense) was selected is that it motivates the realization of responsibility episodes. Beyond that, the main difference of the present account is that it takes the permission sense of justification as more relevant to avoid the second-order free rider problem. My main hypothesis—the social support hypothesis—says that the permission sense arose as part of a mechanism that allowed for the social support enjoyed by a potential responsibility episode to be assessed, in a way that could influence the realization of the episode. The account relies to a great extent on Tomasello's two-step theory of the evolution of human morality and on Boehm's description of punishment practices in LPA bands.

Tomasello (2016) postulates that two steps were crucial for the evolution of the specific form of cooperation found among humans, both of which resulted from an increased interdependence among individuals. Early humans, who lived around 400,000 years ago, had as their main innovation the collaborative hunt in the context of a dyad. That accomplishment depended on new capacities, such as joint intentionality, second-personal agency, and joint commitment, which enabled individuals, for example, to pay attention to common objects in their environment, and to understand the perspective of their partners and the roles they each should play in a collaborative activity (2016, 50). Modern humans, in their turn, who began to develop around 150,000 years ago, achieved the capacity to live in larger and more diverse groups, where cooperation required, among other things, better communication capacities. While individuals in the first step learned to cooperate with well-known partners within small groups, humans in the second step needed to learn how to collaborate with less well-known members of groups that were forging their own cultural identities and that eventually needed to compete with rival groups (2016, 85).

It is relevant to emphasize how Tomasello's account attempts to explain how human morality and cooperation may have become stable and immune to free rider problems.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

Tomasello focuses on processes of mutualism and reciprocity that operate at the individual level in contexts of increased interdependence among individuals, i.e., where collaboration benefits everyone involved (Tomasello 2016, 13–19; Tomasello et al. 2012). One advantage of the mutualistic explanation, in particular, is that free riding, although it still poses some challenges, becomes less salient (see, e.g., Tomasello 2016, 13, 61). The account also posits mechanisms of social selection, according to which individuals who do well in collaborative enterprises—e.g., by communicating well and being helpful to partners—may become preferred partners of joint collaboration and have, as a consequence, increased access to the benefits of collaboration. Tomasello argues that his account is more plausible than classical alternatives, including hypotheses that involve group selection through, for example, competition between groups (Tomasello 2016, 12).

My account centrally hypothesizes that moral responsibility judgment began to develop in the first evolutionary step Tomasello describes. During the first stage, moral responsibility judgment in the demand sense consisted in a capacity to detect violations of proto-norms coupled with a motivation to deliver a (sometimes punitive) response. The permission sense also began to develop in the first step as part of a channel of mutual understanding among cooperators, which enabled them to coordinate a punitive response against free riders. A secondary hypothesis (see section 5) says that moral responsibility judgment is unlikely to have achieved its contemporary form, including its characteristic sensitivity to the mental states of its targets, before the second evolutionary stage. Within the second step, responsibility practices as a whole became more complex as human groups became bigger and cultural. Scrutinizing the mental lives of potential targets of responsibility episodes makes sense in that context, I argue, even though it remains open how close to current assessments of intentions and omissions that scrutiny would be.

In the context of dyadic hunting, responsibility episodes could fit in two main types of situations. First, one could face a greedy hunting partner's attempt at taking more than half of the spoils. Tomasello thinks attempts at non-equal sharing in this context violate an implicit agreement that underlies the whole collaboration. The tacit commitment is that each party should play their role properly for both to achieve their common goal; it also involves accepting that both partners have authority to initiate sanctioning when their joint commitment is violated (2016, 68). The paradigmatic response to violations of this type is a form of resentful, albeit respectful, protest, which "does not seek to punish the partner directly, only to inform her of the resentment, assuming her to be someone who knows better than to do this" (69). That kind of protest requires simple communicative skills and can be

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

expressed by "a simple 'Hey!' or a squawk" (69). As a result, the violator is expected to acknowledge and repair her own fault—thus keeping her cooperative identity—or, less likely given the implicit agreement, to face the threat of being excluded from future collaboration on which her survival may depend.

I take the expression of protest in the scenario just described as an early instance of a moral responsibility episode. But I think that scenario plays a less central role in the evolution of moral responsibility judgment than it plays in the whole story about human morality, as told by Tomasello. A second type of situation is more relevant for moral responsibility judgment, one that Tomasello also describes, although in less detail.

The second type of situation humans of the first step faced was the threat of free riders, i.e. individuals who would attempt to take some of the spoils without having taken part in the hunt themselves. This type of situation arises only after collaborative hunting is achieved: "others could come up after the kill, and they were essentially competitors—from outside the collaboration—so at some point humans also evolved the tendency to deter […] free riders by denying them a share of the spoils" (2016, 61). Tomasello elaborates on how the control of this type of free riding could have produced a sense of desert that later could also guide the division of resources among collaborators. But he does not elaborate on how the denial of a share of the spoils to non-collaborators could be carried out.

In contrast to the situation in which a greedy partner attempts to take more than an equal share, the threat of a non-collaborator is potentially more dangerous. In the context of a collaborative dyad, a vocal protest might suffice because the greedy individual may just need a reminder of a tacit agreement she already made. But, lacking one such agreement, the interaction among collaborators and external free riders might easily involve physical violence from either part. My suggestion is that it would be advantageous for cooperators to have, in addition to a motivation to respond negatively to the violator, a mutual understanding of the fact that they both shared (or not) that motivation.

Returning to Hoffman and Krueger's model, I propose that a primitive moral responsibility judgment in the demand sense arose when early collaborators developed two dispositions that were reactive to having one's own or one's partner's food threatened by external individuals. The first is a disposition of the salience network to display an aversive emotional response, which would work as a proto-norm against food robbery. The second is a disposition to display an aggressive response against violators of that proto-norm, in a way that can influence the central executive network. I propose that the assessment of a violator's mental states (e.g., intentions) that is constitutive of contemporary moral responsibility

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

judgment was absent at this early stage. That is to say, even though some understanding of the mental states of the external violator could be present, it was not, at this point, a factor that could make someone refrain from realizing a responsibility episode. Accordingly, the initial situation was somewhat closer, albeit not identical, to Hoffman and Krueger's (2017, 217) characterization of second-party punishment "as being blame plus an automatic punishment response, all rolled into one, and without the cognitive restraints we see with third-party punishment". In a sense, early humans' punitive response to a free rider was a second-party response, as Tomasello describes them as acting as a collective "we". If this same "we" reacts, then it is implementing second-party punishment, according to the definition (Hoffman and Krueger 2017, 207). But, in another sense, the situation was different because the reaction against the violator was less individualistic and automatic than Hoffman and Krueger take second-party punishment to be.[5] Thus, the initial moral responsibility judgment provided early large-game hunters with the ability to detect a threat to their collective goal and with the motivation to respond aggressively.

On its own, however, a moral responsibility judgment in the demand sense is not likely to become evolutionarily stable, as it does invite the second-order free rider problem. Given the possibility of retaliation from the free rider, each collaborator would be better-off by refraining to punish, if the other party punished alone. A potential solution is available if collaborators, in addition to detecting a violation and being motivated to respond negatively, were also able to assess the motivation of their partners. That would be the birthplace of moral responsibility judgment in the permission sense. The potential authors of responsibility episodes would have not just the motivation to realize those episodes, but also the sensitivity

5 Hoffman and Krueger (2017, 208) see second-party punishment as "widespread throughout the animal kingdom" and include as examples things such as "algae that fire projectiles at would-be predators" and immunological responses. According to them, the key drive for the evolution of third-party punishment is that it provided a slower, more reflective sort of response that allowed for the consideration of the costs and benefits of a punitive episode (211). My account does not address those earlier and fully automatic forms of punishment.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

to the social support enjoyed by the episode, in a way that makes them less open to exploitation.

Even though I assume there was not, at this early stage, an assessment of the mental states of the violator, there are good reasons to assume that collaborators in the first stage could have a capacity to assess the support for a responsibility episode. The context of the joint activity involved a channel of mutual understanding among collaborators themselves (Tomasello 2016, 53). Within this context, then, and under the influence of a first-personal motivation to punish, collaborative hunters could form a further joint goal to protect their collective achievements by punishing eventual attackers. In a sense, that would be simply another intermediary step within their whole joint activity: just as their mutual understanding allowed them to coordinate a set of actions to hunt successfully, their understanding of each other's motivation to punish would allow them to respond to the attacker as a collective "we". One such coordinated sort of response would increase the safety and success of collaborators by reducing the risk of retaliation and by increasing punishment capacity. And because the punitive response was part of a broader collaboration for food each one was already invested in, defecting would be self-defeating. Therefore, even if a moral responsibility judgment in the demand sense alone was not likely adaptive and stable, the addition of the permission sense, in the form of an assessment of the social support for the responsibility episode under consideration, makes adaptiveness and stability more likely. This hypothesis, I propose, offers a more plausible starting point for moral responsibility judgment than the one involving Mameli's second-order dispositions.

Within Tomasello's second evolutionary step, humans formed larger and culturally structured groups. The social support hypothesis fits in that context as well. Collective life in bigger and cultural groups likely affected the practice of holding people responsible. First of all, there were new opportunities for free riding, as cheaters can be harder to detect in bigger groups, where interactions among strangers become more frequent (Tomasello 2016, 98). A cultural organization also means that more collective norms and social control are required to maintain more complex forms of cooperation. In addition to new free riding opportunities and more norms, two other aspects of life in bigger cultural groups could have been relevant. One is what Tomasello refers to as a "common ground", i.e., a shared knowledge base which "meant that everyone in a group knew that everyone in that group had had certain kinds of experiences—and thus skills, knowledge, and beliefs" and that individuals knew "many important things about the minds and likely behavior or others, often without ever interacting with them directly" (93). Another element of cultural life is knowledge of the linguistic

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

conventions that guide communication within the group (95). A shared knowledge base that includes knowledge about mental and behavioral tendencies of other group members and linguistic communication paved the way, among other things, for a new way of assessing the social support enjoyed by responsibility episodes. Boehm's account of gossiping among hunter-gatherers helps to illustrate both points.

Following studies by anthropologist Polly Wiessner, Boehm (2012, 240) notes that the !Kung people "gossip intensively when trouble is shaping up and collective action may have to be taken":

> It's by adding up information that social deviants are identified and people can unite to cope with them. Without safe, private gossiping, free-rider suppression would not be likely to work very effectively in the case of scary bullies, because only a united group is a confident and safe group, and such political unity comes out of finding a consensus. (2012, 240–41)

According to these passages, gossiping helps the group both to detect violations of norms and to coordinate a collective (and consensual) response. The details of that type of communication are illustrated by a specific episode of norm violation Boehm describes, this time based on Colin Turnbull's work on the Mbuti Pygmies. In a collective net hunt—where each hunter would get for his family what he could catch on his net but where cooperation would allow for everyone to succeed—an egoistic man named Cephu quietly repositioned his net in such a way that animals driven by other group members would run first into his net. The hunt worked well for Cephu, but not the aftermath, as his cheating had been witnessed. The episode ended up with Cephu being strongly shamed by the group, facing the threat of expulsion, apologizing, and having to return all of the spoils that he and his family intended to eat. The details of the episode are relevant to the social support hypothesis. Just after the hunt, when the group was returning to the camp in a quiet and bad mood because Cephu's behavior,

> an adult male, Kenge, said to the group, "Cephu is an impotent old fool. No, he isn't, he is an impotent old animal—we have treated him like a man for long enough, now we should treat him like an animal. Animal!" This statement broke the ice, and some serious gossiping began as the score was carefully added up and a group consensus materialized. The result of Kenge's tirade was that everyone calmed down and began criticizing Cephu a little less

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

> heatedly, but on every possible score: the way he always built his camp separately, the way he had even referred to it as a separate camp, the way he mistreated his relatives, his general deceitfulness, the dirtiness of his camp, and even his own personal habits. (Boehm 2012, 38–39)

This is a vivid example of a punishment episode within a group of hunter-gatherers, one that exemplifies a moral responsibility judgment, in the permission sense, playing a role in coordinating a group response that becomes safer for its authors. While individuals could be previously motivated to respond negatively to Cephu, realizing that other group members were similarly motivated helped the group to execute the punitive episode.

In summary, the present account tells a story about how different components of moral responsibility judgment evolved along the two evolutionary stages Tomasello describes. The demand sense evolved as an individual motivation toward realizing negative responsibility episodes against free riders. The permission sense evolved in a way that enabled individuals to coordinate their responses against free riders by assessing the social support enjoyed by a responsibility episode they felt motivated to realize. The permission sense made punishment episodes safer for those implementing them and less open for exploitation. In the first evolutionary stage, social support assessment was part of the mutual understanding among dyadic cooperators; during the second stage, it became linguistic and spread over the group as a whole through the practice of gossiping.

5. Limitations and Explanatory Potential

The account just presented has its own limitations, but it also has some explanatory power. Both are discussed below.

Limitations. The first and most important limitation of the evolutionary account developed thus far is that is leaves unexplained the origin of the contemporary moral responsibility judgment's characteristic sensitivity to the mental states of its targets. Did the sensitivity to intentions evolve and, if so, when and why? Although it does not answer the question directly, I think the present account at least provides three suggestions that may be relevant for the answer. First, the sensitivity to the mental states of potential targets of responsibility episodes is unlikely to have evolved before Tomasello's second evolutionary step. As Tomasello says, the second step creates conditions for the existence of a cultural "common ground" that is relevant for communication and for an understanding of the mental

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

states and experiences of other group members. In the context of the practice of holding responsible, Boehm's discussion of Cephu's case also evidences the consideration of a larger portrait of his faults and, importantly, elements of his psychological profile, his character or "reputation" (Tomasello 2016, 100): the *way* he used to do certain things, his *general* deceitfulness, and other of his *habits*—something Boehm (2012, 167) further describes as "long-term patterns of malfeasance" (167). Even though those considerations still fall short of showing that intentions play a decisive role in the judgment that someone is morally responsible, they are evidence of an increased relevance of the mental life of the target of a potential responsibility episode.

A second suggestion has to do with how intentions help to define what an agent did—as G. E. M. Anscombe (1957) emphasized, actions can be intentional "under a description" but not under others. In a context of more complex cultural norms and relations among group members, knowing more precisely what agents are trying to do may become relevant to the detection of norm violations. Especially in the context of responsibility practices, holding someone responsible may involve punitive acts that are externally (and in isolation) indistinguishable from the very kinds of norm violations they respond to. Consider, for example, how a murder is to be distinguished from capital punishment or a self-defense killing.

A third suggestion is that intentions, in the context of larger and culturally structured groups, may play a role in the assessment of group membership. Mameli's second-order dispositions may help to illustrate the point. From Tomasello's second-step onwards, group identity and individuals' reputation within the group became critical. Studies (McDonald et al. 2017) have suggested that perceiving an outgroup member as having an emotional reaction to an anger-eliciting type of situation that is similar to one's own leads to a more humane and tolerant view of the outgroup member. A possibility, then, is that Mameli's second-order dispositions contributed to give morality a role in the formation of cultural group identities and in the management of group-membership, more than they helped to stabilize early responsibility episodes. It is conceivable that attention to the intentions of agents could play a similar role: one who intentionally fails, say, to share food as expected by the group is not just showing disregard for the victims of his behavior, but is also showing a disregard for the values of the group. Thus, violating a norm of the group *intentionally* is, in a sense, a way of signaling dissonant values and distance from the group's culture. The above considerations surely fall short of saying whether, when, and why the sensitivity to the mental states that characterizes contemporary moral responsibility judgment evolved. But they do

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

offer some hints about how capacities that are involved in that sensitivity may have started to affect the success of individuals once they were living in cultural groups.

A second limitation of the account developed here is its considerable (and somewhat unavoidable) degree of speculation. It is one thing for an account to offer an evolutionary model for a phenomenon like moral responsibility judgment, and quite another to show how the phenomenon actually evolved (Machery and Mallon 2010, 16). My account attempts to get closer to the actual evolution of moral responsibility judgment, but it does so by aligning with some recent evolutionary accounts of human morality and cooperation. As such, it is at most as plausible as those accounts.

Another limitation is that the account developed does not even attempt to explain what makes moral responsibility judgment (and the whole practice of holding responsible) *moral*. On a purely conceptual level, some different claims can be made: one can say that, e.g., a punishment episode has a moral *justification* (i.e., the criteria for justification are of a moral type and not, say, legal), or one can say that it is a justified moral *reaction* (the type of punishment is moral, as shame or expulsion, and not, say, a fine or prison time), or one can say that the reaction is justified because of the violation of a moral *norm* (Zimmerman 2015, 54 offers a similar distinction concerning the first two possibilities). Machery and Mallon (2010, 22) note that showing that specifically moral cognition evolved can be more difficult than showing that some general type of normative cognition evolved. Tomasello (2016, 122), for example, does attempt to explain what makes something a moral norm in terms of its being based on a cultural endorsement of earlier second-personal values, and Boehm (2012, 15) suggests that social control was initially nonmoral. There is also Mameli's account of moral judgment as necessarily involving second-order dispositions. Despite these distinctions and possibilities, I want to leave it open the question of whether the evolution of moral responsibility judgment, as I have described it, was the evolution of something distinctively moral. For the same reason, my account is consistent with the claim that the practice of holding people responsible may serve apparently immoral goals under certain conditions (Raihani and Bshary 2019). I limit myself to the claim that, to the extent that the account is plausible, it shows that some key components of contemporary moral responsibility judgment (which can be insufficient to account for its moral character) evolved.

Explanatory potential. The account offered here helps to explain some facts about human moral psychology and responsibility practices. The first, and most obvious, is the robustness of the assumption that people are usually morally responsible for their actions and, as a result, apt targets of expressions of praise, blame, reward, or punishment. Section 2

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

described how widespread the practice of holding people responsible is: it is present across times and places, and it is learned by children early on. What I add here is just how robust and stable the practice and the assumptions it involves are. Studies on the related topic of belief in free will have documented that people's belief in human freedom is stable and even hard to experimentally manipulate (Schooler et al. 2015, 77; see also Fischborn 2018, 50–51). Some of those studies also include a measurement of beliefs related to moral responsibility and, unsurprisingly, they seem to be even more robust than beliefs about free will. In one study, for example, where belief in free will *was* significantly lowered, no statistically significant change was observed in how much participants blamed another (fictitious) participant for stealing money nor in how much punishment they thought the stealer should receive (Monroe, Brady, and Malle 2017, 193–94).

The present account explains both the universality and robustness of the practice of holding people responsible and its associated assumptions: moral responsibility judgment was selected precisely because it helped to make responsibility episodes viable and, as consequence, a stable practice of the human species. Part of the challenge was to deal with free riders in a way that reduced the risks for cooperators. Moral responsibility judgment, then, was from the start heading toward the affirmation of moral responsibility.

A second, and less evident, aspect of responsibility practices is what can be described as a negative bias. One manifestation of this bias is known as the Knobe-effect (Knobe 2003, 193). The bias manifests itself as a greater readiness to see morally negative side-effects of people's actions as intentional in comparison to morally positive ones, as well as in a greater willingness to blame the agents whose actions bring about negative side-effects than to praise the agents whose actions bring about positive side-effects. The Knobe-effect is robust and has been replicated and extended to different contexts (Cova et al. 2018, Appendix 1; Michael and Szigeti 2019). A related bias manifests itself in greater disposition to see the authors of negative actions as free and responsible for their actions in comparison to the authors of positive actions (Clark et al. 2014; for discussion, see also Monroe and Ysidron 2021; Clark, Winegard, and Shariff 2021). Jay Wallace (1996, 61) also suggests that "praise does not seem to have the central, defining role that blame and moral sanction occupy in our practice of assigning moral responsibility".[6]

---

6 This bias also spills on theoretical investigations. Matthew Talbert (2019) notes that the philosophical 'attention given to blame far exceeds that given to praise'. Similarly, Daniel

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

The negative bias at issue can also be explained in light of the present account. Although disputes remain, there is evidence that punishment and reward (as representatives of negative and positive responsibility episodes) can promote cooperation with statistically indistinguishable strength (Balliet, Mulder, and Van Lange 2011). So the negative bias should not be explained by an asymmetry in the impact positive and negative episodes could have on cooperation. Rather, the way moral responsibility judgment evolved was largely shaped by the challenges the realization of negative responsibility episodes involves. Those challenges include the risks of retaliation, which would require more coordination from the group, especially when the target is an aggressive or stronger individual, but also eventual difficulties for detecting important violations which their authors would have good reason to hide. While it is conceivable that the expression of praise and the distribution of rewards may eventually invite worries about fairness or equity, for example, those worries would likely be much less pressing than the ones related to being punished or suffering retaliation. Returning to Hoffmann and Krueger (2017), given the risks of missing the behavior of a bully or a free rider, it makes sense if a salience network is readier to trigger other neural events that may end up in the realization of a blame or punishment episode than it would be to trigger expressions of praise and reward. Therefore, it seems to make sense that our present moral psychology includes dispositions that are reactive to events categorized as negative that are more easily triggered than ones that are reactive to events categorized as positive, given the challenges and pressures under which the relevant capacities evolved. In other words, it is a suggestion of the present account that the negative bias does not arise because positive and negative actions are equally scrutinized and then the negative ones are judged to involve more responsibility. Instead, the suggestion is that negative actions, because they were evolutionarily more pressing, are more readily and attentively scrutinized and, as a consequence, more likely to be found to justify a response.

As a final note, and adding to the potential relevance of the present account, the practice of holding responsible, despite being central in human life, can be imperfect in many ways. For that reason, many projects are concerned with the modification of the practice (see, e.g., Caruso and Pereboom 2020; Gonzalez et al. 2019; Nadelhoffer 2006; Waller 2015). Part of the challenge those projects face is the stability and robustness of the practice. To the

---

Balliet, Laetitia Mulder, and Paul Van Lange (2011) observe that there are more empirical studies on punishment than on rewards.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

extent that proposed changes to the responsibility system are beneficial, the reasons why it is so stable need to be better understood and taken into consideration in order for the modifications to be viable. The evolutionary origins of morality in general, and of moral responsibility judgment in particular, are part of those reasons, and I hope the present account contributes to a better understanding of them.

References

Anscombe, G. E. M. 1957. *Intention*. Oxford: Basil Blackwell.

Apicella, Coren L., and Joan B. Silk. 2019. "The Evolution of Human Cooperation." *Current Biology* 29 (11): R447–50. https://doi.org/10.1016/j.cub.2019.03.036.

Aquino, Karl, Thomas M. Tripp, and Robert J. Bies. 2001. "How Employees Respond to Personal Offense: The Effects of Blame Attribution, Victim Status, and Offender Status on Revenge and Reconciliation in the Workplace." *Journal of Applied Psychology* 86 (1): 52–59. https://doi.org/10.1037/0021-9010.86.1.52.

Aristotle. 2004. *Nichomachean Ethics*. Translated by Roger Crisp. Cambridge: Cambridge University Press.

Balliet, Daniel, Laetitia B. Mulder, and Paul A. M. Van Lange. 2011. "Reward, Punishment, and Cooperation: A Meta-Analysis." *Psychological Bulletin* 137 (4): 594–615. https://doi.org/10.1037/a0023489.

Boehm, Christopher. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Boyd, Robert, and Peter J. Richerson. 1992. "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and Sociobiology* 13 (3): 171–95. https://doi.org/10.1016/0162-3095(92)90032-Y.

Caruso, Gregg D., and Derk Pereboom. 2020. "A Non-Punitive Alternative to Retributive Punishment." In *Routledge Handbook of the Philosophy and Science of Punishment*, edited by Farah Focquaert, Bruce Waller, and Elizabeth Shaw, 355–65. Routledge.

Clark, Cory J., Jamie B. Luguri, Peter H. Ditto, Joshua Knobe, Azim F. Shariff, and Roy F. Baumeister. 2014. "Free to Punish: A Motivated Account of Free Will Belief." *Journal of Personality and Social Psychology* 106 (4): 501–13. https://doi.org/10.1037/a0035880.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press

Clark, Cory J., B. M. Winegard, and A. F. Shariff. 2021. “Motivated Free Will Belief: The Theory, New (Preregistered) Studies, and Three Meta-Analyses.” *Journal of Experimental Psychology. General* 150 (7): e22–47. https://doi.org/10.1037/xge0000993.

Cova, Florian, Brent Strickland, Angela Abatista, Aurélien Allard, James Andow, Mario Attie, James Beebe, et al. 2018. “Estimating the Reproducibility of Experimental Philosophy.” *Review of Philosophy and Psychology*, June. https://doi.org/10.1007/s13164-018-0400-9.

Fischborn, Marcelo. 2018. “How Should Free Will Skeptics Pursue Legal Change?” *Neuroethics* 11 (1): 47–54. https://doi.org/10.1007/s12152-017-9333-8.

Gonzalez, Miriam, Christine A. Ateah, Joan E. Durrant, and Steven Feldgaier. 2019. “The Impact of the Triple P Seminar Series on Canadian Parents’ Use of Physical Punishment, Non-Physical Punishment and Non-Punitive Responses.” *Behaviour Change* 36 (2): 102–20. https://doi.org/10.1017/bec.2019.7.

Hamlin, J. K., K. Wynn, P. Bloom, and N. Mahajan. 2011. “How Infants and Toddlers React to Antisocial Others.” *Proceedings of the National Academy of Sciences* 108 (50): 19931–36. https://doi.org/10.1073/pnas.1110306108.

Henrich, Joseph, and Robert Boyd. 2001. “Why People Punish Defectors.” *Journal of Theoretical Biology* 208 (1): 79–89. https://doi.org/10.1006/jtbi.2000.2202.

Hoffman, Morris B., and Frank Krueger. 2017. “The Neuroscience of Blame and Punishment.” In *Self, Culture and Consciousness*, edited by Sangeetha Menon, Nithin Nagaraj, and V. V. Binoy, 207–23. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-5777-9_13.

Knobe, Joshua. 2003. “Intentional Action and Side Effects in Ordinary Language.” *Analysis* 63 (279): 190–94. https://doi.org/10.1111/1467-8284.00419.

Laforest, Marty. 2002. “Scenes of Family Life: Complaining in Everyday Conversation.” *Journal of Pragmatics* 34: 1595–1620. https://doi.org/10.1016/S0378-2166(02)00077-2.

Machery, Edouard, and Ron Mallon. 2010. “Evolution of Morality.” In *The Moral Psychology Handbook*, edited by John Michael Doris, 3–46. Oxford University Press.

Malle, Bertram F., Steve Guglielmo, and Andrew E. Monroe. 2014. “A Theory of Blame.” *Psychological Inquiry* 25: 147–86. https://doi.org/10.1080/1047840X.2014.877340.

Mameli, Matteo. 2013. “Meat Made Us Moral: A Hypothesis on the Nature and Evolution of Moral Judgment.” *Biology & Philosophy* 28 (6): 903–31. https://doi.org/10.1007/s10539-013-9401-3.

McDonald, Melissa, Roni Porat, Ayala Yarkoney, Michal Reifen Tagar, Sasha Kimel, Tamar Saguy, and Eran Halperin. 2017. “Intergroup Emotional Similarity Reduces Dehumanization and Promotes Conciliatory Attitudes in Prolonged Conflict.” *Group Processes & Intergroup Relations* 20 (1): 125–36. https://doi.org/10.1177/1368430215595107.

Michael, John Andrew, and András Szigeti. 2019. “‘The Group Knobe Effect’: Evidence That People Intuitively Attribute Agency and Responsibility to Groups.” *Philosophical Explorations* 22 (1): 44–61. https://doi.org/10.1080/13869795.2018.1492007.

Monroe, Andrew E., Garrett Brady, and Bertram F. Malle. 2017. “This Isn’t the Free Will Worth Looking for: General Free Will Beliefs Do Not Influence Moral Judgments; Agent-Specific Choice Ascriptions Do.” *Social Psychological and Personality Science* 8 (2): 191–99. https://doi.org/10.1177/1948550616667616.

Monroe, Andrew E., and Dominic W. Ysidron. 2021. “Not so Motivated after All? Three Replication Attempts and a Theoretical Challenge to a Morally Motivated Belief in

Free Will." *Journal of Experimental Psychology. General* 150 (1): e1–12. https://doi.org/10.1037/xge0000788.

Morris, Norval, and David J. Rothman. 1995. "Introduction." In *The Oxford History of the Prison*, edited by Norval Morris and David J. Rothman, vii–xiv. Oxford: Oxford University Press.

Nadelhoffer, Thomas. 2006. "Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality." *Philosophical Explorations* 9 (2): 203–19. https://doi.org/10.1080/13869790600641905.

Raihani, Nichola J., and Redouan Bshary. 2019. "Punishment: One Tool, Many Uses." *Evolutionary Human Sciences* 1: e12. https://doi.org/10.1017/ehs.2019.12.

Schooler, Jonathan, Thomas Nadelhoffer, Eddy Nahmias, and Kathleen D. Vohs. 2015. "Measuring and Manipulating Beliefs and Behaviors Associated with Free Will: The Good, the Bad, and the Ugly." In *Surrounding Freedom: Philosophy, Psychology, Neuroscience*, edited by Alfred R. Mele, 72–94. New York: Oxford University Press.

Strawson, Peter F. 1962. "Freedom and Resentment." In *Free Will*, edited by Derk Pereboom, 148–71. Indianapolis: Hackett.

Svennevig, Jan. 2012. "On Being Heard in Emergency Calls. The Development of Hostility in a Fatal Emergency Call." *Journal of Pragmatics* 44 (11): 1393–1412. https://doi.org/10.1016/j.pragma.2012.06.001.

Talbert, Matthew. 2019. "Moral Responsibility." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Vol. Winter 2019. Stanford: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/.

Tomasello, Michael. 2016. *A Natural History of Human Morality*. Cambridge, Mass.: Harvard University Press.

Tomasello, Michael, Alicia P. Melis, Claudio Tennie, Emily Wyman, and Esther Herrmann. 2012. “Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis.” *Current Anthropology* 53 (6): 673–92. https://doi.org/10.1086/668207.

Wallace, R. Jay. 1996. *Responsibility and the Moral Sentiments*. Cambridge, Mass.: Harvard University Press.

Waller, Bruce N. 2015. *The Stubborn System of Moral Responsibility*. The MIT Press. https://doi.org/10.7551/mitpress/9780262028165.001.0001.

Zimmerman, Michael. 2015. “Varieties of Moral Responsibility.” In *The Nature of Moral Responsibility: New Essays*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 45–64. New York: Oxford University Press.

https://doi.org/10.1017/psa.2023.60 Published online by Cambridge University Press