Biological Theory
https://doi.org/10.1007/s13752-022-00397-y

ORIGINAL ARTICLE

A Second-Personal Approach to the Evolution of Morality

Carme Isern-Mas[1,2] · Antoni Gomila[2]

Received: 24 June 2021 / Accepted: 3 February 2022

Abstract
Building on the discussion between Stephen Darwall and Michael Tomassello, we propose an alternative evolutionary account of moral motivation in its two-pronged dimension. We argue that an evolutionary account of moral motivation must account for the two forms of moral motivation that we distinguish: motivation to be partial, which is triggered by the affective relationships we develop with others; and motivation to be impartial, which is triggered by those norms to which we give impartial validity. To that aim, we present the second-person standpoint of morality, first as Darwall conceives of it, and then as we reinterpret it from a naturalistic approach. Then we synthesize Tomasello's evolutionary account of morality, and Darwall's objections to it. To reply to those objections, building on Tomasello's proposal, we argue that the motivation to be impartial, and the feeling of obligation to comply with normative requirements, appeared when humans anticipated and, critically, internalized others' sanctions to the violation of social norms. Consequently, we posit that social norms and sanctions appeared first at the community level, and only after that were they internalized in the form of self-directed reactive attitudes. Finally, we derive some corollaries that follow from our proposal.

Keywords Evolution · Feeling of obligation · Human morality · Moral motivation · Moral norms · Second-personal morality

Introduction

An evolutionary account of morality has to address the question of moral motivation: why we feel compelled to act according to our moral judgments (Roskies 2003; Björnsson et al. 2014; Rosati 2016). Moral motivation is the part of our moral psychology by which we feel motivated to act in accordance with our moral judgments, that is, with what we judge to be the morally right action. On the contrary, failing to act according to our moral judgments, without excuse, makes us feel miserable, remorseful, and guilty. For instance, we feel horrified at the very thought of a taboo transgression, and we feel strongly obliged to help a friend in need. Such moral judgments do not need to be explicit. They might be implicit, as happens, for instance, in our moral emotions. When I regret what I did, I am implicitly making a negative valuation of my behavior. Therefore, what characterizes moral judgments is their motivational force.

We often feel obliged to help people we are affectively attached to, such as family and friends. The fact of having these sorts of relationships with other people makes us mind their wellbeing, so that we are motivated to behave in benefit of their interests. For purposes of simplicity, we use the expression "motivation to be partial" to refer to the motivation we feel to act in benefit of our close others' interests. With such cases, our spontaneous tendencies align with our moral judgment. Sympathy, care, or compassion are some of the psychological phenomena that have been identified as illustrative of this inclination towards close others. It does not seem mysterious that we feel motivated to act prosocially towards members of our circle of affection out of a

✉ Carme Isern-Mas
isernmas.carme@gmail.com

[1] Department of Philosophy, Florida State University, Tallahassee, FL, USA

[2] Laboratori de Sistemàtica Humana, University of the Balearic Islands, Palma de Mallorca, Spain

Published online: 04 March 2022

moral judgment when we already feel motivated to benefit them. However, an evolutionary account of the feelings of obligation must explain the motivation that stems from a moral judgment also when no such affective connection occurs. In this latter case we say that morality provides a "motivation to be impartial."

Moral motivation might have, at least, two different forms: motivation to be partial, which stems from our relationships with close others; and motivation to be impartial, which stems from impartial norms. In the case of motivation to be partial, we feel motivated to act morally towards others out of the specific relationships that bind us with them. In such cases, we would not be so motivated were the recipient of our behavior anyone in general. But crucially, we sometimes feel morally motivated to be impartial, that is, we feel motivated to act morally out of an impartial norm. In these cases, we would feel similarly motivated whoever is the person involved in the situation, regardless of our relationship to them.

Both forms of moral motivation could align. Someone could be motivated to help a friend either because of the specific relationship that they both hold (motivation to be partial), or because this is how friends ought to act to one another, regardless of the specific person who is in that position (motivation to be impartial). The main difference between the two of them is their source: either the specific relationship, or a general norm. Contrary to the Kantian approach, which restricts moral motivation to instances of what we call motivation to be impartial, we argue that both cases are instances of moral motivation. The fact that motivation to be partial aligns with our natural inclinations is not a reason to reject it as a moral motivation. What makes a motivation moral is its psychological structure, that is, that it comes with a feeling of obligation and with a normative expectation that has been internalized. Consequently, moral motivation does not need to be contrary to our inclinations. On the other hand, this does not preclude the possibility that co-occurrence of these two forms of moral motivation might pull us to act in incompatible ways, hence the common experience of conflict in moral dilemmas (Christensen and Gomila 2012).

An evolutionary account of moral motivation must account for both forms of moral motivation. However, some accounts just focus on the emergence of the motivation to act in benefit of those to whom we are somehow bonded (for instance, de Waal 2008; Trivers 1971). In these cases, it is assumed that motivation to be impartial somehow emerged as a generalization of the former. Other proposals only focus on the evolution of a motivation to be impartial, which they take to be a requisite for morality (for instance, Gibbard 1982, 1989; Stanford 2018).

Tomasello (2016, 2019) offers a genealogy of both kinds of moral motivation. In his terms, he first accounts for the emergence of sympathy and fairness as motives for moral behavior, which we interpret as motivation to be partial. Next, he characterizes impartiality as a form of generalized partiality, following the classical strategy of the moral sense school, best exemplified by Smith (1759). Tomasello's project is grounded in Stephen Darwall's (2006) second-person view of morality. Yet, as Darwall (2018) has objected, his evolutionary proposal of the motivation to be impartial risks circularity.

In this article, we propose an alternative evolutionary account of moral motivation in its two-pronged dimension, building on the discussion between Stephen Darwall and Michael Tomassello. In reply to Darwall's objections to Tomasello's proposal, we argue that both forms of moral motivation are grounded in the way we interact with others: either in the affective relationships we develop with them, which trigger a motivation to be partial to them, or in the demands that we reciprocally address to, and recognize from, each other, which require some impartial validity. We further argue that the feeling of obligation appeared when humans anticipated and, critically, internalized others' sanctions to the violation of social norms. Consequently, we posit that social norms and sanctions appeared first at the community level, and only after that were they internalized in the form of self-directed reactive attitudes.

In the next section, we present the second-person standpoint of morality, first as Darwall conceives of it, and then as we reinterpret it from a naturalistic approach. In the third section we synthesize Tomasello's evolutionary account of morality; and, in section four, Darwall's objections to it. In the fifth section we present our evolutionary proposal and argue that it avoids those objections. Finally, we derive some corollaries which follow from our proposal.

The Second-Person Standpoint of Morality

We agree with Tomasello that a second-person approach to morality has the potential to account for the evolution of moral motivation in both its forms, motivation to be partial and motivation to be impartial. This second-person approach is owed to Stephen Darwall in *The Second-Person Standpoint* (2006). Yet the aim of Darwall's project is not to give an evolutionary account of morality, not even a descriptive one. Instead, he aims to conceptually analyze some key moral notions such as respect, obligation, right and wrong as involving intrinsically a second-person standpoint, that is, as being grounded in the relationships between subjects. Despite its analytical nature, Darwall's proposal can be developed as a naturalistic project, which can shed light on the psychology of morality, and its evolution.

Darwall's Second-Person Standpoint

The second-person standpoint is "the perspective you and I take when we make and acknowledge claims on one another's conduct and will" (Darwall 2006, p. 3). When someone steps onto my foot, I assume that I have a second-personal authority as a person to demand that the other move their foot. I also assume that they, as a person, have the right to demand something of me; and that we both can hold the other and ourselves accountable if either of us does not comply with the other's demand without an excuse. According to Darwall, morality presupposes this second-person standpoint. Morality consists in the practices of holding each other accountable and responding to those claims. Accordingly, he understands second-personal morality "as normative requirements that obligate all moral agents," and which "consists in demands with which second-personally competent agents are mutually accountable for complying" (Darwall 2018, p. 809). Second-personal morality presupposes that the participants can acknowledge each other's second personal authority to raise demands; and that they can also hold the other, and themselves, accountable for incompliance without excuse. Thus, "second-personal interactions always have the seeds of universalism in them" (Darwall 2018, p. 811).

As a consequence, moral obligation is defined as "what those to whom we are morally responsible have the authority to demand that we do" (Darwall 2006, p. 14), whereas moral motivation is defined as the "intrinsic desire to comply with moral demands to which one may be legitimately held accountable" (Dill and Darwall 2014, p. 14). Therefore, Darwall equals moral motivation with motivation for impartiality, and dismisses motivation for partiality as a source of moral motivation, as it fails the universalizability criterion. According to Darwall, we are motivated to act morally because we are aware of those actions for which we could legitimately be held responsible; and we perceive moral norms as objective because they are those norms the violation of which would be justifiably blameworthy. Therefore, moral motivation comes from the experience of norms as objective, not from concern for others.

Apart from overlooking our motivation for partiality, this view does not aim to describe our moral psychology, rather to analyze our moral concepts. According to Darwall, a norm is moral if in case of incompliance without excuse, a rational and free agent could legitimately be held accountable for any other free and rational agent who is part of the moral community, including themselves. Yet this second-personal interaction is not factual but axiomatic: it does not need to take place, and it does not matter whether it takes place. The value of the second-person standpoint is justificatory. It does not explain the emergence of flesh and blood individuals who are motivated by morality.

A Naturalistic Approach to the Second-Person Standpoint

Despite its analytical nature, Darwall's second-person perspective presupposes a descriptive dimension: it requires agents with a set of psychological capacities for intentional interaction, such as basic perspective-taking, an impartial perspective, some degree of self-control and self-regulation, the ability to hold oneself accountable, and the ability to recognize oneself and others as having shared second-personal authority as mutually accountable second-personally competent agents (Darwall 2006, 2018). Darwall does not make it explicit, yet this set of psychological capacities is presupposed in his notion of "second-personal competence." Second-personal competence is the "capacity to view oneself and another from a second-person standpoint, in which both oneself and the other recognize one another as having the same shared competence and authority to hold themselves accountable to one another" (Darwall 2018, p. 808). In other words, to be second-personally competent, or to be moral, is to have the capacity to be sensitive to the normative claims that any single member of the community can address to me, including myself.

Although Darwall does not aim to provide an evolutionary account of the emergence of such second-personal competence, he concedes that it "is the object of natural selection under the conditions of obligate collaborative foraging" (Darwall 2018, pp. 807–808). The problem with his standpoint, though, is that the process of natural selection does not guarantee the emergence of the sort of rational and deliberative agents his view prescribes.

A naturalistic approach to the second person, on the contrary, is interested in the explicit characterization of the psychological capacities of the agents that did evolve. In particular, it is interested in the capacity for mutual intentional attribution, as well as emotional expression and recognition, required for the face-to-face, intentional, and reciprocal interaction in real time, which our moral concepts presuppose (Christensen and Gomila 2012). This capacity has also been called "the second-person perspective" of psychological attribution (Gomila 2001, 2002, 2015). Morality presupposes agents capable of this kind of intersubjective interactions (Gomila 2008; Isern-Mas and Gomila 2018). Demands are addressed and honored within these kinds of interactions. As a matter of fact, an evolutionary account of moral motivation consists in the effort to make explicit the evolution of these psychological capacities if they make possible the emergence of the required forms of motivation.

Michael Tomasello on the Evolution of Morality

Michael Tomasello's proposal in *A Natural History of Human Morality* (2016), and in "The Moral Psychology of Obligation" (2019) can be interpreted as an account for the emergence of both moral motivation to be partial, and moral motivation to be impartial. Tomasello's proposal relies on two related points: his well-known case for joint agency as the key to human evolution, and his notion of a second-person morality as a basic level of normative regulation that emerged from joint agency.

Tomasello contends that morality emerged out of a group of individuals who competed to survive. He departs from the common ancestor of chimpanzees and humans and proposes that a first transition towards morality was the appearance of cooperation. Cooperation appeared, as is standardly assumed, through reciprocity, understood as a sort of delayed mutualism, when the short-term loss of cooperation was compensated by the long-term gain. Conceding this first transition, the challenge is to explain how self-interested cooperators whose motives for cooperation were prudential gave rise to morally motivated agents.

To address the challenge, Tomasello introduces a twist to the standard view of evolutionary game theory of agents as self-interested and prudential. For before engaging in reciprocal cooperation these cooperative individuals were not strategic, rational, isolated players who could opt out of the social dilemmas they faced. They were already social. The common ancestors between humans and *Pan* already lived in social groups, and they related to each other, as kin, friends, or cooperative partners, in a way similar to how living chimpanzees and bonobos currently do. According to Tomasello, we share with other great apes our concern for those with whom we have close ties. This feature was selected through kin selection, to promote the survival of those carrying our genes; and then reinforced through reciprocity and mutualism, to promote the survival of those who help us. Therefore, when early humans encountered the situation where they were better off if they cooperated, they were not individualistic, rational, and strategic agents. Rather, they were already affectively interconnected, bonded agents with a set of affiliative relationships. For this reason, their motivation to cooperate was not prudential but other-regarding from the start: they already cared for each other, and were capable of empathy.

In this cooperative interaction, agents adjusted their behaviors to each other, formed expectations about others' behaviors, and reacted negatively when those expectations were violated. Yet those expectations were still merely empirical, or statistical: they were about how an agent *would* act. Hence their transgression caused frustration or disappointment (Engelmann et al. 2017). To count as fully moral, interactive partners should form normative expectations as well, that is, expectations about how an agent *should* act and what they *deserved*. Such expectations can be said to be in place when an agent's transgression causes the moral emotion of resentment or indignation which, unlike frustration or disappointment, implicitly addresses a claim to the transgressor, asking for recognition of wrongdoing (Strawson 1974; Darwall 2006, 2013).

Tomasello proposes that this sort of normative demand appeared later in the human lineage because of the change in the ecological conditions of life in the savannah. In adapting to this new environment, the best adaptive strategy was to cooperate through joint intentional activities. Cooperation became necessary to survive, and its more adaptive form was as a joint intentional activity. According to Tomasello, the seeds of morality can be found in this new form of cooperation because it required the appearance –-in our human ancestors— of three new psychological abilities: cognitive processes of joint agency, social-interactive processes of second-personal agency, and self-regulatory processes from joint commitments. The key for the emergence of morality lies in the development of these psychological capacities that made possible joint intentional activities.

Joint intentional activities "can give rise to a 'we-over-me' psychology that represents the beginning of all things moral" (Engelmann and Tomasello 2017, p. 11). Tomasello calls this kind of morality that emerges from joint activity "second-personal morality." According to Tomasello, a second-personal morality is a dyadic morality of face-to-face interactions between agents collaborating, and feeling responsible to one another, as a jointly committed "we." In fact, the phrase "second-personal morality" is meant to emphasize the scope of this kind of morality, which is reduced to the dyad and, specifically, to the dyad's collaborative activity.

It is at this stage that proper motivation to be partial appears, as agents are already capable of some basic level of normative assessment, reduced to the dyad.

Tomasello proposes a final transition from this second-personal morality to the emergence of objective morality, which he views as the precondition for the motivation to be impartial. When two partners engage in joint intentional activity, they get to see each other as part of a cooperative activity which must follow some standards, that is, role ideals about how to perform a part of the joint activity. They also understand what Tomasello calls "self-other equivalence"; both partners know that each could perform the other's role and that the other could perform theirs. Once the partners understand both role ideals, and self-other equivalence, they develop an impartial point of view about the agreements initially reached in the dyad.

To go from the norms of the dyad to the norms of the community, Tomasello resorts to a version of Smith's impartial spectator (1759). According to Tomasello (2016, 2018), the impartial perspective is enhanced through two interrelated processes: generalization of role standards and awareness of third-parties' assessment (Tomasello 2019). According to Tomasello, the clue lies in the fact that the dyadic collaboration "occurs between individuals in a larger pool of collaborators in a loosely structured social group" (Tomasello 2018, p. 825). The impartial perspective is acquired through generalization after several interactions with different partners and through increasing awareness of "how others in the potential pool of collaborators were viewing, or would view, certain kinds of actions within a collaboration" (Tomasello 2018, p. 825). The force of the commitment is thus enhanced by the potential collaborators who witness the collaboration. The role of these bystanders contains the seeds of the "fully moral kind of objectivity and normativity" (Tomasello 2018, p. 825). With this emphasis, Tomasello explains how second-personal morality "had at least some generalizing tendencies–implicit reference to others in the pool of collaborators– that provided the external reference point needed for participants in a collaborative activity to give socially normative forces their due" (Tomasello 2018, p. 826).

Objections Raised by Stephen Darwall

Darwall has raised two circularity objections against Tomasello's evolutionary scenario. First, second-personal morality must already involve universalizing tendencies to be considered morality at all. According to Darwall, it makes no conceptual sense to call morality what only concerns a dyad of agents. Second, joint intentional agency already requires impartiality, and therefore it cannot account for its emergence, on pain of regress.

The first objection is that second-personal morality must already include universalizing tendencies to qualify as morality at all. According to Darwall, even if second-personal processes take place only in the dyad, the demands addressed between participants must be of universal application. Although moral requirements are grounded in second-personal interactions, they are not "restricted to obligations each party has to the other, within the interaction" (Darwall 2018, p. 810). There is no qualitative difference between obligations within the dyad and obligations towards the group. The demands that the participants of the dyad address to each other mean to be valid for any member of the community. To interact "second-personally," individuals need to assume that they would, and should, interact following a general norm about how persons should interact with each other. Accordingly, a second-personal demand is not merely "a bare or naked demand" which is only valid in the dyad, but "a putatively legitimate one" (Darwall 2018, p. 808) which is "committed to presuppositions of universal human morality" (Darwall 2018, p. 809). Tomasello's second-personal morality lacks these universalizing trends.

This is acknowledged by Tomasello, who concedes that second-personal morality has "only partially universalizing tendencies" (2018, p. 825) and hence cannot be considered "full" morality. This difference could be seen as terminological, with Darwall using morality for norms that apply to everybody and Tomasello distinguishing a sense of morality for norms that apply only for each dyad of interacting agents. But this move does not help with Darwall's second point.

Darwall's second objection is that impartiality is in fact a requirement for joint agency, and therefore, it cannot emerge out of it. According to Darwall (2018), impartiality was, for the participants in the dyad, "something that the mutual intelligibility of their collaboration was presupposing," rather than "something it was creating" (2018, p. 812). The reason for this claim has to do with the contractualist view of morals Darwall assumes, according to which the partners to any enterprise must recognize each other as potential partners (Darwall 2018). For two individuals to jointly act, they must assume the second-personal authority (in Darwall's sense) of anyone capable of entering this sort of collaborative activity; they must presuppose that both cooperators have an authority to issue claims and demands which is previous and independent of their joint activity. This kind of independent, impartial, authority is revealed by the fact that both cooperators must assume from the start that anybody can be a partner, and that anybody has a right to refuse the invitation to collaborate. In Darwall's words, "it is only by reciprocally recognizing one another's basic independent second-personal authority that you and I can form a committed we" (2018, p. 807). Therefore, joint agency cannot be the source of impartiality. In Tomasello's account, participants develop a motivation to be impartial by engaging in joint intentional activity: they develop mutual respect, commitment, and trust towards others, as they interact with them. Yet, according to Darwall, these are preconditions for joint agency in the first place.

In sum, Tomasello's proposal fails to provide an account of how standards, or norms of a kind that justify the term, can appear within two-person collaborations, and then extend to the community. Darwall argues that a norm must hold for the community right from the start to count as a norm. Similarly, Tomasello proposes that moral norms emerge out of joint intentional activity, but Darwall objects that the recognition of the normative authority of subjects is a condition of possibility of joint agency in the first place. Since Tomasello's second-personal morality consists in the agreements reached by the dyad, it is difficult to explain both how the dyad can be in a joint activity without such standards; and how those agreements could emerge from, and transcend, the dyadic collaboration of agents. Without such impartial standards, no impartial motivation is possible. Therefore, it seems that Darwall is right and that the evolution of fairness and impartiality are not successfully explained in Tomasello's proposal.

An Alternative Second-Person Account

The difficulties of Tomasello's account come from the assumption that morality emerged in dyadic interactions and was later generalized to the community. Instead, an evolutionary account of moral motivation must be framed at the community level from the start. On the other hand, social norms should not be conceived as emerging to reinforce moral norms as part of a group's identity, as Tomasello proposes. Instead, moral norms, and their distinctive motivational power, must be conceived as a specialization of social norms: those whose sanctions also became internalized.

Our phylogenetic scenario starts from a stage in which norms are not yet present. In this pre-normative stage, our ancestors were already competent at some form of second-person intentional attribution (introduced in the second section, as our naturalistic alternative to Darwall's notion). Our ancestors evolved towards ultrasociality, and second-person intentional attributions developed along with an increase in the expressive bodily cues that grounded them, so that increasing forms of coordination were made possible. Those second-person intentional attributions mediated emotional reactive attitudes that implied some form of implicit appraisal, not yet normative, but often motivated by a broad class of prosocial motivations (attachment, sympathetic concern, empathy in general).

To explain how a further normative level emerged, three aspects must be considered. First, that any two agents may interact repeatedly. Second, that each agent may interact with many other different agents in a group, in succession. And third, that each interaction involves the two partners, and occasionally one (or more) third party–we call it the partial bystander, insofar as that individual might be any of the other members of the group somehow related to each of the interactors. And the question is how to go from the sort of intentional interactions and emotional reactions described to the emergence of moral norms, as a further layer of conceptual articulation that assigns duties, affirms rights, and fixes obligations to each other, that is, duties and obligations that are felt as compelling.

The key notion for such an account, in our view, is that of expectation. Expectations concern physical events, but they also develop for intentional interactions, as one's behavior is often conditional on the behavior of others. Repeated interactions give rise to reciprocal expectations, that is, anticipations of what is about to happen when relating to another agent. These expectations might be related to a particular dyad, but given the scope of interactants, they might also emerge at the group level.

After several interactions, these merely descriptive expectations acquired a proto-normative nature, as their violation started to trigger non-normative reactive attitudes such as disappointment, or frustration, or anger. This is indeed what happens in our ordinary attributions of blame, which depend on how much of the blameworthy behavior is done or how people go about doing it (Bostyn and Knobe 2020). These reactions to violations of expectations work as a social sanction and play a main role in turning those descriptive expectation into normative ones. At this stage, the partial bystander might have had a role in sanctioning or questioning the adequacy of the reaction: whether such a reaction made sense in the situation, and whether some consequence was in order – what is called third-party punishment in accounts of the evolution of cooperation (Fehr and Fischbacher 2004; Gintis and Fehr 2012). In other words, a failed expectation elicited some form of social control to redress the situation, either by the one disappointed or by an ally. Alternatively, it could have been that the expectation did not hold and had to be modified.

These elements are enough to account for the emergence of social norms. According to the influential account of Bicchieri (2006, 2016), social norms involve two levels of agent expectations: empirical expectations about what people normally do in the circumstances, and normative expectations about what the other people think someone ought to do. In her words: "A social norm is a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (empirical expectation), and (b) that most people in their reference network believe they ought to conform to it (normative expectations)" (2016, p. 35).

Notice that these normative expectations need not be explicit. They can be implicit in the estimated probability that the group will force conformity. The relevant "ought" of the norm is cashed out as the expectation that other agents are ready to intervene to make it happen that way (through third-party punishment, reputation tracking, or any other form of social control). Long before discussion about norms and potential agreements on them were possible, social norms emerged when the agents involved developed such twofold expectations. While the normative ones require some robust means of control against deviance, it is not required that a central authority be in place, nor any explicit formulation, nor that such expectations apply across the group. In our view, as stated, social norms begin to operate when agents begin to take such normative expectations into account. And expectations become normative when agents are ready to sanction those failing to comply.

Once norms emerged within a community, some norms became moral at a further stage. At this point, before detailing what else was needed for moral norms to appear, we must show first that such moral norms would improve fitness, that agents and groups capable of moral normative guidance would be better-off than agents with just social normative guidance. In other words, what is to be explained is the emergence of the feeling of obligation, the motivational feeling to comply with some particular social expectations that characterizes the binding force of moral norms. In Bicchieri's terms, social norms are conditional, but moral ones are unconditional (Bicchieri 2006). Our idea is that it is through this psychological twist that norms became moral, when they were internalized, and seen as well as external and objective (Stanford 2018). Internalization of norms has already been shown to be an evolutionarily stable strategy (Gintis 2003; Gavrilets and Richerson 2017).

From an evolutionary point of view, then, morality emerged because the feeling of obligation countervailed the long-term costs of nonconformity, that is, the costs of not doing what is expected by interacting partners and bystanders. These costs might be much bigger than the short-term benefits of not doing so (Barclay 2013; Gaus 2015). Felt obligations were forms of commitment to non-myopic courses of action—a role that has been attributed to emotions in general (Frank 1988). Performing according to normative expectations and group demands signaled the reliability of the agent, thus reinforcing their value as a partner, and the likelihood of receiving help from others. For partners with strong ties, existing motivational mechanisms may be enough to maintain cooperation and coordination, but as the human groups became bigger a new adaptation was selected. Thus, the emergence of feelings of obligation made sense for groups whose members might not always be close kin. If the selective pressure primed bigger groups, groups of agents able to feel obliged to perform as expected would be better-off than those that were just motivated by kin and by the various forms of strategic reciprocity.

What psychological mechanism might be necessary for the feeling of obligation to appear? In our view, the key lies in "internalization" (Gintis 2003; Gavrilets and Richerson 2017). The sense of duty that comes with moral norms emerges when we internalize an anticipated sanction, or blame (Tomasello 2019). This internalization was possible through those emotions whose appraisal involved a self-assessment. Emotions such as guilt, shame, remorse, or pride constitute internal assessments of one's actions in relation to the expectations at stake. As Darwin (1872) already observed, these emotions are typically induced by what one thinks that others will think about oneself. In addition to involving self-assessment, they also motivated the agent to repair the relationship with the complaining partner and so diminish reputational costs (Ketelaar and Tung Au 2003; Sznycer et al. 2016). For instance, if someone felt guilty, they would assess their actions as blameworthy, and they would probably seek to reestablish the relationship with the person who had been wronged (Scanlon 2008). The intensity of the feeling of obligation would vary depending on the severity of the social demands and the anticipation of the costs of failing to comply with the relevant expectations. For this reason, those emotions emerged first towards affiliated others, but their structure is that of motivation to be impartial.

Moral emotions, then, worked as internalized sanctions for not complying. The idea is that moral norms took hold on a set of already existing motivations and preferences for cooperation and constituted a reinforced motivation. It clearly made us receptive to others' demands and claims, and ready to take them into account, when they match our own appraisals–and to rebut them when they do not. Correspondingly, such self-valuations could also be applied to others, as in blame and resentment, motivating these demands and claims, but also in the positive, as in envy and admiration, and trust. In this sense, the second person addresses and demands described by Darwall constitute an important dimension of the evolution of morality, when naturalized. Obligations made our ancestors sensitive to the reactive attitudes of their partners in a very special way.

Importantly, this psychological mechanism can take any normative content. It is a dispositional structure that can be applied to any social group and to any kind of expectation (Sznycer et al. 2018). Accordingly, different groups may moralize different norms; a group may come to change their moral norms as well. On this subject, two comments might be in order. First, from the naturalistic approach that we take, morality does not have a universal content. It is rather a facet of human mental life. Consequently, what we aim to account for is not the universality of the content of morality but the motivation to be impartial. Second, the kind of valuation that is involved in moral emotions might not be explicit, and it does not need to take the form of a judgment. Explicit formulation of norms as general statements probably took much longer to emerge, through language, perhaps in the context of overcoming behavioral conflicts (Gibbard 1989).

Anthropological evidence of such a scenario can be found in the importance of honor in human societies–a term that has to do with reputation. Honor codes can be viewed as an early form of moral norms, in that they are felt as an obligation and their function is also to promote cooperation within a group, imposing a very high cost on noncompliance. While honor is important across the board, honor cultures are those that demand of their members readiness to kill and die (Leung and Cohen 2011). Revenge and purification is the function of the prescribed violence (Handfield and Thrasher 2019). Revenge is addressed at out-group agents; purification, at in-groups who failed to conform to what was expected from them, thus "bringing dishonor" to the whole group. In our view, the importance of honor norms in human societies makes clear that the idea that any human being deserves equal respect is relatively recent. Similarly, apology and other forms of conciliatory behavior are not universal forms of repairing a relationship; recovering honor involves some form of violence. Very often violent acts are performed out of a moralistic motivation (Black 1983; Fiske and Rai 2014). Defending one's honor in this way equally provides information about the reliability of an agent as a potential partner, as somebody to be trusted (Thrasher and Handfield 2018).

Our proposal also finds support in ontogenetic evidence. On the one hand, the psychological structure of obligation and the self-conscious emotions that signal it would have become canalized like the rest of our basic motivational systems. Thus, self-conscious emotions are universal but they may be elicited by a diversity of circumstances (Sznycer et al. 2018). On the other hand, the content of the relevant moral norms must be learnt from the previous generation, in a process that could allow for changes in the norms. The self-reflection required for self-conscious emotions seems to develop at the end of the second year of life (Lewis 1992), and the first self-conscious emotion to appear is embarrassment, in the third year of life (Lewis 1995), which involves a distinctive psychophysiological reaction (Lewis and Ramsay 2002). Shame and guilt follow through after the third anniversary (Lewis 1992; Tangney 1995), revealing the sensitivity to the attitudes of others towards oneself. Children thus begin to discover the relevant norms by discovering the expectations of those with whom they interact and adhering to them.

In summary, we have offered an account of how morality could have emerged out of agents competent at second-person interactions–able to recognize goal-directed behavior and emotional expression in the basic way of second-person attributions. After the emergence of social norms, the key transition was the appearance of self-conscious emotions, which made our ancestors sensitive to their partners' appraisals, and internalized them. Consequently, new patterns of interaction were made possible, and new conceptual developments and linguistic practices probably followed. These new patterns may well be called second-personal, if it is made clear that the sense intended concerns the sensitivity to mutual demands of interacting agents—rather than to their reciprocal attributions per se.

Consequences of our Account

The difference between Tomasello's proposal and ours lies in that his proposal derives impartiality from the structure of dyadic interactions, while we view it as the outcome of the group dynamics. For this reason, the standard is already properly normative: it is endorsed not just by the collaborative dyad, but also by the community at large. Given this common ground, the collaborative agents can imagine how any other member of the moral community would react to their transgressions, and can reason from an impartial, moral, perspective.

Our account avoids Darwall's two objections. On the one hand, our proposal honors his point that norms involve universalizing tendencies. Norms appear as shared expectations revealed in the valuations implicit in emotional reactions. Empirical expectations become proto-normative when the group dynamics make them stable and independent of any specific individual. On the other hand, we avoid the second objection, related to the structure of joint agency, as we do not focus on a single form of interaction.

From the picture we have proposed of the evolution of morality, then, some corollaries follow. First, from an evolutionary point of view, norms are not mind-independent entities. Norms are the unplanned, unexpected result of individuals' interactions; they are the objectification of the implicit normative expectations that we form about others while interacting with them, when the group acquires its own dynamics. In this sense, they are similar to grammar, the norms of languages[1]: they are not mind-independent, because they actually depend on the minds of the speakers of that language and can be changed by them; but they are still objective because they cannot be just made up by any speaker alone because they require interaction. Importantly, this does not mean that all norms are "just conventions": due to their intersubjective nature they involve an emotional mechanism of commitment and valuation which makes us feel them as more binding, objective, and authority-independent (Turiel 1983).

Second, moral emotions are the bridge from mere interactive and reciprocal adjustment to each other to morality.

[1] We are indebted for this example to Shelly Kagan. A similar idea has been developed by John Mikhail (Mikhail 2011). We are indebted for this reference to an anonymous reviewer.

They express and reveal an implicit level of normativity, and hence they might not require language (Rowlands 2012). This kind of proto-normativity might be already present in nonhuman animals (de Waal 1996, 2014; Bekoff 2004; Brosnan 2006; Andrews 2009; Pierce and Bekoff 2012; Brosnan and de Waal 2014; Vincent et al. 2019), and children (Blake and McAuliffe 2011; Castelli et al. 2014; Blake et al. 2015; Engelmann and Tomasello 2019). Nevertheless, one could say that, although moral emotions are not expressed through language, they actually require propositional content as they are propositional attitudes (Gomila 2012). Either if they involve an implicit, nonverbal, normativity, or if they involve propositional content anyway; their intermediate position between explicit norms and behavioral adjustments and between subjective preferences and impartial standards maintains.

Third, moral norms come originally from behavioral dispositions based on emotions, which are lately shaped by interaction with others in a way similar to how traffic norms shape our behavior while driving (Sie 2014). It should not surprise us then to find that we are not always impartially motivated agents, but rather partially motivated ones with some preferences for our "near and dear" (Wolf 2012); and that we see the moral norms of our group as more objective than those of other groups (Sarkissian et al. 2011). From an evolutionary perspective, morality in general, and moral emotions in particular cannot be as impartial as some expect them to be (Prinz 2011; Bloom 2014), and the motivation for partiality must be taken into account. The question about whether an agent with only a motivation for partiality to act morally would count as a fully moral agent becomes a terminological one: with Kantians willing to reserve the term "morality" just for those agents capable of full-blown normative guidance, and those from the school of the moral sense preferring to view it as a graded, fuzzy, term.

Fourth, we can evolutionarily explain why moral judgments are experienced as both motivating, and objective. This apparent contradiction of moral judgments having simultaneously the appearance of both objective statements that state something about the world, and subjective states that motivate us to act, constitutes what Smith (1994) calls "the Moral Problem." Smith's worry in *The Moral Problem* (1994) is to make sense of this paradoxical appearance of moral judgment "with the standard picture of human psychology that we get from Hume" (1994, p. 14). Putting aside the discussion about Hume's description of human psychology, our worry here has been of another kind: in explaining impartial motivation, we have provided an evolutionary account for moral judgments being experienced as both objective, and motivational.

Finally, according to this view morality does not emerge to solve "the problem of cooperation" (Greene 2013). The kind of norms that can emerge from cooperation are just coordination norms, but not necessarily moral ones (Gauthier 1986). The problem of cooperation can be solved in other ways that do not require morality, such as group selection (Sober and Wilson 1998), kin selection (Hamilton 1964), and mutualism or reciprocal altruism (Axelrod 1981; Wilkinson 1990). To explain why we humans are moral, we need a different starting point: morality emerged not from strategic, self-interested individuals who had to cooperate to survive; but from social individuals who related to each other through second-personal mental state attribution. Accordingly, what was first selected in our species was the need to establish long-term bonds with others, and to relate to them. The second-person perspective of psychological attribution contributes to this, especially in nonverbal creatures. Morality emerged within these second-personal interactions; not because of what joint action required, but because of the set of common expectations developed at the community level, which became normative, as we have presented.

Conclusion

While sharing with Tomasello and Darwall the project of an intersubjective grounding of the genealogy of morals, our proposal rejects the common assumption that morality emerged in a scenario of strategic, self-interested, cooperators. Instead, morality emerged in groups of cooperators who already had an interest in the wellbeing of others and had normative expectations. As we have argued, we have been evolutionarily selected to be motivated to bond with others and to take others' interests into account; hence our motivation for partiality to act morally towards others, once we became capable of normative guidance.

As Tomasello contends, those cooperators were already tied to others, and motivated by sympathy to act prosocially towards kin, friends, and potential partners. And, crucially, they were capable of a second-personal mental state attribution. That is, they were able to interact with others through a spontaneous, emotional, and engaged attribution of mental states. Through this sensitivity and adjustment to others, expectations develop. These expectations become normative due to generalization of interactions to the community, and third-party endorsement or sanction. After that, some of these normative expectations involved a new, internal, sanction, in the form of the feeling of obligation and the self-conscious assessment of one's actions. Through this process we came to understand the moral norms as also objective, independent of our own assessment. Their motivational nature becomes now impartial because it is derived not from the preference for the recipient of the behavior, but from the norm itself.

Acknowledgments Versions of this paper were presented at the Second Antwerp Summer School in Philosophy and Society (Belgium, 2017) and at the Summer School “Ethics, Empathy and Errors” (Estonia, 2019). Many thanks go to those attending the events for helpful discussion and feedback. This work was supported by the Spanish Ministry of Education, Culture, and Sports under grant FPU14/01186; and by the Spanish Ministry of Economy, Industry, and Competitiveness under Grant FFI2017-86351-R. We thank Josep Corbí, Stephen Darwall, Katrien Schaubroeck, Maureen Sie, Joaquin Valdivielso, and anonymous reviewers for comments on earlier drafts.

Authors' Contributions Not applicable.

Funding This work was supported by the Spanish Ministry of Education, Culture, and Sports under grant FPU14/01186; and by the Spanish Ministry of Economy, Industry and Competitiveness under Grant FFI2017-86351-R (AEI/FEDER, EU).

Availability of Data and Material Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest The authors declare that they have no conflicts of interest.

References

Andrews K (2009) Understanding norms without a theory of mind. Inquiry 52:433–448. https://doi.org/10.1080/00201740903302584

Axelrod R (1981) The emergence of cooperation among egoists. Am Polit Sci Rev 75:306–318. https://doi.org/10.2307/1961366

Barclay P (2013) Strategies for cooperation in biological markets, especially for humans. Evol Hum Behav 34:164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Bekoff M (2004) Wild justice and fair play: Cooperation, forgiveness, and morality in animals. Biol Philos 19:489–520. https://doi.org/10.1007/sBIPH-004-0539-x

Bicchieri C (2006) The Grammar of society: The nature and dynamics of social norms. Cambridge University Press, New York

Bicchieri C (2016) Norms in the wild: how to diagnose, measure, and change social norms. Oxford University Press, New York

Björnsson G, Eriksson J, Strandberg C, Francén Olinder R, Björklund F (2014) Motivational internalism and folk intuitions. Philos Psychol 28:715–734. https://doi.org/10.1080/09515089.2014.894431

Black D (1983) Crime as social control. Am Sociol Rev 48:34–45. https://doi.org/10.2307/2095143

Blake PR, McAuliffe K (2011) “I had so much it didn't seem fair”: Eight-year-olds reject two forms of inequity. Cognition 120:215–224. https://doi.org/10.1016/j.cognition.2011.04.006

Blake PR, McAuliffe K, Corbit J, Callaghan TC, Barry O, Bowie A et al (2015) The ontogeny of fairness in seven societies. Nature 528:258–261. https://doi.org/10.1038/nature15703

Bloom P (2014) Against Empathy. Bost Rev 1–11

Bostyn DH, Knobe J (2020) The shape of blame: how statistical norms impact judgments of blame and praise. PsyArXiv Preprints. 24. https://doi.org/10.31234/osf.io/2hca8

Brosnan SF (2006) Nonhuman species' reactions to inequity and their implications for fairness. Soc Justice Res 19:153–185. https://doi.org/10.1007/s11211-006-0002-z

Brosnan SF, de Waal FBM (2014) Evolution of responses to (un)fairness. Science 346. https://doi.org/10.1126/science.1251776

Castelli I, Massaro D, Bicchieri C, Chavez A, Marchetti A (2014) Fairness norms and theory of mind in an ultimatum game: judgments, offers, and decisions in school-aged children. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0105024

Christensen JF, Gomila A (2012) Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. Neurosci Biobehav Rev 36:1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Darwall S (2006) The second-person standpoint: morality, respect, and accountability. Harvard University Press, Cambridge

Darwall S (2013) Honor, History, and Relationship: Essays in Second-personal Ethics II. Oxford University Press, Oxford

Darwall S (2018) “Second-personal morality” and morality. Philos Psychol 31:804–816. https://doi.org/10.1080/09515089.2018.1486603

Darwin C (1872) The expression of the emotions in man and animals. University of Chicago Press, Chicago

de Waal FBM (1996) Good natured: the origins of right and wrong in humans and other animals. Harvard University Press, Cambridge

de Waal FBM (2008) Putting the altruism back into altruism: the evolution of empathy. Annu Rev Psychol 59:279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625

de Waal FBM (2014) Natural normativity: The ‘is’ and ‘ought’ of animal behavior. Behaviour 151:185–204. https://doi.org/10.1163/1568539x-00003146

Dill B, Darwall S (2014) Moral psychology as accountability. In: D'Arms J, Jacobson D (eds) Moral psychology and human agency: philosophical essays on the science of ethics. Oxford University Press, Oxford, pp 40–83

Engelmann JM, Tomasello M (2017) Prosociality and morality in children and chimpanzees. In: Helwig C (ed) New perspectives on moral development. Routledge, New York, pp 55–72

Engelmann JM, Clift JB, Herrmann E, Tomasello M (2017) Social disappointment explains chimpanzees' behaviour in the inequity aversion task. Proc R Soc B Biol Sci 284. https://doi.org/10.1098/rspb.2017.1502

Engelmann JM, Tomasello M (2019) Children's sense of fairness as equal respect. Trends Cogn Sci 1–10. https://doi.org/10.1016/j.tics.2019.03.001

Fehr E, Fischbacher U (2004) Third-party punishment and social norms. Evol Hum Behav 25:63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fiske AP, Rai TS (2014) Virtuous violence: hurting and killing to create, sustain, end, and honor social relationships. Cambridge University Press, Cambridge

Frank RH (1988) Passions within reason: the strategic role of the emotions. Norton, New York

Gaus GF (2015) Social philosophy. Routledge, New York

Gauthier D (1986) Morals by agreement. Oxford University Press, Oxford

Gavrilets S, Richerson PJ (2017) Collective action and the evolution of social norm internalization. Proc Natl Acad Sci USA 114:6068–6073. https://doi.org/10.1073/pnas.1703857114

Gibbard A (1982) Human evolution and the sense of justice. Midwest Stud Philos 7:31–46. https://doi.org/10.1111/j.1475-4975.1982.tb00082.x

Gibbard A (1989) Communities of judgment. Soc Philos Policy 7:175–189. https://doi.org/10.1017/S0265052500001072

Gintis H (2003) The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms. J Theor Biol 220:407–418. https://doi.org/10.1006/jtbi.2003.3104

Gintis H, Fehr E (2012) The social structure of cooperation and punishment. Behav Brain Sci 35:28–29

Gomila A (2001) La perspectiva de segunda persona: mecanismos mentales de la intersubjetividad. Contrastes 6:65–86. https://doi.org/10.24310/Contrastescontrastes.v0i0.1448

Gomila A (2002) La perspectiva de segunda persona de la atribución mental. Azafea Rev Filos 1:123–138

Gomila A (2008) La relevancia moral de la perspectiva de segunda persona. In: Pérez D, Fernández, L. (eds) Cuestiones filosóficas: ensayos en honor de Eduardo Rabossi. Catálogos, Buenos Aires, pp 493–510

Gomila A (2012) A naturalistic defense of "human only" moral subjects. Dilemata 69–73

Gomila A (2015) Emociones en segunda persona. X Boletín Estudios de Filosofía y Cultura Manuel Mindán 10:37–50

Greene J (2013) Moral tribes: emotion, reason, and the gap between us and them. Penguin Press, New York

Hamilton WD (1964) The genetical evolution of social behavior. J Theor Biol 7:17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Handfield T, Thrasher J (2019) Two of a kind: are norms of honor a species of morality? Biol Philos 34:1–21. https://doi.org/10.1007/s10539-019-9693-z

Isern-Mas C, Gomila A (2018) Externalization is common to all value judgments, and norms are motivating because of their intersubjective grounding. Behav Brain Sci 41. https://doi.org/10.1017/S0140525X18000092

Ketelaar T, Tung Au W (2003) The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: an effect-as-information interpretation of the role of emotion in social interaction. Cogn Emot 17:429–453. https://doi.org/10.1080/02699930143000662

Leung AKY, Cohen D (2011) Within- and between-culture variation: individual differences and the cultural logics of honor, face, and dignity cultures. J Pers Soc Psychol 100:507–526. https://doi.org/10.1037/a0022151

Lewis M (1992) The socialization of shame: from parent to child. Shame: the exposed self. Free Press, New York, pp 98–118

Lewis M (1995) Embarrassment: the emotion of self-exposure and evaluation. In: Tangney JP, Fischer KW (eds) Self-conscious emotions: the psychology of shame, guilt, embarrassment, and pride. Guilford Press, New York, pp 198–218

Lewis M, Ramsay D (2002) Cortisol response to embarrassment and shame. Child Dev 73:1034–1045. https://doi.org/10.1111/1467-8624.00455

Mikhail J (2011) Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment. Cambridge University Press, New York

Pierce J, Bekoff M (2012) Wild justice redux: what we know about social justice in animals and why it matters. Soc Justice Res 25:122–139. https://doi.org/10.1007/s11211-012-0154-y

Prinz J (2011) Is empathy necessary for morality? In: Coplan A, Goldie P (eds) Empathy: philosophical and psychological perspectives. Oxford University Press, New York, pp 211–229

Rosati CS (2016) Moral motivation. In: Zalta EN (ed) Stanford Encyclopedia of Philosophy (winter 2016 edn). https://plato.stanford.edu/archives/win2016/entries/moral-motivation/

Roskies AL (2003) Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy" [1]. Philos Psychol 16:51–66. https://doi.org/10.1080/0951508032000067743

Rowlands M (2012) ¿Pueden los animales ser morales? Dilemata 9:1–32

Sarkissian H, Park J, Tien D, Wright JC, Knobe J (2011) Folk moral relativism. Mind Lang 26:482–505. https://doi.org/10.1111/j.1468-0017.2011.01428.x

Scanlon TM (2008) Moral dimensions. In: Scanlon TM (ed) Moral dimensions: permissibility, meaning, blame. Belknap Press of Harvard University Press, Cambridge, pp 122–216

Sie M (2014) Self-knowledge and the minimal conditions of responsibility: a traffic-participation view on human (moral) agency. J Value Inq 48:271–291. https://doi.org/10.1007/s10790-014-9424-2

Smith A (1759) The theory of moral sentiments. MetaLibri, São Paulo

Smith M (1994) The moral problem. Blackwell Publishing, Oxford

Sober E, Wilson DS (1998) Unto others: the evolution and psychology of unselfish behavior. Harvard University Press, Cambridge

Stanford PK (2018) The difference between ice cream and Nazis: moral externalization and the evolution of human cooperation. Behav Brain Sci 41:1–57. https://doi.org/10.1017/S0140525X17001911

Strawson PF (1974) Freedom and resentment. In: Strawson PF (ed) Freedom and resentment and other essays. Routledge, Abingdon, pp 1–28

Sznycer D, Tooby J, Cosmides L, Porat R, Shalvi S, Halperin E (2016) Shame closely tracks the threat of devaluation by others, even across cultures. Proc Natl Acad Sci USA 113:2625–2630. https://doi.org/10.1073/pnas.1514699113

Sznycer D, Xygalatas D, Agey E, Alami S, An X-F, Ananyeva KI et al (2018) Cross-cultural invariances in the architecture of shame. Proc Natl Acad Sci USA 115:9702–9707. https://doi.org/10.1073/pnas.1805016115

Tangney JP (1995) Shame and guilt in interpersonal relationships. In: Tangney JP, Fischer KW (eds) Self-conscious emotions: the psychology of shame, guilt, embarrassment, and pride. Guilford Press, New York, pp 114–139

Thrasher J, Handfield T (2018) Honor and violence. Hum Nat 29:378–389. https://doi.org/10.2307/4010934

Tomasello M (2016) A natural history of human morality. Harvard University Press, Cambridge

Tomasello M (2018) Response to commentators. Philos Psychol 31:817–829. https://doi.org/10.1080/09515089.2018.1486604

Tomasello M (2019) The moral psychology of obligation. Behav Brain Sci 43:1–58. https://doi.org/10.1017/S0140525X19001742

Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–57. https://doi.org/10.1086/406755

Turiel E (1983) The development of social knowledge: morality and convention. Cambridge University Press, New York

Vincent S, Ring R, Andrews K (2019) Normative practices of other animals. In: Zimmerman A, Jones K, Timmons M (eds) The Routlege handbook of moral epistemology. Routledge, New York, pp 57–84

Wilkinson GS (1990) Food sharing in vampire bats. Sci Am 262:76–83

Wolf S (2012) "One thought too many": love, morality, and the ordering of commitment. In: Heuer U, Lang G (eds) Luck, value, and commitment. Themes from the ethics of Bernard Williams. Oxford University Press, Oxford, pp 71–92

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.