Intention-Sensitivity in the Evolution of Moral Responsibility Judgment

True Gibson*

Abstract

I provide an evolutionary explanation for how and why intentionality came to figure centrally in human moral responsibility judgment (MRJ). I show that being sensitive to the intentionality of others' actions pre-dates moral cognition, which made it available to the earliest forms of MRJ. I then argue that intention-sensitivity also increased the fitness benefits of MRJ, particularly with respect to adaptive partner choice, by improving evaluations about the desirability of potential cooperative partners. Since intention-sensitivity was both available and useful for MRJ, we ought to infer that it was indeed used in that application from the outset.

CONTENTS

1 Introduction 2
2 The State of Play 3
2.1 Meat and Meta-Anger 3
2.2 Detection and Demand 5
2.3 Fischborn's Error: Omitting Assessment 8
3 Intention-Sensitivity: Available and Useful 9
3.1 Availability 11
3.2 Usefulness 14
3.3 Freedom from Free-Riding 18
4 The Cross-Cultural Objection 20
5 Conclusion 22

* *Department of Logic and Philosophy of Science, University of California–Irvine*

1 INTRODUCTION

Moral responsibility judgment is a central psychological process connecting moral beliefs to everyday moral practices. Moral responsibility judgment (abbreviated MRJ) refers to the assignment of responsibility in response to some morally-charged event one observes. For example, if I see an innocent man executed in the street, I will automatically and intuitively evaluate who (or what) is morally responsible. I may blame the gunman who pulled the trigger, or the fascist commander who ordered the gunman to fire, or I might blame no person in particular, instead blaming some physical condition such as an unexpected gust of wind that knocked the gunman off-balance, causing him to fire accidentally. How my judgment turns out will depend on the details of what I observed, but the psychological process of adjudicating moral responsibility will occur spontaneously when prompted by the detection of a morally-charged event such as this.

MRJ is a ubiquitous part of modern human psychology: all humans spontaneously engage in laying blame, even if they sometimes disagree on where blame should be laid. Some theorists have advanced evolutionary accounts aiming to explain the emergence of MRJ, as well as how such evolutionary details can inform our understanding of modern-day MRJ practices as they are instantiated in our interpersonal relationships and legal institutions. Here I contribute to this growing literature by resolving a crucial deficit of existing accounts: I offer an evolutionary explanation concerning how and why *intentionality* came to figure so centrally in humans' moral responsibility judgments.

I begin in Section 2 by appraising two recent evolutionary accounts of MRJ in order to establish the explanatory gap my contribution fills. I then argue, in Section 3, for the role of intention-sensitivity in the evolution of MRJ. My central hypothesis is that MRJ was sensitive to the intentional character of moral transgressions—whether a violation was done intentionally or not—from the earliest emergence of MRJ as a psychological process. I argue that intention-sensitivity is a capacity that pre-dates moral cognition, and thus would have been available for use in the earliest-emerging form of MRJ. Further, intention-sensitivity would have improved the ability for MRJ to carry out its main fitness-enhancing functions, especially in its role as an adaptive partner choice mechanism. Since intention-sensitivity would have been available upon the emergence of MRJ and useful for the adaptive functions of MRJ, we have reason to suppose that intention-sensitivity *was* in fact a component of MRJ from the outset. This view runs counter to competing evolutionary accounts of MRJ, which either do not discuss the role of intention-sensitivity in the emergence

of MRJ or explicitly posit that early forms of MRJ did not involve considering the intentionality of a transgressor's action. I argue for my hypothesis using a combination of evolutionary theory and evidence from comparative, developmental, and cross-cultural psychology. In Section 4, I contend with a common objection to my view based on the diversity of MRJ practices across cultures. In Section 5, I conclude that hominin evolutionary history offers us an explanation as to why intentionality is so inextricably linked to humans' beliefs about the nature of desert, blame, and moral responsibility.

2 THE STATE OF PLAY

Much scholarly work over the last two decades has sought to illuminate the evolutionary origins of morality itself—the tendency to form moral beliefs, as opposed to normative (but conventional) beliefs—in a pre-existing population without morality (Joyce, 2006; Kitcher, 2011; Tomasello, 2016). By comparison, there is surprisingly little work specifically focused on the evolution of moral responsibility judgment—assessments of who is to blame when a moral norm has been violated. There are two notable exceptions amid this general paucity of work on the evolution of MRJ: the account of Matteo Mameli (2013) and that of Marcelo Fischborn (2023). I discuss each account in turn before identifying the limitation I aim to remedy. These accounts are limited by the fact that they do not recognize the importance of intention-sensitivity in the earliest-emerging forms of MRJ. This results from their focus on the adaptive function of MRJ as a mechanism for *partner control*—as a means of motivating and regulating the delivery of punishment in a fitness-enhancing way. Partner control is indeed an important function that partially explains the emergence of MRJ, but if not more important is the adaptive function of MRJ as a mechanism for *partner choice*. As I will argue in the next section, when we recognize the adaptive importance of partner choice in the emergence story of MRJ, we see why the earliest form of MRJ likely involved considering intent.

2.1 Meat and Meta-Anger

Mameli's account, which goes by the tagline, "Meat made us moral," asserts that the rise of cooperative large-game hunting occasioned the emergence of blame and desert in the hominin lineage. He rightly recognizes that simply holding moral beliefs, e.g., that certain actions are *wrong*, instead of merely prohibited, is not on its own suf-

ficient to generate fitness-enhancing patterns of behavior. There must be some mechanism which connects such beliefs to actions—a psychological drivetrain that connects the moral engine to the behavioral wheels. As part of his account, then, Mameli proposes a model of how moral beliefs became linked up with the relevant fitness-enhancing behaviors through the advent of MRJ.

Mameli claims that to be adaptive, moral cognition needed not only to generate beliefs about the truth of moral propositions, but also produce psychological dispositions to:

> (D1) feel anger toward someone who violates a moral norm, and,
>
> (D2) feel "meta-anger" at others who do not *themselves* experience anger toward the norm violator (905).

Meta-anger is what Mameli thinks enabled punitive responses toward norm-violators to be seen as *merited*, and thus to be protected from the costs usually associated with delivering third-party punishment. Without meta-anger (D2), the first-order anger one feels toward a norm-violator (D1) would generate a punitive response at significant risk of retaliation and reputational damage to the punisher. But if group members all experience meta-anger, then the first-order anger one feels toward the norm-violator is not only permissible in the minds of other group members, but positively *demanded* by group members' dispositions toward meta-anger. Indeed, in a group where all members feel meta-anger, delivering punishments becomes a public signal that the punisher is feeling appropriate anger toward the norm-violator.[1] Mameli sees the adaptive function of meta-anger as its ability to make third-party punishment fitness-enhancing by preventing second-order free-riding—i.e., playing the bystander and letting others deliver punishment, even when you too would like to see the violator punished.

The historical hypothesis Mameli proposes (viz. that "meat made us moral") is plausible, given our repository of archaeological evidence and our trace-based knowledge of early hominins' cooperative practices. But the model Mameli suggests for the psychological process of MRJ itself runs afoul of the empirical evidence. As Fischborn notes, on Mameli's view "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

1 Mameli also includes self-directed dispositions to: (D3) feel anger toward oneself when one has violated a moral norm, and (D4) feel anger toward oneself when one does not feel anger at the moral violation of someone else. For brevity, these dispositions are left out of the foregoing discussion.

or otherwise) in support of that claim" (2023, 823). In particular, Fischborn cites contradictory ethnographic evidence from Christopher Boehm (2012) on this front. Boehm provides evidence from diverse human societies showing that most "Late Pleistocene Appropriate" societies (i.e., extant societies which Boehm thinks best approximate what human societies might have been like in the relevant evolutionary context; see Boehm, 2012, 79) do not display anything like "meta-anger" toward those who fail to express first-order anger toward deviants.

I would add to Fischborn's criticism of Mameli's view that it is unclear what evolutionary dynamic could have enabled (D2) to spread through a population which did not already possess it. Since the stabilizing benefits of meta-anger—the meriting of punishments toward norm-violators—only arise once (D2) has already spread widely throughout the group, it is difficult to see how meta-anger would have produced the marginal fitness advantages required for such a novel trait to spread. Mameli's account helpfully identifies a context in which MRJ may have arisen—cooperative large-game hunting—but it does not how, and in what form, MRJ emerged in a population that previously had no concept of moral responsibility.

2.2 Detection and Demand

Recently, Marcelo Fischborn put forth his own view of MRJ evolution, aiming to improve on Mameli's account with respect to its agreement with well-supported models of MRJ psychology and the plausibility of its evolutionary dynamic.

On the basis of Morris Hoffman and Frank Krueger's (2017) neuropsychological account of blame and punishment, Fischborn considers the following four psychological capacities to be indispensable to MRJ in modern humans. I have added handy labels to these capacities for ease of discussion.

DETECTION	The capacity to detect instances of norm violations.
DEMAND	The motivation to respond punitively to someone who violates a norm.
ASSESSMENT	The assessment of the violator's mental processes relevant to the punitive response.
PERMISSION	The capacity to assess contextual factors relevant to whether/how one should respond punitively.

It is the task of an evolutionary account of MRJ to explain how these capacities came to be constitutive of MRJ psychology in modern humans. Fischborn argues, following Michael Tomasello's (2016) two-step evolutionary account of morality, that these MRJ-constituting capacities evolved in two main steps. The emergence of the DETECTION and DEMAND capacities occurred during the "first step," approximately 400,000 years ago (~400 kya). The "second step" occurred as hominins began living in larger and more socially-cohesive groups around 200 kya, at which point the PERMISSION capacity arose. Developing the conjunction of the DETECTION and DEMAND capacities constituted the emergence of moral responsibility judgment; later on, the addition of the PERMISSION capacity stabilized the adaptive benefits offered by MRJ by protecting third-party punishers from deleterious costs.

Fischborn sees two types of situations in which this early DETECTION–DEMAND form of MRJ would have produced fitness enhancements during the "first step": one member of a cooperative dyad admonishing the other in response to a norm violation, and *both* members of a cooperative dyad coordinating to punish someone attempting to steal their jointly-won resources.

To make this first type of context more concrete, imagine two ancestral hominins cooperating with the joint goal of collecting honey from a beehive high up in a tree. One partner (call them Ha) climbs the tree carrying a hand axe to cut down the beehive. The other partner (call them Lok) stands on the ground below with an outstretched animal hide, ready to catch the falling beehive and seal it quickly with the hide to suffocate the bees.[2] Suppose Ha successfully cuts the hive from the tree, but Lok gets frightened as the hive is falling and dives away, letting the hive splatter on the ground, thus ruining the honey and thwarting the cooperative endeavor for both individuals.

From our modern psychological perspective, it is intuitive to imagine that Ha would not just feel generally upset about losing the honey, but would feel upset at Lok specifically and be motivated to punish Lok through admonishment. Probably, most of us even feel that Ha would be *justified* in their anger at Lok. But the very psychological profile we now possess is the target of evolutionary explanation here, so we must ask how ancestral hominins would have come to form judgments such that they respond to this event with anger directed at Lok, leading to punishment, *without* appealing to the obviousness of such judgments from our modern psychological profile (Farrell, 2025). Fischborn's proposal is that the capacities required to form such judgments were DETECTION and DEMAND.

2 Characters borrowed from William Golding's prehistoric science fiction novel, *The Inheritors* (1955).

In this dyadic example, there is no need for Ha's anger at Lok to be seen by third parties as being justified or merited. If Ha delivers a punishment upon Lok, Ha incurs no danger of retribution by others in their social group, because there are none. Thus, on Fischborn's account, Ha's ability to detect Lok's violation of their role-specific norm (catching the falling beehive) and the demand Ha feels to punish this norm violation are sufficient to produce the retributive punishment that follow from the MRJ. This will also apply to situations where the violated norm is less instrumental and more squarely "moral," such as if Lok succeeded in catching the hive but then tried to keep all the honey for himself instead of sharing it with Ha. The fitness enhancements come from the effect that Ha's punishment has to disincentivize Lok from violating norms in future cooperative interactions, leading to higher expected payouts for Ha in the long run.

The second type of scenario in which Fischborn thinks MRJ would have functioned adaptively during "step one" (~400 kya) is in the context of punishing food thieves after a successful dyadic hunt. We may imagine here an uninvolved third person attempting to take the spoils of a dyadic hunt for themselves, despite not having contributed whatsoever to their procurement. The thought is that punitive responses toward would-be food thieves would have deterred food theft, and thus benefited the members of the dyadic hunting party by minimizing the loss of their hard-earned spoils.

Fischborn is quick to acknowledge one way in which this second scenario is open to a kind of free-riding: if one is part of a successful dyadic hunt facing the threat of food theft by an uninvolved third person, it benefits them if their partner, rather then they themselves, deters the would-be food thief via punishment. It is thus adaptive to *not* deliver punishment in response to attempted food theft in this type of situation, on the assumption that one's dyadic partner is willing to punish the food thief. This, of course, is just an instance of the classic second-order free-riding problem.

Fischborn sees two mitigating factors here. First, it is plausible that a coordinated punishment delivered by both members of the dyad would more effectively deter food theft than the individualistic punishment described above. If so, then the second-order free-riding issue may not arise. Secondly, Fischborn hypothesizes that because dyadic hunting was facilitated by the capacity for joint intentionality—the cognitive capacity to represent the dyad as a single agent—the delivery of coordinated punishments could be faciliated in the same way: "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' "(828).

Fischborn claims that the early DETECTION–DEMAND form of MRJ later evolved, around 200 kya, into an MRJ psychology more closely resembling that of modern humans. This occurred through the advent of the PERMISSION: the ability to evaluate the degree of social support for the delivery of punishment, with special attention to determining whether one is *permitted* to deliver the punishment. The adaptive value added by the PERMISSION capacity, says Fischborn, is to stabilize the delivery of punishments in group settings. Once hominins began living in larger groups, punishments were likely to be witnessed by numerous onlookers, opening the punisher to potential retribution from those who did not think the original punitive act was warranted and thus interpreted it as a norm violation in and of itself. The ability to "read the room," so to speak, before going through with punishments one felt motivated to realize minimized these potential costs of delivering punishment in group settings.

Fischborn's view is that "moral responsibility judgment in the demand sense alone was not likely adaptive and stable," but "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" (Fischborn 828). Thus Fischborn holds that, although MRJ did emerge upon the appearance of the DETECTION and DEMAND capacities ~400 kya, it was the PERMISSION capacity ~200 kya that allowed full-fledged MRJ to get off the evolutionary ground and find a stable adaptive trajectory.

2.3 Fischborn's Error: Omitting Assessment

Fischborn presents an adaptive hypothesis that enjoys greater plausibility than does Mameli's, particularly in its evolutionary dynamic, and especially if one is friendly to Tomasello's two-step account of the evolution of moral cognition. He also does well to square the capacities he posits to be constitutive of MRJ with the psychological literature on MRJ in modern humans, which represents a marked improvement over Mameli's empirically inadequate model of MRJ psychology. But notice there was no mention of the ASSESSMENT capacity in my description of Fischborn's account above; this is because ASSESSMENT plays no role in Fischborn's account of the emergence of MRJ. Fischborn asserts that sensitivity to intent was either absent from or unimportant to the ancestral forms of MRJ psychology he describes.

Fischborn acknowledges that "some understanding of the mental states of the external violator" may have been present in the DETECTION–DEMAND form of MRJ during "step one," but says such con-

siderations were "not, at this point, a factor that could make someone refrain from realizing a responsibility episode" (Fischborn, 827). In other words, upon the emergence of MRJ and for a long time thereafter, considering a transgressor's intent played no constitutive or regulative part in guiding judgments about whom to punish, when, and under what conditions. The intentionality of their action did not factor into the process by which they were judged blameworthy and/or punishable at this stage.

In what follows, I argue it is a mistake to treat intention-sensitivity as an extraneous feature in the evolutionary emergence of MRJ. Intention-sensitivity was in fact an essential feature of MRJ psychology from the get-go, and was moreover indispensable to the evolutionary dynamic that enabled MRJ to emerge. Fischborn acknowledges that "the assessment of mental states (e.g., intentions) ... is constitutive of modern moral responsibility judgment" (827), and that his account offers no explanation for how that came to be the case. This is meant as an identification of future avenues for research, a sort of loose end that should be tied up by later work. But the role of intention-sensitivity in the evolution of moral responsibility judgment is not just a loose thread to be woven in later; it forms the very seam which holds the whole evolutionary picture together. Fischborn's account errs in ignoring the adaptive importance of intention-sensitivity in the evolution of MRJ, but the larger point is that *any* evolutionary account of moral responsibility judgment ought to recognize and explain the centrality of intention-sensitivity to the evolutionary dynamic by which MRJ emerged. Our understanding of the process by which we came to be blame-layers in just the ways we are, something which is only recently starting to take shape, will be greatly improved for it.

3 INTENTION-SENSITIVITY: AVAILABLE AND USEFUL

Before diving into the argument for my hypothesis, we require a brief prelude to establish the following methodological point. To support the hypothesis that a particular capacity (C) was used by an ancestral population (A) for a particular fitness-relevant purpose (P) in the evolutionary history of a particular lineage, it suffices to show that C was *available* to A and that C would have been *useful* to employ for P.

Ron Planer and Kim Sterelny employ this methodology in their recent account of the evolution of language (2021). Their inferential strategy is to use empirical evidence "to identify the availability of

cognitive and social resources for particular communicative capacities, and to use that same record to identify communicative needs that select for those capacities. *If capacities were both available and useful, they probably had them*" (48, emphasis mine). When a population faces recurrent challenges of reproductive significance due to their existing lifeways—such as the communicative demands of plastic prosocial cooperation Planer and Sterelny describe, or in our case, the demand to safely disincentivize counternormative behavior similarly brought on by plastic prsocial cooperation—and there exist capacities that 1) would have been useful in overcoming those challenges, and 2) we have good reason to think were available for such uses given the cognitive, social, cultural and/or ecological resources accessible to them, then we are licensed to think that the population likely *did* employ those capacities to contend with those reproductively-significant challenges.

Note here that saying C would have been useful for P must take into account the costs associated with the use of C. If C particularly costly to deploy, such as the venom injection of snakes or the inking behavior of some cephalopods, then there may be many cases in which using C would have been successful at accomplishing some P, but not worth the cost of deployment, and thus not *useful* for P, on this definition. Costliness is particularly difficult to judge in the case of cognitive capacities, as we have no guarantee that the ideas and inferences modern humans find easy, intuitive, and automatic would have been similarly low-cost for the ancestral hominins in question (Farrell, 2025).

There is further danger in this inferential approach due to its reliance on theorists' assessments concerning what capacities were or were not available to a particular population at a particular time, given the evidence at hand, and how those capacities might or might not have been useful in particular applications. But note that I am only claiming these are sufficient conditions for supporting a particular hypothesis about what capacities were used for which purposes in a lineage's evolutionary history; that support, of course, need not establish certainty or even a great deal of confidence in the historical claim. If we know C was available to A and C would have been useful for P, then we have *some* support for thinking that C was used for P, but that support will come in various degrees and may in many cases be exceedingly weak. Nevertheless, with this methodological commitment stated, let us build as strong a case as possible for my claim that intention-sensitivity was available to and useful for moral responsibility judgment, and thus that intention-sensitivity was likely a feature of MRJ from the outset.

3.1 Availability

The relevant starting point in hominin evolutionary history here is the emergence of the cognitive capacity I claim to have been available to and useful for the earliest form of MRJ. The cognitive capacity under discussion here is the ability to detect intentional actions and discriminate them from non-intentional ones. I take this to at least approximate what Fischborn's notion of ASSESSMENT—the "capacity to assess the mental involvement of the target with the violation in a way that can affect the motivation to respond" (2023, 819)—though intention-sensitivity as I imagine it is a thinner notion than is "mental state assessment" proper, since intention-sensitivity does not strictly require mental state ascription. The capacities relevant to our evolutionary discussion here are actually two: 1) discriminating agents from non-agents, and 2) sensitivity to whether an agent's actions are intentional or not. I argue in the following subsections that both capacities pre-date MRJ, likely by several million years.

3.1.1 Agent-Discrimination

Chimpanzees, humans' closest extant relatives, discriminate between agents and non-agents in a similar though less sophisticated way as humans do. Though more controversial, there is also significant evidence that chimps possess the ability to reason about others' desires, beliefs—even their false beliefs in some contexts (Krupenye et al., 2016)—and intentions (Call and Tomasello, 2008; Krupenye, 2021; Royka and Santos, 2022). The existance of these homologous cognitive traits in chimpanzees make it likely that the tendency to cognitively categorize things in one's environment as either agents or non-agents emerged prior to the split between the Pan and Homo lineages approximately six million years ago (~6 mya). Since chimpanzees do not further possess moral cognition or an MRJ psychology, as humans do, we can conclude that MRJ emerged sometime after that split.

Let us then grant that by 6 mya, the last common ancestor of humans and chimpanzees possessed a basic capacity for agent-discrimination. In that context, and for most of its history, agent-discrimination served the very general adaptive function of enhancing individuals' ability to predict and interpret other agents' behaviors. Representing certain entities—cooperators, prey, predators, and other suitably agential things—as having choices of possible actions among which they can intentionally select enhanced fitness because it enhanced prediction, and so improved the chances of success (survival, acquisition of resources, etc.) in interactions with those entities (see Dennett, 1987).

In the evolutionary context, "successful" predictions were those which helped ancestral hominins correctly anticipate the behaviors of other humans (or weather systems, or predators, or prey) in situations with fitness consequences. These hominins would not have been directly motivated by reproductive success in their discrimination of agents from non-agents, of course. Their thought processes about, e.g., whether that thing on the ground is a stick or a snake, were presumably prompted by more tangible goals like avoiding harm, acquiring resources, and protecting kin. But these proximate motivations would have frequently correlated with increased fitness, and this would have led to robust and profligate use of agent-discrimination by hominins in pre-moral contexts where the right cues and predictive demands presented themselves.

3.1.2 Intention-Sensitivity

The capacity for agent-discrimination allows users to form different expectations about the behaviors of different types of entities. This will have been quite useful in many cases, for example, determining what will transpire when an entity tumbles off of a platform it was previously sat upon. If the entity is represented as agent (e.g., a bird perched in its nest), then the user will expect it to return itself atop the platform; if not represented as agent (e.g., an egg sitting in the aforementioned nest), then there shall be no expectation that it will reestablish its previous position.

But this minimal form of reasoning and representation about agents, so far described, does not guarantee that users can or will differentiate between intentional and unintentional behaviors of entities *already represented as agents*. Such differentiation is intuitive and automatic to modern humans, but it is not unreasonable to ask whether our hominin ancestors might have initially been blind to whether putative agents' actions were intentional or not. It could be that birds were represented as agents, but that this did not enable individuals to detect or represent the difference between a bird that falls from its nest and a bird that flies down to the forest floor. However, there is good reason to think that the capacity for intention-sensitivity—the capacity to discriminate between the intentional and unintentional behaviors of putative agents—was at least present by the Pan-Homo split approximately 6 mya.

This phylogenetic claim is supported by comparative evidence: chimpanzees can and do discriminate between the intentional and accidental behaviors of agents. Call and Tomasello (1998) studied chimpanzees trained to use a particular mark to tell which of three boxes contained a reward. When they then observed an experimenter

mark two of the three boxes—one marked accidentally and the other marked intentionally—chimps preferentially selected the intentionally-marked box, suggesting they understood that the accidentally-marked box was less likely to contain a reward, since the mark was not meant to be placed there. In a later study, Call et al. (2004) found that chimpanzees were much less likely to get upset at an experimenter who presented them a grape but accidentally dropped it than an experimenter who presented them a grape but then pulled the grape away from them before they could take it.[3] Finally, in a social learning context, Tomasello and Carpenter (2005) tested chimpanzees who had observed an experimenter complete a novel task to get a reward. They saw the experimenter make two actions, one accidental and one intentional, both of which were unfamiliar to the chimp. When given the chance to try the task themselves, the chimpanzees only imitated the experimenter's intentional action, and not their accidental one.

Being able to distinguish between agents' behaviors which are intentional and those which are not is of immense predictive value. Since agents are represented as being able to exert intentional causal control over their actions, detecting which of an agents' actions were done intentionally is understood to reflect their projectable behavioral tendencies. Detecting whether you jumped in to the lake or whether you fell into the lake is crucial to predicting how likely you are to enter the water in the future. Intentional actions are far more informative for predictive purposes than unintentional or unavoidable actions, a fact which even infants understand (Eason et al., 2018). The comparative evidence reviewed above makes it seem quite likely that chimpanzees share with humans a homologous capacity not only to represent certain entities as agents, but also to distinguish between the intentional and unintentional behaviors of agents.

Most agree, or at least find plausible, that agent-discrimination emerged as way to enhance predictions about the behaviors of suitably complex entities that individuals might encounter. Beyond being able to discriminate agents from non-agents, it is likely that by the Pan-Homo split approximately 6 mya, individuals could also discriminate between the intentional and unintentional actions of putative agents, further enhancing the predictive power of this representational system. In the pre-moral context of their emergence, these capacities enabled successful interactions (both cooperative and competitive) with conspecifics, predators, prey, and other complex entities. Knowing whether the hyena turned its head my way because it saw me or just due to happenstance might mean the difference be-

3 As evidence for the particular claim I'm arguing here, it would have been better if the experimenter in the intentional condition had intentionally *dropped* the grape, instead of pulling it away, in order to match the exact action taken in the accidental condition. Unfortunately no such condition was tested.

tween life and death. Thus, both agent-discrimination and intention-sensitivity were favored by selection and, with time, spread to fixation in some population ancestral to modern humans.

With the pre-existence of intention-sensitivity to MRJ established, we may now leap forward in time to the emergence of MRJ.[4] Intention-sensitivity would have been *available* to MRJ upon the latter's emergence; how, though, would it have been useful? If, as I will now argue, being sensitive to the intentionality of a transgression would have appreciably improved the ability of MRJ to fulfill its adaptive functions, it would be puzzling if early forms of MRJ did not do so.

3.2 Usefulness

Mameli and Fischborn both see facilitating safe and effective punishment of moral transgressions as the adaptive function which explains the emergence of MRJ. Both authors think MRJ emerged through natural selection because it helped maximize the payoffs of cooperation by modifying cooperative partners' behaviors for the better and/or preventing the loss of hard won resources by free riders such as food thieves while also minimizing the risks associated with delivering punishment, especially in large group settings. MRJ is thus, according to these authors, primarily a mechanism of adaptive partner control. But MRJ is also a mechanism for adaptive partner *choice*, and it is in this role that intention-sensitivity improves the fitness-enhancing power of MRJ.

It is of utmost adaptive importance for cooperators to be able to judiciously choose which members of their social group they are willing to engage with in high-stakes, high-risk cooperative endeavors (Sperber and Baumard, 2012). Considering others' moral reputations and using them to curate one's social network is a core adaptive function of moral cognition, since it enables prosocial cooperators to avoid serial exploitation by free-riders and antisocial individuals across diverse cooperative contexts (Stanford, 2018). Forming moral responsibility judgments about others' behaviors works as a sort of running register of the facts relevant to assessing them as potential cooperative partners. Thus, beyond partner control, MRJ enhances fitness by improving judges' evaluations regarding who the desirable cooperative partners are, and more importantly, who the undesirable ones are. But this is only to identify an additional adaptive function of way MRJ which is underappreciated within existing evolutionary ac-

4 While it is clear MRJ emerged long after the Pan-Homo split, I have no strong view on when, exactly, MRJ emerged. The 400 kya date claimed by Fischborn (following Tomasello) is plausible, but my view does not turn on whether that particular date is accurate.

counts. I have not yet said exactly how *intention-sensitivity* improves MRJ in its role as a mechanism of adaptive partner choice, and thus, why exactly we should think that the earliest form of MRJ likely did involve sensitivity to the intentionality of a transgressor's actions.

Given the importance of high-risk cooperative endeavors in our species' evolutionary history, many of the systems in human social cognition seem to be tailored to some degree toward assessing the quality of potential partners (Baumard et al., 2013). For example, the emergence of language was likely driven in part by the need for gossip brought on by increasing group size, which greatly enhanced reputational assessments about others with whom one has not directly interacted (Dunbar, 2004). Some have even argued that reasoning itself—or at least, the dialogical, rhetorical form of reasoning sometimes called 'reason giving'—is adapted primarily for reputational protection through public justifications of one's own actions by appeal to reasons, and for reputational assessment of others through judgments about reasons they offer to justify their actions (Mercier and Sperber, 2017). Let us consider whether MRJ fits this mold of a cognitive mechanism tailored to the assessment of others in the service of evaluating them as potential cooperative partners.

There are many routes to deciding whether it is a good idea to cooperate with a particular member of one's social group. One may use superficial cultural markers or physical features: perhaps wearing a particular style of dress signals that they are a member of some reputable family, or perhaps their physical stature and impressive strength signal that they will be useful in hunting down large game. Different partner choice strategies will be apt for different partner choice contexts. However, one very general task germane to nearly all partner choice contexts is determining whether they will follow moral norms—whether they are the kind of good-natured, rule-abiding individual with whom it is a good idea to get involved. Will this individual share the spoils of our hunt when we have the prize in hand? Will they have my back while we are out in the wilderness and something goes wrong? When it comes to partner choice, no factor looms larger than predicting a potential partners' likely behavior during a high-stakes cooperative endeavor, and no information is more valuable to such predictions than information about the potential partner's history of intentionally-taken, morally-relevant actions.

The MRJ psychology of modern humans follows this same logic. Psychologist Fiery Cushman has shown that while peoples' judgments about whether someone deserves punishment for causing a negatively valenced moral outcome does not much depend on whether their causal contribution was intentional or not, the exact opposite is true when it comes to judgments about their moral character and social de-

sirability (Cushman, 2008, 2015). In such moral character assessments, intentionality take precedence. When an agent causes harm unintentionally, people feel that the agent is not necessarily a bad person at their core, but nevertheless do deserve some kind of punishment. But if an agent attempts to cause harm, even if no harm occurs, people consider the agent to have betrayed their poor moral character, despite also asserting that the agent does not deserve punishment (because they did not in fact cause any harm). We should not overstate this, though: there is no perfect dissociation here. Accidentally causing harm can lead to judgments of poor moral character, especially when the accidental harm results from recklessness or negligence. Similarly, attempting (but failing) to cause harm can lead to judgments that punishment is deserved, though the degree of punishment deserved is invariably weaker than if the attempt to harm had been successful. Still, these results suggest that in modern human psychology, moral character and social desirability judgments are evaluated primarily on the basis of intentionality, rather than actual actions and outcomes.

These findings have been subsequently validated by independent researchers on a larger scale (Kneer and Machery, 2019; Kneer and Skoczeń, 2023). However, both Cushman's studies and these large-scale replications were done using adult Westerners, so a natural rejoinder here is to question whether these results show anything beyond a cultural artifact specific to the moral systems typical of WEIRD (Western, Educated, Industrialized, Rich and Democratic; see Henrich et al., 2010) societies.[5] But the psychological salience of intent over outcome when judging social desirability and moral character appears a strikingly young age. J. Kiley Hamlin (2013) conducted a series of experiments on infants showing that, within the first year of life, children robustly privilege intent over outcome in their social desirability evaluations. When given the choice of a puppet they saw accidentally harm another puppet, versus a puppet they saw try (but fail) to harm another puppet, infants preferred the puppet that accidentally caused harm. Infants thus judged that merely intending to cause harm, regardless of the outcome, makes you undesirable; actually causing harm, but doing so unintentionally, does not.[6] These results are corroborated by other studies showing that very young children are keenly sensitive to the intentional character of others' morally-valenced behaviors (Chernyak and Sobel, 2016; Li and Tomasello, 2018; Vaish et al., 2010; Woo et al., 2017), even in moral domains other than harm-avoidance, such as fairness norms about resource distribution (Geraci et al., 2022; Strid and Meristo, 2020).

5 I take up this challenge in greater detail in the following section.

6 A similar pattern of results also held for helping behaviors. Attempting (but failing) to help someone makes you desirable, but accidentally helping does not.

All this suggests that the intent-based evaluation of moral character and social desirability employed by WEIRD adults is present even in preverbal, largely unenculturated infants. It has long been thought, going back to Piaget (1965), that children begin life with an outcome-oriented approach to moral judgment and only later acquire an intent-based morality. But in recent work, Francesco Margoni and Luca Surian (2016) have argued that this is a spurious artifact of the methods traditionally used to investigate sociomoral evaluation in toddlers, which present too high of cognitive demands on executive functioning for children in this age range. Subsequently, Margoni & Surian (2020) provided direct evidence that children between 2-4-y.o. readily and preferentially make sociomoral evaluations on the basis of intent when evaluation tasks were modified to reduce executive functioning demands. These results confirm earlier studies suggesting that children between 3-8-y.o. use intentionality as a primary guide in social evaluation (Nobes et al., 2009). Thus, sociomoral evaluation seems to display developmental continuity: the intent-based sociomoral assessment humans display in infancy persists, at its core conceptually unchanged, all the way through to full-fledged MRJ in adulthood.

Returning now to evolutionary considerations, we see that judgments regarding the social desirability of an agent need not be especially dependent on whether they actually causally contributed to a negatively-valenced moral outcome or not. What matters for adaptive partner choice purposes is just whether the agent intended to cause such an outcome. Attending to the *intended* outcome of an agent's actions is a better compass for assessing their likely future behavior, and is thus more useful for partner choice, than is attending to the *actual* outcome of their actions. That you tried to kill your last cooperative partner is all I need to know; further information about whether your attempt at murder succeeded or failed is superfluous to my evaluation of your desirability as a partner, though it may helpfully inform my assessment of your proficiency for murder.

There are two types of situations where an intention-sensitive MRJ creates clear fitness advantages over a hypothetical form of MRJ which is not sensitive to intentionality. The first involves missing out on good cooperative partners and the second involves getting stuck with bad cooperative partners.

For an example of the first type of situation, imagine that during a hunt Ha witnesses Lok throw a spear at the prey which inadvertently wounds Mal, who was wrangling with the animal at close quarters. Lok clearly caused harm to Mal by accident, which makes him far less likely to cause similar harms in the future than if he had caused the harm to Mal intentionally. If Ha sensitive to this fact, he

will recognize that Lok's actions do not reflect deep about his moral character. If Ha is not sensitive to this fact, he would see 'causing harm to cooperative partners' as a projectable behavioral disposition of Lok, and thus will likely avoid cooperative interactions with Lok going forward despite the fact that Lok is by all means a good cooperative partner and was unlucky this time.

For an example of the second type of situation, imagine that Mal scavenged a large carcass and brings it back to camp, where he intends to selfishly hoard the meat all to himself. However, Mal quickly gives up on trying to hoard the meat when it becomes clear that he can't defend the resource successfully, and each tribe member ends up taking a share of meat. Ha, witnessing this scene with an intention-sensitive form of MRJ, will realize that Mal sharing their resources was not intentional, and so should not reflect a strong likelihood that Mal will abide by fairness norms in the future. In fact, depending on sophistication of Ha's mindreading capacity (i.e., if Ha can tell not just that the sharing was unintentional but that Mal's goal was specifically to *not* share), Ha may take Mal's behavior to reflect a projectable antisocial tendency, despite the actual fair outcome that resulted from this event. By comparison, if Ha's MRJ psychology does not involve consideration of intent, the fair outcome resulting from Mal bringing a carcass back to camp will suggest that Mal abide by the relevant fairness norms; thus, Ha would erroneously conclude that Mal is a good, fair cooperative partner to interact with for future hunts.

Intention-sensitivity improves MRJ as a mechanism for adaptive partner choice by the same very route through which intention-sensitivity conferred fitness benefits in the pre-moral context: by improving users' predictions of the future behaviors of other agents. In the moral context, these predictions informed partner choice, while in pre-moral contexts (and concurrently amoral contexts, which of course did not stop after the emergence of moral cognition), they informed the strategies one should use to escape the hyena or corner the hare.

3.3 Freedom from Free-Riding

My focus on the importance of intention-sensitivity in the evolution of MRJ foregrounds the adaptive significance of partner choice. This contrasts with the other views we have discussed, from Mameli and Fischborn, which see the primary adaptive role of MRJ as being a mechanism for facilitating partner control, i.e., punishment. While Fischborn, especially, does recognize partner choice as one function

of MRJ, it still plays far too small a role in his overall evolutionary account.[7]

There is good reason why these authors are so attuned to the punishment-facilitating role of MRJ. The main challenge in any evolutionary account of morality is to explain how third-party punishment could be fitness-enhancing, when it so naturally exposes the punisher to fitness-reducing risks. In our survey of Fischborn's view (2.2), we considered how MRJ-guided motivations to deliver punishments might generate opportunities for second-order free-riding in two types of contexts: admonishing a norm-flauting hunting partner, and deterring a would-be food-thief. The risk of free-riding in these scenarios arose from the fact that if one decides to deliver a punishment, they do so at the risk of letting others enjoy the increase in expected payoff that results from the punishment without having to pay that same cost.

Because I emphasize the adaptive function of MRJ as a mechanism for partner choice, I do not suppose deterring food thieves or correcting greedy partners to have been the essential use cases which explain the emergence of MRJ. Though such situations likely did contribute to the evolutionary dynamic leading to the emergence of MRJ, I argue that these partner-control-centric contexts are of less relevance to a complete explanation of the evolutionary emergence of MRJ than are situations in which MRJ functioned as a mechanism for partner choice. Insofar as experiencing a motivation to punish the food thief is fitness enhancing, it enhances fitness primarily because it offers the punisher an opportunity to demonstrate their value as a good cooperative partner, and secondarily because it will modify the future behaviors of the thief and others similarly tempted to experiment with thievery. Viewing MRJ as primarily an adaptation suited for partner choice avoids the recalcitrant free-riding issues that are brought on by supposing, as existing accounts do, that MRJ is an adaptation suited primarily or exclusively for partner control.

There is one other way my proposal helps to fill in our patchwork picture of how and why MRJ emerged in our hominin lineage's cognitive evolution. Fischborn considers the DETECTION and DEMAND capacities to have been jointly sufficient for the emergence of MRJ, without any immediate need for the ASSESSMENT or PERMISSION capacities until later, when hominins began living in larger groups. At that later time the PERMISSION capacity emerged, Fischborn hypothesizes, which allowed punishments to be delivered in a safe and co-

7 One might reciprocally argue that my account gives too small of a role for the partner control function of MRJ. However, I am not advancing a full-fledged account of the evolution of MRJ here; I am only seeking to remedy a deficit present in the full-fledged accounts on offer.

ordinated manner in large group settings by enabling punishers to check whether the punishment they mean to deliver will be seen as "permitted" by other group members. I suggest that, in fact, the intention-sensitive nature of MRJ filled this warranting role that Fischborn sees the ASSESSMENT capacity as having played. The punishments that were permitted, and thus safe to deliver, were those licensed by widely-accepted facts about the relevance of intentionality to moral responsibility judgments. Everyone knows causing harm is all the worse if done intentionally, and they use this knowledge to their individual benefit via-a-vis partner choice. That this knowledge is widely possessed means that if one observes another stealing from a third person in a clearly intentional manner, they are permitted to punish them—and further, they can be sure that others will see the punishment as merited too. We need not hypothesize a separate psychological capacity (PERMISSION) to detect whether others agree that punishment is warranted; seeing that a transgressive harm was done intentionally is frequently all that is needed to infer that the ensuing punishment will be recognized as warranted.

4 THE CROSS-CULTURAL OBJECTION

We now have an idea of how intention-sensitivity, a capacity which pre-dates moral responsibility judgment, could have improved the adaptive function of MRJ in its role as a mechanism of partner choice. Thus, intention-sensitivity was both available to and useful for MRJ upon the latter's emergence. Following the methodological commitment stated at the start of the previous section, we have support for the claim that intention-sensitivity probably *was* used in the earliest form of MRJ. There is however a potentially devastating objection to my view that has not yet been fully addressed. Recent cross-cultural work has demonstrated the diversity of intentionality's salience in MRJ across cultures. In some societies, adults' moral responsibility judgments rely primarily and sometimes entirely on outcome alone, with no import whatsoever placed on intentionality (Barrett et al., 2016; McNamara et al., 2019). In such societies, involuntary manslaughter is seen as being pretty much just as bad as premeditated murder with respect to both the deserved punishment and the moral character of the transgressing individual. Notable examples of cultures where an action's intentionality does not significantly affect people's moral judgments include the Yasawan peoples of Fiji, the Hadzabe of Tanzania, and the Himba of northern Namibia and southern Angola. This clearly challenges my claim that MRJ is fundamentally linked to intention-sensitivity. Although this is a legitimate

concern, it presents no real danger to the foregoing evolutionary hypothesis.

The developmental evidence reviewed above (3.2) shows that the salience of intent in sociomoral assessment is present even during infancy. This suggests developmental canalization and perhaps even innateness, which is often considered to be indicative of an adaptive origin for the canalized trait (Khalidi, 2002; Lorenz, 1965, though see Griffiths and Machery, 2008). The presence of intent-based sociomoral evaluation in infancy (and continuously throughout development, if Margoni & Surian are correct) suggests the outcome-based moralities displayed by adults in some cultures have been instilled through enculturation processes which modify MRJ practices later on in development.[8] We ironically emerge, then, with a complete reversal of the Piagetian orthodox story for cultures in which MRJ is intention-insensitive: instead of starting with outcome-oriented judgments and then learning during development to consider intent, children in these cultures start off (as all children do) with intent-based judgments as infants, and then through enculturation shift to outcome-based judgments, which they employ into adulthood. Whether this explanation is correct or not is, of course, an empirical question which cross-cultural developmental research looking at the sociomoral evaluations of Yasawan, Hadza, and Himba infants could settle.

To argue this way is to suggest that the intention-sensitive form MRJ we see in WEIRD adults aligns with the way MRJ psychology has been 'designed' to function by hominin evolutionary history. This frames the handful of non-WEIRD cultures whose MRJ practices place little importance on intent as 'divergent' from how MRJ psychology 'naturally' develops. This will likely put some readers on edge, given the problematic history of Western evolutionary theorists creating capricious 'just-so' adaptive explanations for psychological tendencies they unreflectively take to be universal features of human nature (Buller, 2005). But in this instance, the developmental evidence does suggest that sociomoral evaluation is intention-sensitive from infancy on through later development. The claim that MRJ is innately intention-sensitive is thus not being justified here by the intuitive plausibility assessment of the WEIRD evolutionary theorist writing this, but by the results of studies in which enculturation into a WEIRD system of MRJ practices are unlikely to have played any meaningful factor.

8 It has been suggested that cultures with outcome-based moralities often have "mental opacity" norms which discourage reasoning and speech about others' mental lives (Robbins and Rumsey, 2008). However, there are most likely many different cultural processes that lead to outcome-based moralities beyond just mental opacity norms (see Barrett and Saxe (2021)).

There are, moreover, candidate explanations for when and why cultures tend to deviate from the intention-sensitive MRJ psychology with which humans, as infants, start out. H. Clark Barrett and Rebecca Saxe, in recent work (2021), claim that "there is no evidence for mental-state-disregarding cultural groups" when it comes to MRJ; instead, moral judgment "even within 'Western' moral, legal and philosophical traditions, depends on how situations are appraised and for what reasons judgements are being made" (2). Indeed, Barrett & Saxe argue that context and function determine to what degree a MRJ will privilege what these authors call "mind-mindedness," and these contexts and functions can be empirically shown to exist cross-culturally. Across societies, they conclude, "[m]oral judgements depend more on mental states when people are judging high status, competent individuals, when the violation was theft or injury, and when the purpose of the judgement is to express indignation. Moral judgements depend less on mental states when people are judging someone who is incompetent, or a whole group of people, when the action was inherently dangerous or involved a taboo about sex or food, and when the purpose of the judgement is to restore social cohesion" (6). There is much left to explore in the full landscape of MRJ practices across cultures, contexts, and functions. But we may safely consider intention-sensitivity a core psychological component of MRJ, despite the fact that other factors frequently generate variation in the MRJ practices we observe within and across cultures.

5 CONCLUSION

On the basis of evidence from comparative and developmental psychology, I have argued that intention-sensitivity arose much earlier in phylogeny than did moral responsibility judgment. Upon the emergence of MRJ, the capacity to discriminate between the intentional and non-intentional behaviors of agents was already present and thus available for use from the earliest stages in MRJ psychology. I then argued that incorporating intention-sensitivity into MRJ psychology would have improved fitness-enhancing functions of MRJ, especially with respect to partner choice, by enhancing reputational assessments through predictions about the likely future behavior of potential cooperative partners. If intention-sensitivity it was both available to and useful for moral responsibility judgment, as I claim it was, then we ought to conclude that the relevant ancestral population probably used it in just that way.

Assessing a transgressor's intent was not added to MRJ psychology late in hominin evolutionary history, after the core adaptive

functions of MRJ were well established. Discriminating between intentional and non-intentional behavior was a crucial aspect of MRJ from the outset, and was indispensable to attaining the fitness enhancements generated by MRJ. This perspective makes sense of the wide variety of comparative, developmental, and cross-cultural psychological evidence, all of which suggest that moral responsibility judgment is inextricably linked up to the assessment of intentions. More work will be needed to refine the granular details of the evolutionary picture and explain the bumpy contours of MRJ practices across the globe, but at this point, we should think that considering intent has been part of moral responsibility judgment from its very emergence in hominin evolution right up through to the present day.

REFERENCES

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., and Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, 113(17):4688–4693. Publisher: Proceedings of the National Academy of Sciences.

Barrett, H. C. and Saxe, R. R. (2021). Are some cultures more mind-minded in their moral judgements than others? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1838):20200288. Publisher: Royal Society.

Baumard, N., André, J.-B., and Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1):59–78.

Boehm, C. (2012). *Moral origins: the evolution of virtue, altruism, and shame*. Basic books, New York, NY.

Buller, D. (2005). *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. MIT Press.

Call, J., Hare, B., Carpenter, M., and Tomasello, M. (2004). 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. *Developmental Science*, 7(4):488–498.

Call, J. and Tomasello, M. (1998). Distinguishing Intentional From Accidental Actions in Orangutans (Pongo pygmaeus), Chimpanzees (Pan troglodytes), and Human Children (Homo sapiens). *Journal of Comparative Psychology*, 112(2):192–206.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–192.

Chernyak, N. and Sobel, D. M. (2016). "But he didn't mean to do it": Preschoolers correct punishments imposed on accidental transgressors. *Cognitive Development*, 39:13–20.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2):353–380.

Cushman, F. (2015). Punishment in Humans: From Intuitions to Institutions. *Philosophy Compass*, 10(2):117–133. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12192.

Dennett, D. C. (1987). Three kinds of intentional psychology. In *The Intentional Stance*, pages 43–68. ISBN: 026204093X.

Dunbar, R. I. (2004). Gossip in evolutionary perspective. *Review of General Psychology*, 8(2):100–110.

Eason, A. E., Doctor, D., Chang, E., Kushnir, T., and Sommerville, J. A. (2018). The choice is yours: Infants' expectations about an agent's future behavior based on taking and receiving actions. *Developmental Psychology*, 54(5):829–841.

Farrell, M. E. (2025). What would imaginary ancestors do? Thought experiments and intuitive plausibility in human cognitive evolution. *Biology & Philosophy*, 40(4):14.

Fischborn, M. (2023). The Evolutionary Roots of Moral Responsibility. *Philosophy of Science*, 90(4):817–835.

Geraci, A., Simion, F., and Surian, L. (2022). Infants' intention-based evaluations of distributive actions. *Journal of Experimental Child Psychology*, 220:105429.

Golding, W. (1955). *The Inheritors*. Faber and Faber.

Griffiths, P. E. and Machery, E. (2008). Innateness, Canalization, and 'Biologicizing the Mind'. *Philosophical Psychology*, 21(3):397–414. Publisher: Routledge _eprint: https://doi.org/10.1080/09515080802201146.

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, 128(3):451–474.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Hoffman, M. B. and Krueger, F. (2017). The Neuroscience of Blame and Punishment. In Menon, S., Nagaraj, N., and Binoy, V. V., editors, *Self, Culture and Consciousness*, pages 207–223. Springer Singapore, Singapore.

Joyce, R. (2006). *The Evolution of Morality*. MIT Press.

Khalidi, M. A. (2002). Nature and Nurture in Cognition. *The British Journal for the Philosophy of Science*, 53(2):251–272.

Kitcher, P. (2011). *The Ethical Project*. Harvard University Press, Cambridge, Massachusetts.

Kneer, M. and Machery, E. (2019). No Luck for Moral Luck. *Cognition*, 182:331–348.

Kneer, M. and Skoczeń, I. (2023). Outcome effects, moral luck and the hindsight bias. *Cognition*, 232:105258.

Krupenye, C. (2021). The Evolution of Mentalizing in Humans and Other Primates. In Gilead, M. and Ochsner, K. N., editors, *The Neural Basis of Mentalizing*, pages 107–129. Springer International Publishing, Cham.

Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308):110–114. Publisher: American Association for the Advancement of Science.

Li, J. and Tomasello, M. (2018). The development of intention-based sociomoral judgment and distribution behavior from a third-party stance. *Journal of Experimental Child Psychology*, 167:78–92.

Lorenz, K. (1965). *Evolution and the Modification of Behavior*. Chicago University Press, Chicago.

Mameli, M. (2013). Meat made us moral: a hypothesis on the nature and evolution of moral judgment. *Biology & Philosophy*, 28(6):903–931.

Margoni, F. and Surian, L. (2016). Explaining the U-Shaped Development of Intent-Based Moral Judgments. *Frontiers in Psychology*, 7.

Margoni, F. and Surian, L. (2020). Conceptual continuity in the development of intent-based moral judgment. *Journal of Experimental Child Psychology*, 194:104812.

McNamara, R. A., Willard, A. K., Norenzayan, A., and Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition*, 182:95–108.

Mercier, H. and Sperber, D. (2017). *The Enigma of Reason*. The enigma of reason. Harvard University Press, Cambridge, MA, US. Pages: vi, 396.

Nobes, G., Panagiotaki, G., and Pawson, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology*, 104(4):382–397.

Piaget, J. (1965). *The Moral Judgment of the Child*. The Free Press, New York.

Planer, R. J. and Sterelny, K. (2021). *From signal to symbol: the evolution of language*. Life and mind: philosophical issues in biology and psychology. The MIT Press, Cambridge, Massachusetts.

Robbins, J. and Rumsey, A. (2008). Introduction: Cultural and Linguistic Anthropology and the Opacity of Other Minds. *Anthropological Quarterly*, 81:407–420.

Royka, A. and Santos, L. R. (2022). Theory of Mind in the wild. *Current Opinion in Behavioral Sciences*, 45:101137.

Sperber, D. and Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective. *Mind & Language*, 27(5):495–518.

Stanford, P. K. (2018). The Difference Between Ice Cream and Nazis: Moral Externalization and the Evolution of Human Cooperation. *Behavioral and Brain Sciences*, 41(e95).

Strid, K. and Meristo, M. (2020). Infants Consider the Distributor's Intentions in Resource Allocation. *Frontiers in Psychology*, 11. Publisher: Frontiers.

Tomasello, M. (2016). *A Natural History of Human Morality*. Harvard University Press, Cambridge, Massachusetts.

Tomasello, M. and Carpenter, M. (2005). The Emergence of Social Cognition in Three Young Chimpanzees. *Monographs of the Society for Research in Child Development*, 70(1):1–152.

Vaish, A., Carpenter, M., and Tomasello, M. (2010). Young Children Selectively Avoid Helping People With Harmful Intentions. *Child Development*, 81(6):1661–1669. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8624.2010.01500.x.

Woo, B. M., Steckler, C. M., Le, D. T., and Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition*, 168:154–163.