How inferred motives shape moral judgements

Ryan W. Carlson[1,3]✉, *Yochanan E. Bigman*[1,3], *Kurt Gray*[2], *Melissa J. Ferguson*[1] *and M. J. Crockett*[1]✉

Abstract | When people judge acts of kindness or cruelty, they often look beyond the act itself to infer the agent's motives. These inferences, in turn, can powerfully influence moral judgements. The mere possibility of self-interested motives can taint otherwise helpful acts, whereas morally principled motives can exonerate those behind harmful acts. In this Review, we survey research showcasing the importance of inferred motives for moral judgements, and show how motive inferences are connected to judgements of actions, intentions and character. This work suggests that the inferences observers draw about peoples' motives are sufficient for moral judgement (they drive character judgements even without actions) and functional (they effectively aid observers in predicting peoples' future behaviour). Research that directly probes when and how people infer motives, and how motive properties guide those inferences, can deepen our understanding of the role of inferred motives in moral life.

When people engage in moral or immoral actions, there is one question observers often ask: why? Both everyday conversations about morality and legal proceedings often revolve around people's motives[1–3] — the psychological forces that guide their actions. One key reason for this is that motives often clarify what a person's actions say about their character. People usually condemn violence, but few would condemn a mother for punching a paedophile if she was trying to protect her child. People often praise human rights initiatives, yet many condemn corporate campaigns that support LGBTQ rights during Pride Month, dismissing such efforts as motivated by profit-seeking[4].

The idea that people's motives matter in social life should not surprise any psychologist. A wealth of psychology research has found that people frequently and spontaneously infer others' motives[5–9], even from a young age[10–13]. The ability to infer people's motives is an important social skill, as motives reveal people's character and predict their downstream actions. When evaluating moral or immoral actions, motive inferences are especially vital. For instance, people with a genuine motive to help others in one situation are likely to want to help others in other situations[14], and people who want to harm others on a whim in one situation are likely to do so again in the future[15]. By contrast, people who want to help others only when it benefits themselves, or who would harm others only to protect someone they care about, should be less likely to act the same across settings. Thus, accurately inferring others' motives can be extremely useful, if not vital, for discerning what an action says about a person, and in determining whom to trust and befriend, and whom to avoid.

Given the real-world importance of human motives and their central role in other areas of psychology[16–20], understanding the impact of inferred motives on moral judgement is an important domain for moral psychology. Although a growing literature shows that inferences about motives powerfully shape moral judgements[21–26], moral psychology has traditionally focused more on judgements of actions (whether an action was morally right or wrong[27–29]) and moral character (whether a person is good or bad[30–33]).

In this Review, we highlight research showcasing the many influences that inferred motives can have on moral judgement, and how motive inferences play a functional part in predicting others' future behaviour. First, we briefly review research on moral judgements of actions and character. Next, given past ambiguity surrounding these concepts, we situate motives in relation to actions and character, and other key targets of moral cognition, including intentions and outcomes (see TABLE 1). Then, we review existing research on the role of inferred motives in moral judgement, and how motive properties serve as inputs to these inferences. Finally, we consider how motive and action multiplicity complicate these judgements, and outline directions for future research.

The morality of actions and character

Most research in moral psychology focuses on either judgements of actions or judgements of character. We briefly review these research traditions to situate judgements of motives within these literatures.

[1]*Department of Psychology, Yale University, New Haven, CT, USA.*

[2]*Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.*

[3]*These authors contributed equally: Ryan W. Carlson, Yochanan E. Bigman*

✉*e-mail: ryan.carlson@yale.edu; mj.crockett@yale.edu*

https://doi.org/10.1038/s44159-022-00071-x

Table 1 | **Five key targets of moral cognition**

Target	Content	Related concepts	Relevant question	Example
Actions	Overt sequences of movements an agent makes[165,166]	Behaviour, conduct, acts	What did they do?	Ace kicked a goose
Outcomes	Results of an agent's action[29,40]	Consequences, benefits, side-effects	What were the consequences?	The goose is injured
Intentions	Action plans based on an agent's beliefs about how those actions will serve their motives[89,90]	Plans, purposes, strategies, decisions	Was it deliberate?	Ace planned to kick the goose
Motives	Outcomes (or states) that an agent is actively attracted to[16,72]	Goals, desires, needs	What did they want?	Ace wanted to harm the goose
Character	Stable qualities of an agent that summarize and predict their actions[32,49]	Dispositions, traits, values, personality	What kind of person are they?	Ace is a cruel person

Moral judgements of actions. Two major philosophical schools have influenced research on moral judgements of actions. According to consequentialist theories, the morality of an action depends solely on its consequences[34]. By contrast, deontological theories hold that the morality of actions depend on whether actions themselves are right or wrong, irrespective of their consequences[35]. Philosophers ask the normative question of what principles should guide human moral life. Psychologists instead focus on describing how philosophical views (such as deontology and consequentialism) connect to the moral judgements that ordinary people make about actions[28,36].

Robust evidence supports the idea that people are, in some respects, naive consequentialists, such that they show concern for the consequences of actions. For instance, the more harm an action causes, the more negative people typically judge the action to be[27,37–40]. Indeed, evidence from a range of experimental paradigms (some shown in FIG. 1) suggest that people's moral judgements are sensitive to how much an action negatively (or postively) influences the welfare of others[27,29,39,41–45].

However, there is also evidence that people are naive deontologists, such that they judge some actions as immoral regardless of their outcomes. For example, in the classic trolley problem an agent must decide whether to kill one person to save five people on the tracks of a runaway trolley[46,47] (FIG. 1). Researchers find that people deem the act of killing one person to save five to be considerably worse when it involves physically pushing someone off a bridge to halt the trolley versus flipping a switch to divert the trolley to a different track where it will run over someone[28,48].This finding suggests that some actions, especially those involving direct physical harm, are judged to be worse than others, even when the outcome is the same.

Moral judgements of character. Another prominent tradition in moral psychology focuses on people's moral character (or lack thereof). Moral character reflects the morally relevant dimensions of a person's personality, such as their tendency to be empathic, trustworthy, loyal, modest and equitable[32,49,50]. Moral character is conceptually related to, yet distinct from, concepts in person perception such as warmth[51] and communion[52], which include traits that are not morally relevant (such as sociability and expressiveness)[53]. People use inferences about others' moral character to predict their future behaviour across a range of morally relevant settings[54]. For instance, if we judge a person to be kind, honest or loyal, we might predict that this person will support us in times of need, engage with us authentically and have our back in a conflict[32,55,56].

In contrast to the action-based tradition, which aligns with deontological and consequentialist moral theories, research on judgements of moral character aligns with the normative theory of virtue ethics, which focuses on the agent's moral character as a core basis for moral judgement[57]. According to virtue ethicists, morality is fundamentally about how people tend to act across situations (their character), rather than individual acts of kindness or cruelty. In line with this normative view, person-centred accounts of moral judgement suggest that people are motivated to evaluate people's moral character, more so than their actions[32]. Evidence of 'act–person dissociations' (cases where the morality of actions and character disconnect) highlight the uniqueness of moral character as a target of moral judgement. For instance, observers judge the act of privately uttering a racial slur as less blameworthy than physical assault; however, they perceive the use of a slur as a stronger signal of poor moral character than physically assaulting someone[58]. Moreover, in sacrificial dilemmas, observers judge the act of throwing a dying man overboard to keep a lifeboat full of people from sinking to be moral, but not indicative of positive moral character[59].

Past work highlights two key factors that shape how an agent's actions influence judgements of their moral character: diagnosticity[31] and intentionality[60,61]. Both diagnosticity and intentionality stem from classic work on attribution[62–64] and impression formation[65]. Diagnosticity refers to how informative a specific action is in revealing an agent's character[66–69]. Such work shows that some actions are more diagnostic than others. For instance, a person who physically harms their partner's cat is judged as having a worse moral character than a person who harms their partner[31] (FIG. 1). Even though harming one's partner is judged to be a worse action, animal cruelty is rarer and therefore more diagnostic of negative traits (such as blunted empathy and antisocial tendencies) unique to that person, driving harsher character judgements.

Another key factor for how an agent's actions reflect their moral character is the inferred intentionality of their actions. For instance, observers judge intentional actions as worse than unintentional actions, even when the outcomes of those actions are identical[70], suggesting that moral judgements can change based on aspects of the person (the agent's inferred mental state), independently of any action and outcome. However, inferred intentions are not the only, or perhaps even the most important, mental state that reflects one's moral character[71]. Indeed, a growing body of research suggests that inferred motives are crucial for moral judgements of character, and that they shape moral character judgements, even when an agent's intentions are held constant.

Motives and moral cognition

Motives are relatively less well understood in moral psychology than actions or character, but emerging work reveals their numerous influences on moral judgement. In this section we examine these influences, and situate them relative to actions, intentions, outcomes and character.

Motives and related constructs. Motives refer to psychological states that direct an individual towards an end that they actively want to obtain, be it a state (such as pleasure or avoiding pain) or an outcome (such as the fair distribution of a resource)[16,62,72–74]. Thus, we treat motives as dynamic: motives change from moment to moment, and reflect an interaction between the person and the situation[75]. According to this view, motives encapsulate concepts such as desires[8,40,76–80] and goals[81–86], which are more frequently used in different subdisciplines of psychology. Moreover, under this definition, motives are related to (but distinct from) reasons that explain agents' actions[87,88]. More specifically, reasons can refer to either the motives (for instance, shooting someone out of revenge) or beliefs (for instance, not knowing the gun was loaded) that guide an agent's decision to act. Next, we focus on distinguishing motives from other key targets of moral cognition: actions, outcomes, intentions and character.

One key distinction between actions, outcomes, intentions, motives and character (TABLE 1) is whether they refer to external states of the world that observers can see, or internal mental states that observers can only infer (or learn about indirectly). Imagine an agent, Ace, kicks a goose. Although people can directly observe actions and their outcomes (Ace kicks the goose; the goose gets injured), people cannot see the character of others (whether Ace is a cruel person). Thus, observers typically make inferences about agents' underlying character. To do so, observers also need to consider the agent's intentions (whether Ace acted deliberately) and motives (did Ace want to kick the goose out of sadism or self-defence).

Although an agent's motives are connected to their specific intentions and their broader character, these targets of moral cognition are also meaningfully distinct. Intentions are typically viewed as directed towards specific actions, whereas motives are not[89]. Intentions arise from a combination of an agent's motives (Ace's motive to harm the goose) and beliefs about how a specific action and its outcomes will serve their motives (Ace's belief that kicking the goose will achieve this end)[90,91]. Intentions are therefore psychologically closer to action

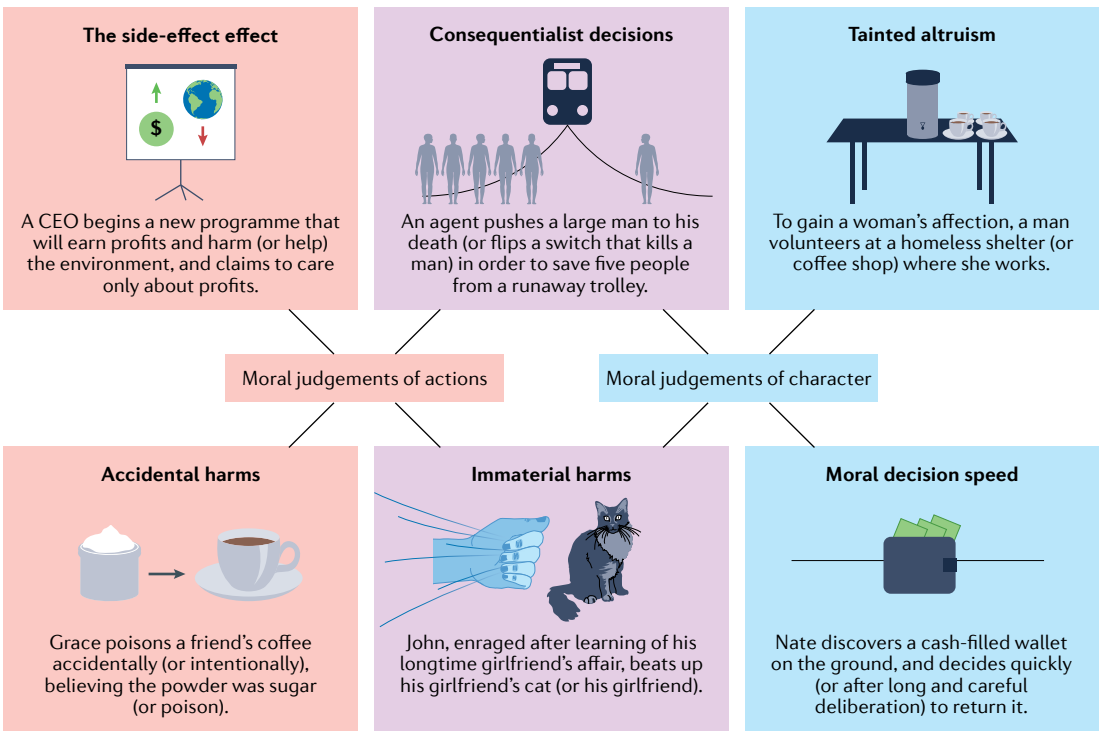

Fig. 1 | **Typical scenarios in moral psychology for judging actions and character.** A selection of common scenarios used in moral psychology research to assess moral judgements of actions and character[28,31,115,145,167,168].

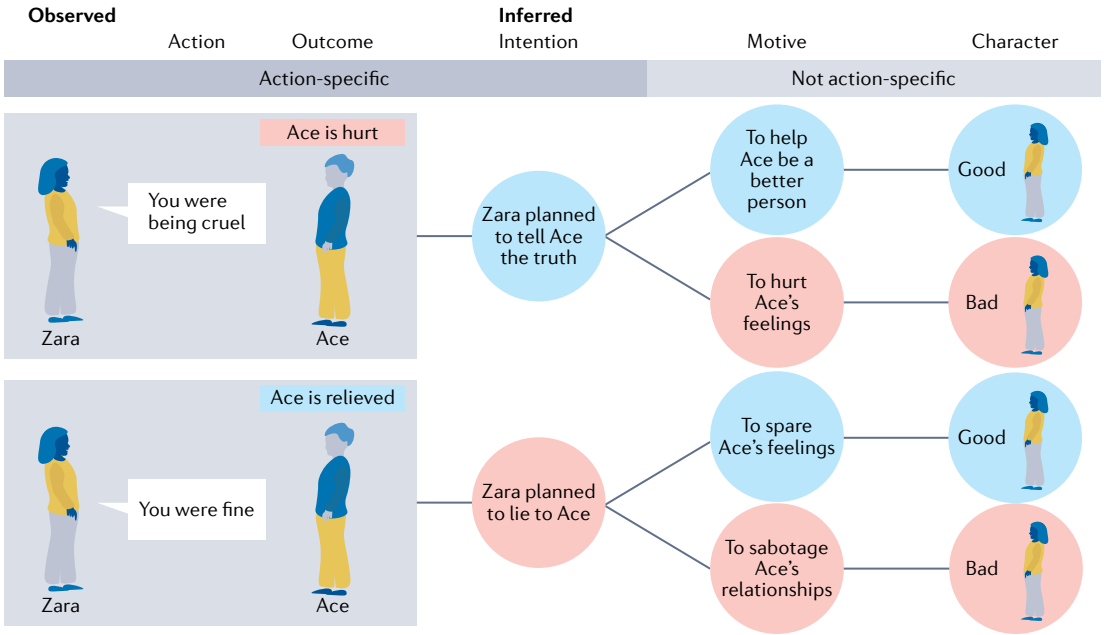

Fig. 2 | **Key targets of moral judgement.** Observers witness actions (Zara lying or being honest with Ace) and outcomes (Ace's reaction). Observers then make several inferences. Observers typically infer that actions are intentional, and proceed to make inferences about the agent's motives. Crucially, whether observers infer that Zara has a positive or negative moral character depends on the motive inferences they make (positive or negative). Thus, motive inferences connect moral judgements of actions and characters.

than motives. To infer motives from observed actions, observers must first infer that the action was intentional[87]. However, observers can also infer an agent's motives purely from what they know about the agent's character (if Ace is a cruel person, he would probably want to harm an animal). In such cases it is not necessary to infer the agent's intentions first.

An agent's character typically refers to more general, recurring patterns in their motives and actions (whether Ace typically wants to harm, or acts harmfully towards, animals), and how these patterns adhere to (or violate) moral norms. Importantly, an agent's motives can reflect their traits and broader character (Ace wanted to kick the goose because he's a sadistic person), but they could also reflect the situation (Ace harmed the goose only because he felt the need to defend himself). In this way, motive inferences are a key diagnostic tool: they clarify for observers whether an agent's actions reflect their true character, allowing them to better predict when an agent's actions should reflect a recurring pattern (whether Ace will harm others).

Moral judgements of motives and their effect on judgements of character. Motives usually drive peoples' actions, which can make it difficult for researchers to assess the unique influence of motives on moral judgements. However, a growing body of research suggests that motives matter for moral judgement, even when they do not produce any tangible actions or outcomes[23,26,73,77,92,93]. Several lines of evidence support this idea. First, research shows that motives can be judged as morally wrong in and of themselves, absent any action or outcome. For instance, people deem it morally wrong to merely possess a desire to harm others, even if no harm ultimately results from this desire[40]. Second, motives alone can drive judgements of moral character. For instance, people condemn the moral character of agents who want to profit from harm inflicted on others, even if the events they wish for are uncontrollable, such as a fund manager wishing for a natural disaster in order to profit on an investment[94]. Additionally, some theorize that motives, not actions, are a core basis by which we judge people to be selfish or not[77]. Finally, even motives directed toward one's own motives (meta-motives) can shape moral judgements of character, independent of actions. For instance, agents who have an impulsive motive (drug craving) but wish they did not have this impulse (wish they were not addicted to drugs) are seen as more moral than those who possess only an impulsive motive[95]. Overall, these findings suggest that motives constitute a distinct target of moral judgement, and inferred motives can shape people's judgements of moral character, even in the absence of any tangible action.

Motives also play an important part in connecting moral evaluations of an agent's actions to judgements of their character[71]. Imagine a scenario in which two moral values (honesty and preventing harm) clash: Ace's friend, Zara, is deciding whether to give honest, but hurtful, feedback to Ace (telling him he was cruel; FIG. 2; top panel), or to give dishonest feedback that will spare his feelings (telling him that what he did was fine; FIG. 2; bottom panel). In the event of either action or outcome (harmful honest feedback or harmless dishonest feedback), observers will probably (and in many cases spontaneously) infer Zara's mental states. If observers infer that Zara's action was intentional (as observers usually do with most actions[62,96–98]) they will probably draw inferences about Zara's motives.

An observer's inference about Zara's motives (why Zara called Ace cruel) can determine how they connect Zara's action to her moral character. Zara's intention to give honest feedback could have arisen from many possible motives, including some that are helpful (wanting to help Ace be a better person), and some that are harmful (wanting to hurt Ace's feelings). The same is true if Zara decides to lie. Zara might intentionally lie to Ace based on a helpful motive (wanting to spare Ace's feelings) or a harmful one (wanting to sabotage Ace's relationships). Although Zara probably has helpful motives if Zara is Ace's friend, this inference quickly becomes less certain if Zara is Ace's distant colleague, boss or nemesis. This example illustrates that for the same intentional action, the inferred motives (for instance, to help or to harm) can determine whether the action reflects positively or negatively on an agent's moral character.

Indeed, research confirms that people's motives often vary for the same moral or immoral action. Although people often help others based on other-oriented or morally principled motives (empathy or fairness)[16], they also often help out of self-interest[99]. For instance, people help because they want to avoid feeling guilty[100] or to signal their virtue to others[101–103]. Moreover, although people harm, punish and lie to others on the basis of self-interested motives[104–106], they also frequently engage in such acts on the basis of other-oriented or morally principled motives[107–109]. For instance, people tell white lies to help others[110], and are motivated by fairness to act punitively towards moral transgressors[108]. Because motives can vary for the same action, inferred motives are important for connecting actions to character judgements[71].

Finally, when the moral valence of actions and motive inferences are at odds, motive inferences can reverse — and even dominate — the expected relationship between actions and character judgements[21,102,109,111–115]. When people engage in harmful actions with other-oriented motives, observers judge them more positively. For instance, lying is often morally condemned, and viewed as reflecting poor moral character[116]. However, if observers infer that an agent's lie was driven by a prosocial motive ('prosocial lying'), they judge the agent positively[109,117]. Similarly, a person who physically harms someone out of self-defence is judged more positively than a person who physically harms someone to benefit themselves[118].

Observers are also highly sensitive to the motives behind prosocial actions. When helpful actions are driven strictly by emotion or empathic concern, observers tend to praise such actions and the moral character of those carrying them out[21,24,119]. However, people will readily derogate prosocial acts that seem motivated by self-interest ('tainted altruism'). For instance, prosocial agents who personally benefit from helping others tend to be judged more negatively than agents who help without benefiting (FIG. 1). Moreover, when prosocial actions are known to be motivated by personal benefits (such as seeking praise), observers judge agents as morally worse than agents who did not act helpfully at all[22,24,25,115,120].

Together, these findings illustrate how inferred motives underlying helpful and harmful actions can powerfully shape judgements of moral character, sometimes even overriding the typical relationship between actions and moral character.

Motive properties and moral judgements

The research reviewed above indicates the importance of motives for moral judgement. We now examine how different properties of motives (strength, direction and conflict) provide cues for motive inferences, and shape subsequent character judgements. To this end, we synthesize research on motivation[18,83,121,122] and models of social inference[79,123] to show how researchers can systematically integrate the influence of motives into the study of moral judgement.

Motive direction and strength. Observers perceive the motives of others on two key dimensions: direction and strength. Motive direction refers to the state or outcome the agent is striving to achieve. For instance, the direction of an agent's motive for not littering might be to benefit themselves, to benefit another person, to benefit their group or to uphold a moral norm[124]. By contrast, motive strength refers to the relative importance of that state or outcome to the agent at the moment[62].

Observers often infer the direction of an agent's motives through the outcomes the agent caused (or could cause) through their action[79,125], and the emotions they display in response to whether those outcomes are realized or not[16,126]. For instance, observers come to suspect self-interested motives when the outcomes of an agent's prosociality are self-oriented (such as receiving praise or a tax break)[24]. Moreover, observers can readily infer whether or not an agent desired their coworker to die on the basis of the agent's emotional expression (happy or sad) in response to news of his death in a plane crash[127].

Observers infer the strength of others' motives through the effort they exert in performing intentional actions, and the costs they are willing to incur doing so. Children[79,128] and adults[6,7] attribute stronger motives to agents that show evidence of exerting more (versus less) effort through their actions. Moreover, when agents incur greater costs to help others, people infer that the agents are more prosocially motivated[79,129].

Crucially, these observable cues to motive direction and strength (effort, costs, outcomes and emotions) can be inputs to moral judgements. Observers are more likely to infer that the agent's motives were self-interested, and to judge their character more negatively, when prosocial actions yield positive material or reputational outcomes for agents than when the aganets obtain no personal benefit from their actions[24,130]. Similarly, when a harmful action (throwing an injured man off a lifeboat to save others) benefits the group (consequentialism), but also the agent (saving themselves), observers judge the agent to have worse moral character than if the agent receives no self-benefits, owing to the suspicion of self-interested motives[59]. Emotions experienced in response to morally relevant actions also shape moral character judgements. For instance, the more warm, positive feelings donors report experiencing after giving to a charity, the more observers infer that their

donations were authentically motivated, which leads to more positive evaluations of the donor's moral character[21].

Inferred motive direction and strength also influence downstream moral judgements of character. For example, when helping requires a lot (versus a little) effort from an agent, observers infer that the motive to help is stronger and therefore judge the agent more positively[23,79]. Similarly, when an agent exerts a lot of (versus a little) effort in a harmful act such as stealing, observers infer that the agent has a strong motive to steal, and therefore judge the agent more negatively[23]. Moreover, observers judge prosocial agents who incur a higher (versus lower) personal cost to helping others to have greater moral character[131,132]. However, the reputational benefits of generosity might asymptote with increasing degrees of costliness. For instance, the reputational boost from donating $80,000 rather than $70,000 is much smaller than the boost from donating $10,000 rather than nothing[133].

These findings are broadly consistent with the theory of dyadic morality[33], which suggests that people evaluate the morality of others' character along the dimensions of valence (whether the agent's motives are generally good or bad) and strength (the power of the agent's motives to translate into motive-congruent acts). In this view, those evaluated as heroes and villains — the most extreme judgements of good and evil character — are seen to have both the strongest motives for good or evil, and to be able to act on their desire to help or harm others regardless of situational constraints. By contrast, more everyday good- and evil-doers are perceived to have less extreme valence and strength — they are seen to be both less motivated to help and harm others, and to act less upon those motives, only doing so when given the appropriate opportunity. For example, a serial killer (villain) wants to kill many other people and develops plans to fulfill this goal no matter the circumstances. By contrast, a person who sees a package on someone's doorstep and impulsively grabs it (more everyday evil-doer) does not have a strong motive to harm others (they are motivated by greed and not cruelty) and steals only if an easy situation presents itself.

Finally, people's expectations about others' motives are themselves an important factor in moral judgement. For instance, observers place greater trust in agents who make deontological decisions as opposed to consequentialist decisions[113,134]. Expectations about the strength and direction of motives can explain this finding. Specifically, consequentialists might be judged more negatively because people expect others' actions to reflect respect for people (namely, not to treat them as a means to an end). Consequentialist decisions (for instance, pushing a man to his death to save five others) often signal a motivational calculus that violates this expectation[113].

Cases of negligence, where people are held responsible for and morally condemned for harm they could have prevented, also highlight how perceived motive strength and expectations shape moral judgement. By one account, moral condemnation of negligence is based on the amount of harm caused, such that the more harmful the agent's negligence is, the more observers condemn the agent[135]. However, it is also possible that people who exhibit negligence are morally condemned because they are viewed as having little motivation to prevent harm to others, or perhaps even possessing some degree of motivation to cause harm. The latter possibility converges with work suggesting that inferred motives are important for judging the negative side-effects of actions[136]. This research finds that people perceive agents who purport 'not to care' about negative side-effects of their actions — in this case, harming the environment (FIG. 1) — to actually possess a moderate desire to cause the negative outcome[136]. In other words, observers think that it takes active motivation to harm the environment for someone to claim that they do not care about it at all. This finding suggests that people might evaluate how strong an agent's motives ought to be given the situation, and judge an agent's character by comparing their inference of actual motive strength against expected motive strength.

Motive conflicts. In many situations (and especially in moral dilemmas), people might have multiple motives pulling them towards different actions[62,72,83,121,137]. For instance, Ace might simultaneously wish to comfort and hurt Zara during a harsh verbal disagreement. Because these motives pull Ace towards different actions (helping and harming Zara), they constitute a motive conflict. Such conflicts are often resolved on the basis of the relative strength of each motive, with the stronger motive ultimately driving the action[62,138]. However, when conflicting motives are of similar strength, agents tend to experience tension and engage in active deliberation about which motive to pursue[139], all of which can be discerned by an astute observer.

Importantly, moral evaluations of character are sensitive to the perception of motive conflicts[140]. One well documented cue that observers use to detect motive conflicts is decision-making speed[141–144]. For instance, agents who make slow decisions to help others and quick decisions to harm others are judged more negatively than agents who are quick to help and slow to harm[93,145,146]. Because fast decisions reflect a lack of conflict, quick decisions to harm signal that the agent has a much stronger desire to harm others (or to benefit from harming them) than to avoid harming others. One caveat to this finding is that in some cases observers assign moral credit for overcoming motives via self-control[147], for instance when an agent overcomes the desire to lie about being responsible for breaking a lamp. However, self-control over motives appears to be less praiseworthy than simply having pure motives when the motive requiring self-control is itself is deemed immoral, such as being motivated to act harmfully[140].

Another important cue to motive strength is whether an agent seems to consider their self-interest when pursuing a moral action. For example, people who cooperate without choosing to look at the potential costs and benefits of doing so are viewed as more trustworthy social partners than those who choose to view the costs and benefits of cooperating[148]. In this case, 'not looking' is consistent with a lack of conflict between motives,

specifically, that one's concern for others is sufficiently strong to be unhindered by self-interested motives.

Together, this work suggests that cues to motive conflicts can be an important input to motive inferences by revealing the relative strength of prosocial and self-oriented motives. In turn, these inferences are important for moral judgements of character, which can have functional consequences for who people select as social partners.

Motive and action multiplicity

Although motives are sometimes treated as deterministic drivers of action, the relationship between motives and actions is complicated by at least two factors: different motives can produce the same action, and different actions can serve the same motive. In this section, we discuss how these aspects of the motive–action relationship complicate motive inferences and moral judgements.

Motive multiplicity. Although some actions have one likely motive behind them (bank robberies are probably motivated by a desire to acquire money), other actions might be motivationally ambiguous. For example, Ace could help Zara move apartments on the basis of several motives depending on the dynamics of their social relationship. He might feel empathy for Zara, wish to gain her social approval, or simply feel it's fair to help her as she's helped him in the past. The same is true for motives to harm. Revenge, sadism or justice could all motivate Zara to harm Ace. Even when it's clear to observers that an action was intentional (Ace intentionally helped Zara), the agent's motives might nonetheless be ambiguous. This ambiguity is a product of 'motive multiplicity' in actions[62,72,121]: the same action can result from many possible motives[121] (FIG. 3).

From a functional perspective, motive multiplicity is a problem that observers would prefer to avoid altogether. People prefer predictability in others[149], and judge predictable agents to have better moral character than less predictable agents[30,150]. It follows that observers should prefer prosocial agents whose aid, trust and support come with fewer (and more noble) motives attached. That is, they should prefer agents whose kindness is dispositional (reflecting their stable moral character), not situational. Indeed, the importance of predictability of moral character offers an elegant account of why people place greater trust in deontologists than consequentialists in moral dilemmas[134,151,152]. People might trust deontologists more because their decisions reflect stronger motives to respect others, and reveal a less complex — and therefore more predictable — set of motives than do utilitarians[150,151].

When the outcomes of an action point to several possible motives, prior beliefs about what motives tend to drive people in general are an important guide for motive inferences. Past work suggests that people over-attribute self-interest as a motive for others' actions[153–155]. For example, when people learn about blood donors who were paid for giving blood, they underestimate how many of those donors would give blood without monetary compensation[155]. However, people also vary in their beliefs about the prevalence of certain motives, such as the extent to which they believe people act on altruistic motives[111]. These beliefs influence how much evidence they need to infer that a prosocial agent acted out of self-interest. For instance, observers are more likely to infer that a prosocial agent had self-serving motives based on the presence of self-serving outcomes if they believe that altruistic motives are rare than if they believe altruistic motives are common[111].

Action multiplicity. Motives can be realized through many possible actions in a situation[62,72,121] (FIG. 3), and this principle of 'action multiplicity' in motives also influences moral judgement. For instance, if Ace wants to improve

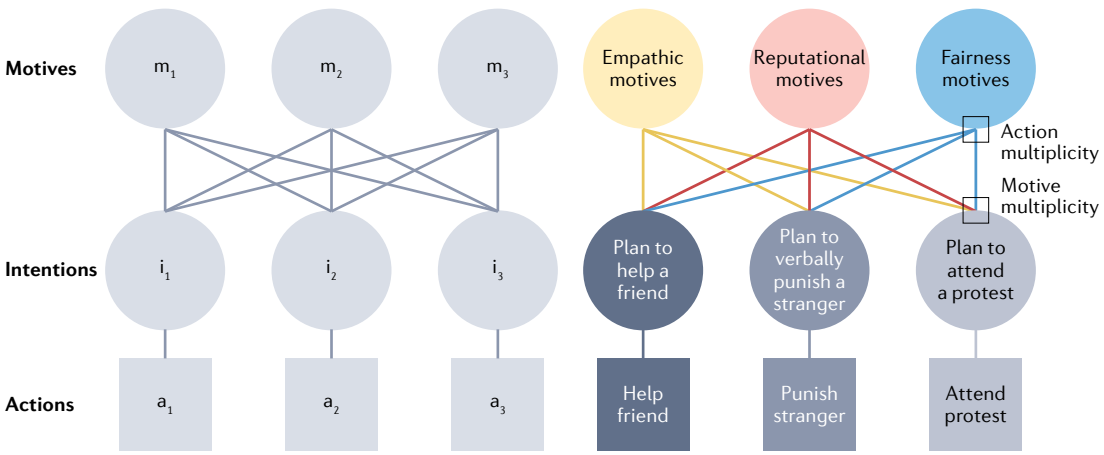

Fig. 3 | **Motive and action multiplicity.** Relationships between an agent's motives (m), intentions (i), and actions (a) are shown as a minimal framework (left), and with specific examples (right). Action multiplicity reflects how one motive (for instance, empathic concern) can map forward to many intentions. Motive multiplicity reflects how one intention (for instance, attending a protest) can map backward to many motives. Intentions do not always lead to actions (for instance, an agent might change their intentions or procrastinate), but we omit this representation here for simplicity. Social perceivers face the problem of motive multiplicity when inferring others' motives from actions, and the problem of action multiplicity when predicting future actions based on motives.

Box 1 | **Motives and emerging social challenges**

Researchers and ethicists are expressing growing concern about autonomous technologies and their rapidly increasing role in human life[169]. Robots and other artificial agents are perceived as less driven by motives than humans[170,171]. These agents are increasingly tasked with decisions that have moral implications, such as allocating scarce medical resources[172], informing parole decisions[173] and guiding autonomous vehicles[171,174]. Understanding the influence of motives in moral judgement can shed light on how the motiveless existence of artificial agents influences how people respond to the decisions of such artificial agents. On the one hand, people are averse to having artificial agents make morally relevant decisions[175,176], which can be explained by people perceiving robots as lacking helpful motives. On the other hand, people see artificial agents as less capable of discrimination[177], and are less outraged when they do discriminate[178], which can be explained by people perceiving robots as lacking harmful motives, such as prejudice[178].

Intergroup relations is another societally important domain in which progress has been made in examining both motives[19,158] and motive inferences[73,162,179]. For instance, people attribute more self-interested motives[180] and harmful motives[162] to political outgroups than to ingroups, suggesting that group membership is crucial for drawing more morally negative motive inferences of others, independent of actions. These negative motive inferences could in turn reinforce intergroup hostility and conflicts. For instance, when probed about their group's motives for being involved in the regional conflict, Israelis and Palestinians are both more likely to ascribe their own group's motives in the conflict to love of their ingroup, and their outgroup's motives to hatred of their ingroup[162]. Moreover, in the USA, Democrats and Republicans both tend to hold inaccurate perceptions of how much the other group dislikes them in competitive settings, and the greater this misperception, the more they tend to infer that their outgroup's collective actions are motivated by obstructionism[179].

Algorithms and intergroup relations come together in considering the ways in which algorithms mediate how people interact with each other in online social networks. Social networks use algorithms to maximize user engagement by amplifying content that increases time spent on their platforms[181]. Such content includes extreme messaging and displays of moral outrage[182]. These interventions distort how people interact with each other[183–185], and could in turn shift the inferences people make about the motives of outgroups. Increasing public awareness of the newsfeed manipulations of online networks, and more careful inferences about outgroup motives, could depolarize and increase intergroup tolerance[186].

Zara's wellbeing, Ace could perform several actions, such as inviting Zara on a hike or taking Zara out for dinner. The same goes for more sinister motives. If Zara wanted to impair Ace's well-being, Zara could perform several actions, such as inviting Ace to help her clean her car, or asking Ace to take out her overflowing garbage bin. This aspect of motives highlights the flexibility with which motives translate into intentional actions. It also provides a perspective on action prediction: when observers already know (or have inferred) an agent's motives, they can use this information to predict (or make assumptions about) an agent's future actions by considering different actions that would serve their motive.

Although motives can lead to many possible actions, some actions serve a given motive better than others. For instance, donating publicly is more consistent with the motive to gain reputationally than is donating privately. People seem to be well aware of how the outcomes of their actions might reflect their motives. For instance, people forgo incentives for prosocial behaviour specifically to signal that they have altruistic motives[156]. Moreover, people are sometimes more prosocial when doing so is painful and effortful, which signals more altruistic motives[157]. These findings suggest that agents consider how different action outcomes reflect their motives to observers. In turn, observers effectively use knowledge of an agent's relative motives to benefit themselves (versus others) to predict that agent's actions[129], and even leverage these predictions to make more strategic choices in social dilemmas[158].

Taken together, motive multiplicity and action multiplicity showcase the complexity of the connection between motives and actions. Motive multiplicity complicates the task of inferring an agent's motives from their actions (did he help out of self-interest, duty or altruism?), whereas action multiplicity complicates the task of predicting an agent's future actions from their inferred motives (will his self-interest motivate him to effectively support me in times of need, or lead only to half-hearted attempts at support?). Interestingly, the work reviewed suggests a game of cat and mouse: observers seeking to detect self-interested or harmful motives in agents might often encounter agents who are motivated to act in ways that strategically conceal such motives.

Summary and future directions

Moral psychology has long emphasized the importance of actions and character in moral judgements. However, observers frequently go beyond judging actions and seek to understand peoples' motives. Moral psychology paradigms often feature cues to motives which carry moral weight, such as an agent's desire to harm others physically, or their lack of motivation to prevent harm to others. The inferences people draw about others' motives are crucial for moral judgement in two respects. First, the mere presence of certain motives can drive moral judgements of character, even in the absence of any action. Second, inferred motives shape what an agent's actions reveal about their character to observers, and thereby allow observers to better predict others' future actions[158]. To integrate past work and guide future research in moral psychology, we reviewed research connecting motives with actions, character and other key constructs. These insights can enrich our understanding of moral judgement, and shed light on emerging social phenomena that are relevant to moral psychology (see BOX 1). The motive properties reviewed (motive strength, direction and conflict), as well as motive and action multiplicity, offer a guide for future work.

First, researchers should measure inferred motives when studying moral judgements. As outlined above, motives have numerous influences on moral judgements. However, knowledge about the role of motives is limited, mainly because inferences about motives often go unexamined. Future research in moral psychology should incorporate motives into theories and research by measuring the inferences observers might make about others' motives in moral situations.

Second, a central implication of the work reviewed above is that the motives people attribute to others shape their moral character judgements. However, much of this past work explicitly revealed motives to observers, rather than having observers infer motives. Consequently, this prior work does not capture the natural process by which people make inferences about motives. One important future direction is to further characterize the cognitive and computational mechanisms by which people infer others' motives in moral situations[73,79,158]. For instance,

when do people project their own motives onto others through simulation (what would my motives be in this situation?), and when do they consider others' motives through perspective taking (given what I know about this person, what would their motives be in this situation?). Another important line of inquiry is determining the extent to which people draw on different social inference mechanisms such as covariation[64] (how often an agent has acted similarly across different settings), parallel constraint satisfaction[159] (which motive fits best with the situation, prior beliefs about the agent, and/or the observer's stereotypes), social structure learning (which motives can be inferred on the basis of the agent's likely group memberships)[158,160,161], and cost–benefit analysis[79]. Further investigations will further enrich our understanding of the cognitive mechanisms observers rely on to infer motives across different situations.

Third, the work reviewed above suggests that motives can tell a different story from someone's behaviour, which can 'reverse' moral judgements of character from good to bad (or vice versa). For instance, observers can infer that an agent acted generously from self-interested motives, and acted harmfully from moral motives. A crucial question for future work is when and for whom people make such 'overriding' motive inferences. For example, group membership can be an important factor in drawing more morally negative motive inferences of others, independent of actions, and these negative motive inferences could have a key role in reinforcing intergroup hostility and conflict[162] (BOX 1). Further exploring group membership and other psychological factors that could shape the valence of motive inferences — such as social distance, distrust, suspicion and paranoia — is a vital direction for future research.

Although we suggest that inferred motives are a powerful force in moral judgement, there are certainly exceptions to this claim. For instance, it seems unlikely that any motive inference could exonerate individuals behind extreme acts of violence, such as mass shootings or genocide (in such cases, people's interest in the agent's motives may instead stem from a desire to understand why the event occurred). Moreover, some research suggests that an agent's mental states have less bearing on how they are judged in certain religious traditions[163]. For example, Jewish and Catholic participants judged a person who dislikes their parents but treats them well more positively than did Protestant participants[163]. There is also cultural variation in how people react to the outcomes of prosocial behaviour. For instance, although American participants tend to infer more self-interested motives when a prosocial agent benefits reputationally from their act (compared to when they reap no benefits)[24], Japanese participants do not infer greater self-interest from the mere presence of reputational benefits[164]. These findings suggest that the role of motive inferences in moral judgement might vary across religions and cultures. Such variations are important for future research to examine.

Psychologists have long proposed that motives powerfully shape human behaviour. Here we have applied this insight to the study of moral judgement. We described how inferred motives might reveal what an agent's actions say about their character for observers, and charted some paths forward for deeper investigations into this topic. By pursuing such avenues, future work will shed further light on the influence of motive inferences on moral life.

Published online: 08 June 2022

1. Liefgreen, A., Yousif, S. R., Keil, F. C. & Lagnado, D. A. Motive on the mind: explanatory preferences at multiple stages of the legal-investigative process. *Cognition* **217**, 104892 (2021).
2. Nadler, J. & McDonnell, M.-H. Moral character, motive, and the psychology of blame. *Cornell Rev.* **97**, 255 (2011).
3. Verstein, A. The failure of mixed-motives jurisprudence. *Univ. Chicago Law Rev.* **86**, 725–796 (2019).
4. Zheng, L. Your rainbow logo doesn't make you an ally. *Harvard Business Review* https://hbr.org/2021/06/your-rainbow-logo-doesnt-make-you-an-ally (2021).
5. Aarts, H., Gollwitzer, P. M. & Hassin, R. R. Goal contagion: perceiving is for pursuing. *J. Pers. Soc. Psychol.* **87**, 23–37 (2004).
6. Dik, G. & Aarts, H. Behavioral cues to others' motivation and goal pursuits: the perception of effort facilitates goal inference and contagion. *J. Exp. Soc. Psychol.* **43**, 727–737 (2007).
7. Hassin, R. R., Aarts, H. & Ferguson, M. J. Automatic goal inferences. *J. Exp. Soc. Psychol.* **41**, 129–140 (2005).
8. Malle, B. F. & Holbrook, J. Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *J. Pers. Soc. Psychol.* **102**, 661 (2012).
9. Moskowitz, G. B. & Olcaysoy Okten, I. Spontaneous goal inference (SGI). *Soc. Personal. Psychol. Compass* **10**, 64–80 (2016).
10. Baillargeon, R. et al. *Psychological and Sociomoral Reasoning in Infancy* (American Psychological Association, 2015).
11. Gergely, G., Nádasdy, Z., Csibra, G. & Bíró, S. Taking the intentional stance at 12 months of age. *Cognition* **56**, 165–193 (1995).
12. Hamlin, J. K., Wynn, K. & Bloom, P. Social evaluation by preverbal infants. *Nature* **450**, 557–559 (2007).
13. Liu, S., Ullman, T. D., Tenenbaum, J. B. & Spelke, E. S. Ten-month-old infants infer the value of goals from the costs of actions. *Science* **358**, 1038–1041 (2017).
14. Davis, M. H. Measuring individual differences in empathy: evidence for a multidimensional approach. *J. Pers. Soc. Psychol.* **44**, 113–126 (1983).
15. Buckels, E. E., Jones, D. N. & Paulhus, D. L. Behavioral confirmation of everyday sadism. *Psychol. Sci.* **24**, 2201–2209 (2013).
16. Batson, C. D. *Altruism in Humans* (Oxford Univ. Press, 2011).
17. Deci, E. L. & Ryan, R. M. Self-determination theory: a macrotheory of human motivation, development, and health. *Canad. Psychol.* **49**, 182 (2008).
18. Fiske, S. T., Gilbert, D. T. & Lindzey, G. *Handbook of Social Psychology* Vol. 2 (Wiley, 2010).
19. Rai, T. S. & Fiske, A. P. Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychol. Rev.* **118**, 57–75 (2011).
20. Weisz, E., Ong, D. C., Carlson, R. W. & Zaki, J. Building empathy through motivation-based interventions. *Emotion* **21**, 990–999 (2021).
21. Barasch, A., Levine, E. E., Berman, J. Z. & Small, D. A. Selfish or selfless? On the signal value of emotion in altruistic behavior. *J. Pers. Soc. Psychol.* **107**, 393–413 (2014).
22. Berman, J. Z. & Silver, I. Prosocial behavior and reputation: when does doing good lead to looking good? *Curr. Opin. Psychol.* **43**, 102–107 (2022).
23. Bigman, Y. E. & Tamir, M. The road to heaven is paved with effort: perceived effort amplifies moral judgment. *J. Exp. Psychol. Gen.* **145**, 1654–1669 (2016).
24. Carlson, R. W. & Zaki, J. Good deeds gone bad: lay theories of altruism and selfishness. *J. Exp. Soc. Psychol.* **75**, 36–40 (2018).
25. Raihani, N. J. & Power, E. A. in *Evolutionary Human Sciences* Vol. 3 (Cambridge Univ. Press, 2021).
26. Woolfolk, R. L., Doris, J. M. & Darley, J. M. Identification, situational constraint, and social cognition: studies in the attribution of moral responsibility. *Cognition* **100**, 283–301 (2006).
27. Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R. & Hütter, M. Consequences, norms, and generalized inaction in moral dilemmas: the CNI model of moral decision-making. *J. Pers. Soc. Psychol.* **113**, 343–376 (2017).
28. Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. An fMRI investigation of emotional engagement in moral judgment. *Science* **293**, 2105–2108 (2001).
29. Malle, B. F., Guglielmo, S. & Monroe, A. E. A theory of blame. *Psychol. Inq.* **25**, 147–186 (2014).
30. Crockett, M. J., Everett, J. A., Gill, M. & Siegel, J. Z. in *Advances in Experimental Social Psychology* Vol. 64, 1–64 (Elsevier, 2021).
31. Tannenbaum, D., Uhlmann, E. L. & Diermeier, D. Moral signals, public outrage, and immaterial harms. *J. Exp. Soc. Psychol.* **47**, 1249–1254 (2011).
32. Uhlmann, E. L., Pizarro, D. A. & Diermeier, D. A person-centered approach to moral judgment. *Perspect. Psychol. Sci.* **10**, 72–81 (2015).
33. Hartman, R., Blakey, W. & Gray, K. Deconstructing moral character judgments. *Curr. Opin. Psychol.* **43**, 205–212 (2022).
34. Mill, J. S. *Utilitarianism* (Cambridge Univ. Press, 1861).
35. Kant, I. *Groundwork for the Metaphysics of Morals* (Oxford Univ. Press, 1785).
36. Kahane, G. et al. Beyond sacrificial harm: a two-dimensional model of utilitarian psychology. *Psychol. Rev.* **125**, 131–164 (2018).
37. Gray, K., Schein, C. & Ward, A. F. The myth of harmless wrongs in moral cognition: automatic dyadic completion from sin to suffering. *J. Exp. Psychol. Gen.* **143**, 1600–1615 (2014).
38. Djeriouat, H. & Trémolière, B. The dark triad of personality and utilitarian moral judgment: the mediating role of honesty/humility and harm/care. *Personal. Individ. Differ.* **67**, 11–16 (2014).

39. Schein, C. & Gray, K. The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personal. Soc. Psychol. Rev.* **22**, 32–70 (2018).
40. Cushman, F. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* **108**, 353–380 (2008).
41. Cohen, D. J. & Ahn, M. A subjective utilitarian theory of moral judgment. *J. Exp. Psychol. Gen.* **145**, 1359–1381 (2016).
42. Graham, J., Haidt, J. & Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**, 1029–1046 (2009).
43. Mikhail, J. Universal moral grammar: theory, evidence and the future. *Trends Cogn. Sci.* **11**, 143–152 (2007).
44. Miller, R. M., Hannikainen, I. A. & Cushman, F. A. Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion* **14**, 573–587 (2014).
45. Yudkin, D. A., Prosser, A. M. B. & Crockett, M. J. Actions speak louder than outcomes in judgments of prosocial behavior. *Emotion* **19**, 1138–1147 (2019).
46. Foot, P. The problem of abortion and the doctrine of the double effect. *Oxf. Rev.* **5**, 5–15 (1967).
47. Thomson, J. J. Killing, letting die, and the trolley problem. *Monist* **59**, 204–217 (1976).
48. Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. & Cohen, J. D. The neural bases of cognitive conflict and control in moral judgment. *Neuron* **44**, 389–400 (2004).
49. Goodwin, G. P., Piazza, J. & Rozin, P. Moral character predominates in person perception and evaluation. *J. Pers. Soc. Psychol.* **106**, 148–168 (2014).
50. Wojciszke, B. Morality and competence in person-and self-perception. *Eur. Rev. Soc. Psychol.* **16**, 155–188 (2005).
51. Fiske, S. T., Cuddy, A. J. C. & Glick, P. Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
52. Abele, A. E. & Wojciszke, B. Agency and communion from the perspective of self versus others. *J. Pers. Soc. Psychol.* **93**, 751–763 (2007).
53. Goodwin, G. P. Moral character in person perception. *Curr. Dir. Psychol. Sci.* **24**, 38–44 (2015).
54. Brambilla, M. & Leach, C. W. On the importance of being moral: the distinctive role of morality in social judgment. *Soc. Cogn.* **32**, 397–408 (2014).
55. Buchanan, A. *Our Moral Fate: Evolution and the Escape from Tribalism* (MIT Press, 2020).
56. Enke, B. Kinship, cooperation, and the evolution of moral systems. *Q. J. Econ.* **134**, 953–1019 (2019).
57. Hursthouse, R. & Pettigrove, G. Virtue ethics. In *The Stanford Encyclopedia of Philosophy* (ed. Zalta, E. N.) (Metaphysics Research Lab, Stanford University, 2018).
58. Uhlmann, E. L., Zhu, L. & Diermeier, D. When actions speak volumes: the role of inferences about moral character in outrage over racial bigotry. *Eur. J. Soc. Psychol.* **44**, 23–29 (2014).
59. Uhlmann, E. L., Zhu, L. L. & Tannenbaum, D. When it takes a bad person to do the right thing. *Cognition* **126**, 326–334 (2013).
60. Epley, N., & Waytz, A. in *Handbook of Social Psychology* 5th edn (eds. Fiske, S. T., Gilbert, D. T. & Lindzey, G.) 498–541 (Wiley, 2010).
61. Pizarro, D. A. & Tannenbaum, D. in *The Social Psychology of Morality: Exploring the Causes of Good and Evil* 91–108 (American Psychological Association, 2012).
62. Heider, F. *The Psychology of Interpersonal Relations* (Psychology Press, 1958).
63. Kelley, H. H. in *Nebraska Symposium On Motivation* (Univ. Nebraska Press, 1967).
64. Kelley, H. H. The processes of causal attribution. *Am. Psychol.* **28**, 107–128 (1973).
65. Skowronski, J. J. & Carlston, D. E. Social judgment and social memory: the role of cue diagnosticity in negativity, positivity, and extremity biases. *J. Pers. Soc. Psychol.* **52**, 689 (1987).
66. Cone, J. & Ferguson, M. J. He did what? The role of diagnosticity in revising implicit evaluations. *J. Pers. Soc. Psychol.* **108**, 37–57 (2015).
67. Reeder, G. D. & Brewer, M. B. A schematic model of dispositional attribution in interpersonal perception. *Psychol. Rev.* **86**, 61–79 (1979).
68. Reeder, G. D., Pryor, J. B. & Wojciszke, B. in *Language, Interaction And Social Cognition* 37–57 (Sage, 1992).
69. Trafimow, D. & Trafimow, S. Mapping perfect and imperfect duties onto hierarchically and partially restrictive trait dimensions. *Pers. Soc. Psychol. Bull.* **25**, 687–697 (1999).
70. Ames, D. L. & Fiske, S. T. Intentional harms are worse, even when they're not. *Psychol. Sci.* **24**, 1755–1762 (2013).
71. Reeder, G. Mindreading: judgments about intentionality and motives in dispositional inference. *Psychol. Inq.* **20**, 1–18 (2009).
72. Lewin, K. *The Conceptual Representation and the Measurement of Psychological Forces* (Duke Univ. Press, 1938).
73. Reeder, G. D. & Trafimow, D. in *Other Minds: How Humans Bridge the Divide Between Self and Others* 106–123 (Guilford, 2005).
74. Yuill, N. & Perner, J. Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Dev. Psychol.* **24**, 358–365 (1988).
75. Lewin, K. Defining the 'field at a given time'. *Psychol. Rev.* **50**, 292–310 (1943).
76. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).
77. Carlson, R. W., Adkins, C., Crockett, M. J. & Clark, M. S. Psychological selfishness. *Perspect. Psychol. Sci.* https://doi.org/10.1177/17456916211045692 (2022).
78. Dennett, D. C. *The Intentional Stance* (MIT Press, 1987).
79. Jara-Ettinger, J., Gweon, H., Schulz, L. E. & Tenenbaum, J. B. The naïve utility calculus: computational principles underlying commonsense psychology. *Trends Cogn. Sci.* **20**, 589–604 (2016).
80. Kotabe, H. P. & Hofmann, W. On integrating the components of self-control. *Perspect. Psychol. Sci.* **10**, 618–638 (2015).
81. Berkman, E. T. & Lieberman, M. D. in *The Psychology Of Goals* 98–126 (Guilford, 2009).
82. Carlson, R. W. & Crockett, M. J. The lateral prefrontal cortex and moral goal pursuit. *Curr. Opin. Psychol.* **24**, 77–82 (2018).
83. Fishbach, A. & Ferguson, M. J. in *Social Psychology: Handbook of Basic Principles* 2nd edn (eds. Kruglanski, A. W. & Higgins, E. T.) 490–515 (Guilford, 2007).
84. Kruglanski, A. W. in *The Psychology of Action: Linking Cognition and Motivation io Behavior* 599–618 (Guilford, 1996).
85. Moskowitz, G. B. & Grant, H. *The Psychology of Goals* (Guilford, 2009).
86. O'Reilly, R. C. Unraveling the mysteries of motivation. *Trends Cogn. Sci.* **24**, 425–434 (2020).
87. Malle, B. F. in *Advances in Experimental Social Psychology* Vol. 44 (eds. Olson, J. M. & Zanna, M. P.) Ch. 6, 297–352 (Academic, 2011).
88. Korman, J. & Malle, B. F. Grasping for traits or reasons? How people grapple with puzzling social behaviors. *Pers. Soc. Psychol. Bull.* **42**, 1451–1465 (2016).
89. Malle, B. F. & Knobe, J. in *Intentions and Intentionality: Foundations of Social Cognition* 45–67 (MIT Press, 2001).
90. Bratman, M. E. *Faces of Intention: Selected Essays on Intention and Agency* (Cambridge Univ. Press, 1999).
91. Malle, B. F. & Knobe, J. The folk concept of intentionality. *J. Exp. Soc. Psychol.* **33**, 101–121 (1997).
92. Choshen-Hillel, S., Shaw, A. & Caruso, E. M. Lying to appear honest. *J. Exp. Psychol. Gen.* **149**, 1719–1735 (2020).
93. Critcher, C. R., Helzer, E. G. & Tannenbaum, D. Moral character evaluation: testing another's moral-cognitive machinery. *J. Exp. Soc. Psychol.* **87**, 103906 (2020).
94. Inbar, Y., Pizarro, D. A. & Cushman, F. Benefiting from misfortune: when harmless actions are judged to be morally blameworthy. *Pers. Soc. Psychol. Bull.* **38**, 52–62 (2012).
95. Pizarro, D., Uhlmann, E. & Salovey, P. Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychol. Sci.* **14**, 267–272 (2003).
96. Cushman, F. Deconstructing intent to reconstruct morality. *Curr. Opin. Psychol.* **6**, 97–103 (2015).
97. Baker, C. L., Saxe, R. & Tenenbaum, J. B. Action understanding as inverse planning. *Cognition* **113**, 329–349 (2009).
98. Malle, B. F. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction* (MIT Press, 2004).
99. Cialdini, R. B. Altruism or egoism? That is (still) the question. *Psychol. Inq.* **2**, 124–126 (1991).
100. Charness, G. & Dufwenberg, M. Promises and partnership. *Econometrica* **74**, 1579–1601 (2006).
101. Ariely, D., Bracha, A. & Meier, S. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am. Econ. Rev.* **99**, 544–555 (2009).
102. Kraft-Todd, G., Kleiman-Weiner, M. & Young, L. Differential discounting of virtue signaling: public virtue is perceived less favorably than private virtue for generosity but not impartiality. Preprint at *PsyArXiv* https://psyarxiv.com/zqpv7 (2020).
103. Berman, J. Z., Levine, E. E., Barasch, A. & Small, D. A. The braggart's dilemma: on the social rewards and penalties of advertising prosocial behavior. *J. Mark. Res.* **52**, 90–104 (2015).
104. Crockett, M. J., Özdemir, Y. & Fehr, E. The value of vengeance and the demand for deterrence. *J. Exp. Psychol. Gen.* **143**, 2279 (2014).
105. Shalvi, S., Gino, F., Barkan, R. & Ayal, S. Self-serving justifications: doing wrong and feeling moral. *Curr. Dir. Psychol. Sci.* **24**, 125–130 (2015).
106. Marshall, J., Yudkin, D. A. & Crockett, M. J. Children punish third parties to satisfy both consequentialist and retributive motives. *Nat. Hum. Behav.* **5**, 361–368 (2021).
107. West, S. J., Parton, D. M. & Chester, D. Harming in order to help: an empirical demonstration of prosocial aggression. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/phsve (2022).
108. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
109. Levine, E. E. & Schweitzer, M. E. Prosocial lies: when deception breeds trust. *Organ. Behav. Hum. Decis. Process.* **126**, 88–106 (2015).
110. Erat, S. & Gneezy, U. White lies. *Manag. Sci.* **58**, 723–733 (2012).
111. Carlson, R. W. & Zaki, J. Belief in altruistic motives predicts prosocial actions and inferences. *Psychol. Rep.* https://doi.org/10.1177/00332941211013529 (2021).
112. Dhaliwal, N. A., Skarlicki, D. P., Hoegg, J. & Daniels, M. A. Consequentialist motives for punishment signal trustworthiness. *J. Bus. Ethics* **176**, 451–466 (2022).
113. Everett, J. A. C., Faber, N. S., Savulescu, J. & Crockett, M. J. The costs of being consequentialist: social inference from instrumental harm and impartial beneficence. *J. Exp. Soc. Psychol.* **79**, 200–216 (2018).
114. Gorsira, M., Denkers, A. & Huisman, W. Both sides of the coin: motives for corruption among public officials and business employees. *J. Bus. Ethics* **151**, 179–194 (2018).
115. Newman, G. E. & Cain, D. M. Tainted altruism: when doing some good is evaluated as worse than doing no good at all. *Psychol. Sci.* **25**, 648–655 (2014).
116. Tyler, J. M., Feldman, R. S. & Reichert, A. The price of deceptive behavior: disliking and lying to people who lie to us. *J. Exp. Soc. Psychol.* **42**, 69–77 (2006).
117. Levine, E. E. & Schweitzer, M. E. Are liars ethical? On the tension between benevolence and honesty. *J. Exp. Soc. Psychol.* **53**, 107–117 (2014).
118. Reeder, G. D., Kumar, S., Hesson-McInnis, M. S. & Trafimow, D. Inferences about the morality of an aggressor: the role of perceived motive. *J. Pers. Soc. Psychol.* **83**, 789–803 (2002).
119. Levine, E. E., Barasch, A., Rand, D., Berman, J. Z. & Small, D. A. Signaling emotion and reason in cooperation. *J. Exp. Psychol. Gen.* **147**, 702–719 (2018).
120. Alcala, V. et al. The tainted altruism effect: a successful pre-registered replication. *R. Soc. Open. Sci.* **9**, 211152 (2022).
121. Kruglanski, A. W. et al. in *Advances In Experimental Social Psychology* Vol. 34 (ed. Zanna, M. P.) 331–378 (Academic, 2002).
122. Kruglanski, A. W., Chernikova, M., Babush, M., Dugas, M. & Schumpe, B. M. in *Advances in Motivation Science* Vol. 2 (ed. Elliot, A. J.) Ch. 3, 69–98 (Elsevier, 2015).
123. Olcaysoy Okten, I. & Moskowitz, G. B. Goal versus trait explanations: causal attributions beyond the trait-situation dichotomy. *J. Pers. Soc. Psychol.* **114**, 211–229 (2018).
124. Batson, C. D., Ahmad, N. & Tsang, J.-A. Four motives for community involvement. *J. Soc. Issues* **58**, 429–445 (2002).
125. Jones, E. E. & Davis, K. E. in *Advances in Experimental Social Psychology* Vol. 2 (ed. Berkowitz, L.) 219–266 (Academic, 1965).
126. Ong, D. C., Zaki, J. & Goodman, N. D. Affective cognition: exploring lay theories of emotion. *Cognition* **143**, 141–162 (2015).

127. Wu, Y., Baker, C. L., Tenenbaum, J. B. & Schulz, L. E. Rational inference of beliefs and desires from emotional expressions. *Cogn. Sci.* **42**, 850–884 (2018).
128. Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B. & Schulz, L. E. Children's understanding of the costs and rewards underlying rational action. *Cognition* **140**, 14–23 (2015).
129. Davis, I., Carlson, R. W., Dunham, Y. & Jara-Ettinger, J. Reasoning about social preferences with uncertain beliefs. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/au5gc (2021).
130. Lin-Healy, F. & Small, D. A. Nice guys finish last and guys in last are nice: the clash between doing well and doing good. *Soc. Psychol. Personal. Sci.* **4**, 692–698 (2013).
131. Johnson, S. Dimensions of altruism: do evaluations of prosocial behavior track social good or personal sacrifice? Preprint at *SSRN* https://doi.org/10.2139/ssrn.3277444 (2018).
132. Siegel, J. Z., Mathys, C., Rutledge, R. B. & Crockett, M. J. Beliefs about bad people are volatile. *Nat. Hum. Behav.* **2**, 750–756 (2018).
133. Klein, N. & Epley, N. The topography of generosity: asymmetric evaluations of prosocial actions. *J. Exp. Psychol. Gen.* **143**, 2366–2379 (2014).
134. Bostyn, D. H. & Roets, A. Trust, trolleys and social dilemmas: a replication study. *J. Exp. Psychol. Gen.* **146**, e1–e7 (2017).
135. Kneer, M. & Machery, E. No luck for moral luck. *Cognition* **182**, 331–348 (2019).
136. Guglielmo, S. & Malle, B. F. Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Pers. Soc. Psychol. Bull.* **36**, 1635–1647 (2010).
137. Dai, X. & Fishbach, A. When waiting to choose increases patience. *Organ. Behav. Hum. Decis. Process.* **121**, 256–266 (2013).
138. Kruglanski, A. W. et al. The rocky road from attitudes to behaviors: charting the goal systemic course of actions. *Psychol. Rev.* **122**, 598–620 (2015).
139. Luce, M. F. Choosing to avoid: coping with negatively emotion-laden consumer decisions. *J. Consum. Res.* **24**, 409–433 (1998).
140. Berman, J. Z. & Small, D. A. Discipline and desire: on the relative importance of willpower and purity in signaling virtue. *J. Exp. Soc. Psychol.* **76**, 220–230 (2018).
141. Diederich, A. Decision making under conflict: decision time as a measure of conflict strength. *Psychon. Bull. Rev.* **10**, 167–176 (2003).
142. Kleiman, T. & Hassin, R. R. Non-conscious goal conflicts. *J. Exp. Soc. Psychol.* **47**, 521–532 (2011).
143. Konovalov, A., Hu, J. & Ruff, C. C. Neurocomputational approaches to social behavior. *Curr. Opin. Psychol.* **24**, 41–47 (2018).
144. Stillman, P. E., Krajbich, I. & Ferguson, M. J. Using dynamic monitoring of choices to predict and understand risk preferences. *Proc. Natl Acad. Sci. USA* **117**, 31738–31747 (2020).
145. Critcher, C. R., Inbar, Y. & Pizarro, D. A. How quick decisions illuminate moral character. *Soc. Psychol. Personal. Sci.* **4**, 308–315 (2013).
146. Evans, A. M. & van de Calseyde, P. P. F. M. The effects of observed decision time on expectations of extremity and cooperation. *J. Exp. Soc. Psychol.* **68**, 50–59 (2017).
147. Starmans, C. & Bloom, P. When the spirit is willing, but the flesh is weak: developmental differences in judgments about inner moral conflict. *Psychol. Sci.* **27**, 1498–1506 (2016).
148. Jordan, J. J., Hoffman, M., Nowak, M. A. & Rand, D. G. Uncalculating cooperation is used to signal trustworthiness. *Proc. Natl. Acad. Sci. USA* **113**, 8658–8663 (2016).
149. Walker, A. C., Turpin, M. H., Fugelsang, J. A. & Bialek, M. Better the two devils you know, than the one you don't: predictability influences moral judgments of immoral actors. J. *Exp. Soc. Psychol.* **97**, 104220 (2021).
150. Turpin, M. H. et al. The search for predictable moral partners: predictability and moral (character) preferences. *J. Exp. Soc. Psychol.* **97**, 104196 (2021).
151. Everett, J. A. C., Pizarro, D. A. & Crockett, M. J. Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* **145**, 772–787 (2016).
152. Sacco, D. F., Brown, M., Lustgraaf, C. J. & Hugenberg, K. The adaptive utility of deontology: deontological moral decision-making fosters perceptions of trust and likeability. *Evol. Psychol. Sci.* **3**, 125–132 (2017).
153. Heath, C. On the social psychology of agency relationships: lay theories of motivation overemphasize extrinsic incentives. *Organ. Behav. Hum. Decis. Process.* **78**, 25–62 (1999).
154. Miller, D. T. The norm of self-interest. *Am. Psychol.* **54**, 1053–1060 (1999).
155. Miller, D. T. & Ratner, R. K. The disparity between the actual and assumed power of self-interest. *J. Pers. Soc. Psychol.* **74**, 53–62 (1998).
156. Kirgios, E. L., Chang, E. H., Levine, E. E., Milkman, K. L. & Kessler, J. B. Forgoing earned incentives to signal pure motives. *Proc. Natl. Acad. Sci. USA* **117**, 16891–16897 (2020).
157. Olivola, C. Y. & Shafir, E. The martyrdom effect: when pain and effort increase prosocial contributions. *J. Behav. Decis. Mak.* **26**, 91–105 (2013).
158. van Baar, J. M., Nassar, M. R., Deng, W. & FeldmanHall, O. Latent motives guide structure learning during adaptive social choice. *Nat. Hum. Behav.* **6**, 404–414 (2021).
159. Read, S. J., Vanman, E. J. & Miller, L. C. Connectionism, parallel constraint satisfaction processes, and gestalt principles: (re)introducing cognitive dynamics to social psychology. *Personal. Soc. Psychol. Rev.* **1**, 26–53 (1997).
160. Gershman, S. J. & Cikara, M. Social-structure learning. *Curr. Dir. Psychol. Sci.* **29**, 460–466 (2020).
161. Shin, Y. S. & Niv, Y. Biased evaluations emerge from inferring hidden causes. *Nat. Hum. Behav.* **5**, 1180–1189 (2021).
162. Waytz, A., Young, L. L. & Ginges, J. Motive attribution asymmetry for love vs. hate drives intractable conflict. *Proc. Natl. Acad. Sci. USA* **111**, 15687–15692 (2014).
163. Cohen, A. B. & Rozin, P. Religion and the morality of mentality. *J. Pers. Soc. Psychol.* **81**, 697–710 (2001).
164. Kawamura, Y., Sasaki, S. & Kusumi, T. Cultural similarities and differences in lay theories of altruism: replication of Carlson and Zaki (2018) in a Japanese sample. *Asian J. Soc. Psychol.* https://doi.org/10.1111/ajsp.12502 (2021).
165. Cushman, F. Action, outcome, and value: a dual-system framework for morality. *Personal. Soc. Psychol. Rev.* **17**, 273–292 (2013).
166. Jara-Ettinger, J., Schulz, L. E. & Tenenbaum, J. B. The naïve utility calculus as a unified, quantitative framework for action understanding. *Cogn. Psychol.* **123**, 101334 (2020).
167. Knobe, J. Intentional action and side effects in ordinary language. *Analysis* **63**, 190–194 (2003).
168. Young, L. & Saxe, R. Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* **47**, 2065–2072 (2009).
169. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press, 2014).
170. Gray, H. M., Gray, K. & Wegner, D. M. Dimensions of mind perception. *Science* **315**, 619 (2007).
171. Bigman, Y. E. & Gray, K. Life and death decisions of autonomous vehicles. *Nature* **579**, E1–E2 (2020).
172. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
173. Angwin, J. A., Larson, J., Kirchner, L. & Mattu, S. Machine bias. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
174. Awad, E. et al. The moral machine experiment. *Nature* **563**, 59–64 (2018).
175. Bigman, Y. E. & Gray, K. People are averse to machines making moral decisions. *Cognition* **181**, 21–34 (2018).
176. Young, A. D. & Monroe, A. E. Autonomous morals: inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *J. Exp. Soc. Psychol.* **85**, 103870 (2019).
177. Jago, A. S. & Laurin, K. Assumptions about algorithms' capacity for discrimination. *Pers. Soc. Psychol. Bull.* https://doi.org/10.1177/01461672211016187 (2021).
178. Bigman, Y. E., Gray, K., Waytz, A., Arnestad, M. & Wilson, D. Algorithmic discrimination causes less moral outrage than human discrimination. *J. Exp. Psychol. Gen.* https://doi.org/10.1037/xge0001250 (2022).
179. Lees, J. & Cikara, M. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nat. Hum. Behav.* **4**, 279–286 (2020).
180. Reeder, G. D., Pryor, J. B., Wohl, M. J. & Griswell, M. L. On attributing negative motives to others who disagree with our opinions. *Pers. Soc. Psychol. Bull.* **31**, 1498–1510 (2005).
181. Brady, W. J., Crockett, M. J. & Van Bavel, J. J. The MAD model of moral contagion: the role of motivation, attention, and design in the spread of moralized content online. *Perspect. Psychol. Sci.* **15**, 978–1010 (2020).
182. Brady, W. J., McLoughlin, K., Doan, T. N. & Crockett, M. J. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).
183. Levy, R. Social media, news consumption, and polarization: evidence from a field experiment. *Am. Econ. Rev.* **111**, 831–870 (2021).
184. Santos, F. P., Lelkes, Y. & Levin, S. A. Link recommendation algorithms and dynamics of polarization in online social networks. *Proc. Natl. Acad. Sci. USA* **118**, e2102141118 (2021).
185. Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021).
186. Kubin, E., Puryear, C., Schein, C. & Gray, K. Personal experiences bridge moral and political divides better than facts. *Proc. Natl. Acad. Sci. USA* **118**, e2008389118 (2021).

Acknowledgements
The authors thank A. Morris, V. Chituc and C. Kealoha for helpful comments on prior drafts of this manuscript.

Author contributions
M.J.C., R.W.C. and Y.E.B. researched data for the article. All authors contributed substantially to discussion of the content. R.W.C. and Y.E.B. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Psychology thanks Michał Białek, Emma Levine, and the other, anonymous, reviewers for their contribution to the peer review of this work.