1 Modern Moral Psychology

A Guide to the Terrain

Bertram F. Malle and Philip Robbins

The term *moral psychology* is commonly used in at least two different senses. In the history of philosophy, moral psychology has referred to a branch of moral philosophy that addresses conceptual and theoretical questions about the psychological basis of morality, often (but not always) from a normative perspective (Tiberius, 2015). In the empirical investigations of psychology, anthropology, sociology, and adjacent fields, moral psychology has examined the cognitive, social, and cultural mechanisms that serve moral judgment and decision making, including emotions, norms, and values, as well as biological and evolutionary contributions to the foundations of morality. Since 2010, over six thousand articles in academic journals have investigated the nature of morality from a descriptive-empirical perspective, and this is the perspective the handbook emphasizes. Our overarching goal in this volume, however, is to bring philosophical and psychological perspectives on moral psychology into closer contact while maintaining a commitment to empirical science as the foundation of evidence. Striving toward this goal, we have tried to cast a wide net of questions and approaches, but naturally we could not cover all topics, issues, and positions. We offer some guidance to omitted topics later in this introduction, which we hope will allow the reader to take first steps into those additional domains.

The chapters try to strike a balance between being up to date in a fast-moving field and making salient insights that have garnered attention for an extended time. The reader may consult some of the main journals publishing on moral psychology to follow the latest research in the field. Table 1.1 lists the journals that, in a recent database search, published the largest number of articles on "moral psychology." We see frequent contributions from journals in social and cognitive psychology but also from generalist and philosophy journals.

As several of the chapters illustrate, much research in moral psychology has been informed in one way or another by the work of moral philosophers. For this reason, it may be helpful for the reader to bear in mind the theoretical perspectives on morality that tend to dominate the philosophical literature in ethics and metaethics. Four of these perspectives have been especially influential in moral psychology. First, there is act utilitarianism, the idea that right actions are those actions that have the best consequences, as measured by aggregate utility (Mill, 1998). Second, there is deontology, according to which the moral permissibility of an action is determined by whether it conforms to a set of

Table 1.1 *Journals that publish the largest proportion of research on moral psychology*

10 general journals (in alphabetical order)	*Social Psychology*
Cognition	*Zeitschrift für Sozialpsychologie*
Developmental Psychology	
Frontiers in Psychology	**Applied or Domain Focus**
Journal of Experimental Psychology: General	*Ethics & Behavior*
Journal of Experimental Social Psychology	*Journal of Business Ethics*
Journal of Personality and Social Psychology	*Journal of Moral Education*
Personality and Individual Differences	*Nursing Ethics*
Personality and Social Psychology Bulletin	*Psychological Trauma*
PLoS ONE	*Social Science & Medicine*
Social Psychological and Personality Science	*Traumatology*
Next 10	***Additionally:* Theoretical Focus**
British Journal of Social Psychology	*Ethics*
Emotion	*Journal of Philosophy*
European Journal of Social Psychology	*Mind and Language*
Journal of Applied Psychology	*Personality and Social Psychology Review*
Judgment and Decision Making	*Philosophical Review*
Philosophical Psychology	*Psychological Review*
Psychological Science	*Review of Philosophy and Psychology*
Social Cognitive and Affective Neuroscience	

Note. The first three groups stem from a search for keyword *moral** conducted July 31, 2023, on all Academic Search Premier databases, showing the peer-reviewed journals that published the highest number of empirical articles between 2010 and 2023. The fourth category is a listing of important outlets for theoretical work in the field.

abstract rules, such as the rule that people should be treated as ends in themselves, rather than solely as means to an end (Kant, 1785/1998). The contrast between act utilitarian and deontological commitments is vividly illustrated by sacrificial dilemma cases, in which prioritizing the good of the many would violate the rights of the one (Thomson, 1985). A third option, which effectively splits the difference between act utilitarianism and deontology, is rule utilitarianism, according to which an action is morally right just in case it is required by an optimific social rule, that is, a rule that would tend to maximize aggregate utility if everyone were to follow it. Finally, there is virtue ethics, which shifts the focus from how people should behave to what sort of character traits people should cultivate (namely, the virtues). On this view, moral standards for behavior are determined by what a hypothetical virtuous person or persons would do in the context: Right actions are actions that all virtuous people would do, wrong actions are actions that no virtuous person would do, and merely permissible (i.e., neither right nor wrong) actions are actions that some virtuous people would do.

We put no constraints on authors to align themselves with a particular metaethical position. Some of the chapters could be assigned to well-known positions that have influenced the field: utilitarian (Baron, in Chapter 8; Niemi

& Nichols, in Chapter 7), deontological (Andrighetto & Vriens, in Chapter 4; Malle, in Chapter 15), and virtue-based (Narvaez, in Chapter 17). Other chapters do not take any particular position but speak to topics relevant to those positions (Demaree-Cotton & Kahane, in Chapter 5; Goodwin & Landy, in Chapter 2; Robbins, in Chapter 9; Shweder et al., in Chapter 20).

1.1 The Landscape of Morality

The broad topic of morality encompasses a variety of more specific phenomena, such as moral judgment, moral decision making, moral emotions, moral norms, and more. In this section we briefly discuss and distinguish these different phenomena that make up the landscape of morality (following Bello & Malle, 2023) and highlight which of the chapters speak to each of the phenomena.

Morality exists only against the background of a community's moral standards. Moral judgment would not be *moral* judgment unless it is made relative to a set of moral standards, which are typically referred to as norms and values; the same holds for what makes decision making moral, what makes communication moral, and so on. Though all are embedded in norms and values, several phenomena of morality must be distinguished, and Figure 1.1 highlights some of the more important distinctions. There is diversity within each of the phenomena: Within moral communication, for instance, we would find forgiving, justifying, and praising, and numerous emotions have been considered *moral* emotions (e.g., guilt, outrage, contempt, and disgust). But the boundaries between the phenomena can be drawn in meaningful ways, at least to organize the sets of questions and psychological mechanisms under investigation.

1.1.1 Moral Behavior

Moral behavior includes intentional acts (often studied under the label *moral decision making*) but also unintended or negligent behavior. The processes underlying an agent's moral behavior are distinct from the processes underlying an observer's moral judgments of *another* person's behavior. This distinction helps sharpen terms like *moral sense* (Marazziti et al., 2013; Wilson, 1993), which can refer to behavioral phenomena such as altruism and moral disengagement or to evaluations and judgments of other people's behavior (and sometimes one's own). Moral judgments and moral behaviors take the same norms into account and are responsive to similar kinds of information (e.g., justifying reasons, causal counterfactuals), but their underlying processes are distinct, and conclusions from one do not necessarily apply to the other.

Philosophy and psychology have long focused on moral decision making as the primary driver of moral behavior. Decision making is moral if it refers to choices between possible paths of action in light of moral norms. In principle, moral decisions are no different from other decisions (Zeelenberg et al., 2012). But because of the deep involvement of norms, moral decisions take on key

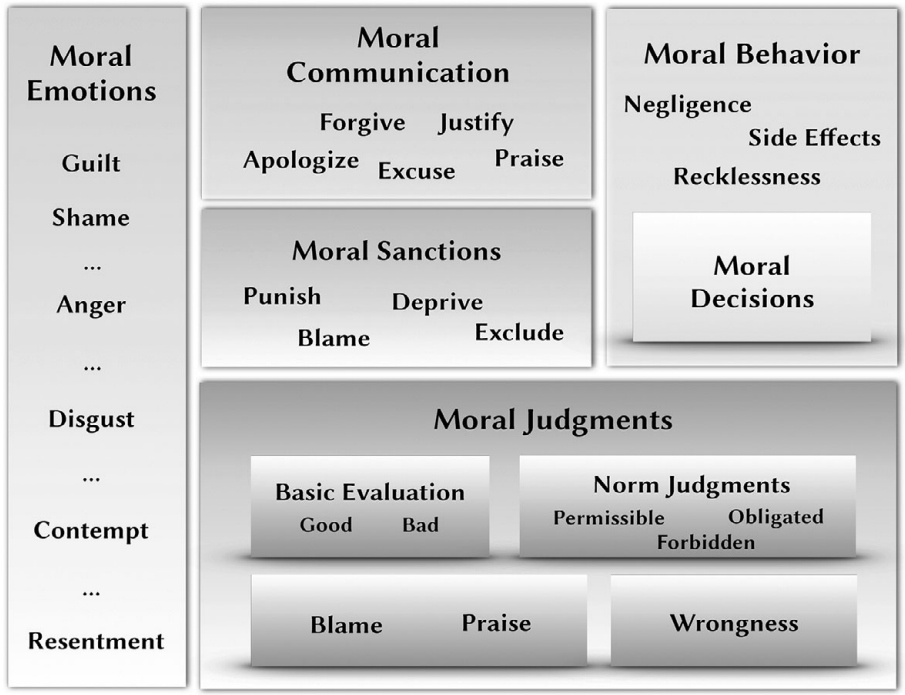

Figure 1.1 *The landscape of morality and its five major territories: moral behavior (including moral decision making), moral judgments (including multiple types, such as evaluation, wrongness, and blame), moral sanctions, moral emotions, and moral communication (expanded from Bello & Malle, 2023, Figure 31.1).*
Source: Sun, The Cambridge Handbook of Computational Cognitive Sciences*, 2023 ©, published by Cambridge University Press, reproduced with permission.*

properties of norms, including their substantial context sensitivity (Bartels et al., 2015) and keen responsiveness to what the community thinks and does (Bicchieri, 2006). A long tradition of psychological work has also examined moral decision making in light of moral values and principles (Kohlberg, 1981; Schwartz, 1992). Although such abstract guides can undoubtedly influence concrete moral decisions (and the justifications of those decisions), the question of whether a particular principle applies to a given problem is still guided by context-specific normative considerations. In fact, details in the setting and type of action under considerations can distinctly affect which moral principles dominate other principles (Christensen & Gomila, 2012).

However, moral decisions – intentional acts by their nature – are only one part of the territory of moral behavior. Many morally significant behaviors are unintentional (such as negligence, recklessness, preventable accidents, or unintended side effects), and moral communities respond strongly to such behaviors (Laurent et al., 2016; Monroe & Malle, 2019). These responses, in turn, constitute the second major territory of the moral landscape: moral judgments.

1.1.2 Moral Judgments

When people make a moral judgment, they appraise an object in light of moral norms. These appraisals differ considerably depending on the object of appraisal – an event, behavior, or person – and the information that guides the appraisal – about an action, its reasons, caused outcomes, counterfactuals, and more (Alicke, 2000; Cushman, 2008; Malle, 2021). In the philosophical literature, the term *moral judgment* often refers to one of these kinds: first-person appraisals that a behavior one might perform is right or wrong (e.g., Ratoff & Roskies, Chapter 3 in this volume). In this meaning of the term, moral judgment directly underlies moral (intentional) action. We follow here the broader use of the term (more common in empirical moral psychology), in which moral judgments can refer to both first-person and third-person good–bad evaluations, norm judgments (what is prescribed or prohibited), and wrongness judgments, as well as blame judgments and character judgments (Malle, 2021).

One might distinguish these kinds of moral judgments by their position in a processing hierarchy. Very often, the flow of information processing begins with the detection of a norm violation, so norms may already be cognitively activated when the other moral judgments are formed. The simplest and fastest judgments are evaluations (Yoder & Decety, 2014), followed by wrongness judgments (Cameron et al., 2017); more complex are judgments of blame and character (Malle et al., 2014; Murray et al., 2024), which build on the simpler ones. Another way to distinguish moral judgments is by the functions they serve when expressed in social settings. Norm judgments serve to persuade others to (not) take certain actions ("That's not allowed!"), declare applicable norms (e.g., posted rules of conduct), and teach others ("The appropriate thing to do here is . . ."). Stating a behavior's moral wrongness mainly serves to mark a behavior as a moral transgression, especially when it is seen as intentional. Blame, finally, criticizes, influences reputation, and regulates relationships (Coates & Tognazzini, 2012).

While blame has been investigated extensively, less research is available on praise. Praise and blame are by no means mirror images of one another, and scholars have documented numerous asymmetries between the two judgments (Bostyn & Roets, 2016; Guglielmo & Malle, 2019; Hindriks, 2008; Pizarro et al., 2003). Both take into account the agent's mental states and the performed behavior's relation to relevant norms. But whereas blame tries to bring an agent who violated a norm back in line with the norm, praise identifies an action that exceeds normative expectations (Monroe et al., 2018) and rewards the agent for that action, helping to build social relationships (Anderson et al., 2020).

1.1.3 Moral Sanctions

Aside from examining moral decisions and judgments, scholars have examined another class of responses to morally significant behavior: *moral sanctions*.

Whereas moral judgments are typically considered in the perceiver's head, moral sanctions are social acts that express a moral judgment, impose a cost on the transgressor, and regulate the transgressor's and other people's future behavior. Most prominent among sanctions is punishment, often cast as the backward-looking act of retribution (literally payback), said to fulfill a desire to hurt the transgressor (Goodwin & Gromet, 2014). However, punishment is more complex. First, punishment can be an act of affirming a norm system (Fehr & Fischbacher, 2004) and of teaching (Cushman, 2013); if done properly, it can maintain cooperation in a community (Fehr & Gächter, 2000), mainly when it is accompanied by communication about the relevant norms (Andrighetto et al., 2013). Second, many forms of punishment have emerged rather recently in cultural human history, primarily as institutional behavior regulation closely tied to the law (Cushman, 2013; Malle, Chapter 15 in this volume). By contrast, everyday moral sanctions x are rarely as harsh and physical as the institutional ones; instead, they range from complaints (Drew, 1998) and acts of moral criticism (Moisuc & Brauer, 2019) to shaming or exclusion (Panagopoulou-Koutnatzi, 2014). In further contrast to institutional punishment, these informal sanctions are often negotiable and can even be taken back, and they are subject to social scrutiny to ensure they are appropriate and fair (Friedman, 2013; Malle et al., 2022).

1.1.4 Moral Emotions

Emotions play at least two roles in the landscape of morality. First, many scholars have identified so-called moral emotions, such as guilt, shame, or contempt (Haidt, 2003; Prinz & Nichols, 2010; Tangney et al., 2007). Which emotions fall under this special label has long been debated, and Russell (Chapter 10, this volume) examines in detail the possible criteria one might apply to such designations. Second, scholars have considered the role of emotions as either causes, concomitants, or consequences of moral judgments and decisions (Monin et al., 2007). We see historical oscillations between opposing positions on this matter, between philosophers such as Kant and Hume as well as, more recently, between waves of empirical research, claiming moral judgments to be either primarily a matter of reason or primarily a matter of emotion. Over the past half century, early research cast moral judgment as deliberate, mature, and complex (Kohlberg, 1981). Then a revolution occurred in which morality was reframed as based primarily on evaluations, emotions, and unreasoned intuitions (Greene, 2008; Haidt, 2001; Prinz, 2006). Over the past decade, several scholars have cast doubt on previous evidence for this "unreason" picture of moral judgment (Guglielmo, 2015; Royzman et al., 2009; Sauer, 2012), provided new evidence for the significant role of cognition and reasoning (Guglielmo & Malle, 2017; Martin & Cushman, 2016; Monroe & Malle, 2019; Royzman et al., 2011) and even evidence for the possible temporal precedence of moral cognition over emotion (Cusimano et al., 2017; Yang et al., 2013). Models that ascribe primary causality to affect and emotion have been

called out as underspecified (Huebner et al., 2009; Mikhail, 2008), and perhaps in response, information processing models have aimed to offer more theoretical detail while still allowing room for the role of emotion (Malle et al., 2014; May, 2018; Sauer, 2011). Interestingly, many researchers dedicated to the study of emotion per se (whether or not involved in morality) have developed models that integrate cognitive and affective processes (e.g., Scherer, 2009). A newly emerging position is that moral emotions have pronounced social functions, including to signal norm commitments and express moral judgments (Grappi et al., 2013; Kupfer & Giner-Sorolla, 2017; Sorial, 2016).

1.1.5 Moral Communication

This brings us to the final territory of morality, the tools and practices of communicating about moral norms, violations, and their sequelae. When a norm violation occurs, people almost automatically make moral judgments in their heads, but at least some of the time they express their moral criticism (Molho et al., 2020; Przepiorka & Berger, 2016), ask transgressors to account for their actions (Semin & Manstead, 1983), and sometimes grant forgiveness even for grave atrocities (Gobodo-Madikizela, 2002). Transgressors, on their part, will often try to explain or justify their violations (Gollan & Witte, 2008; Riordan et al., 1983) and mitigate others' criticism with remorse, apologies, or compensation (Tedeschi & Reiss, 1981; Yucel & Vaish, 2021). Even though it is through communication that people typically regulate other community members' moral behavior (Andrighetto et al., 2013; Shank et al., 2019), we see overall less research in moral psychology dedicated to these social-communicative processes than to the cognitive processes that undergird them. In this handbook, therefore, one chapter (Funk & McGeer, Chapter 16) directly speaks to the important communicative sphere, and several others draw connections (Guan et al., Chapter 22; Malle, Chapter 15; Shweder et al., Chapter 20).

1.2 Guide to Additional Topics

Given the vast landscape of morality, this handbook cannot be a complete map of its territory. In this section, we provide pointers to topics that did not end up in the handbook but present exciting and valuable directions of work.

1.2.1 Moral Psychology of Artificial Agents

The first topic of interest, centering on facets of "artificial morality," has seen a rapid rise over the past 10 years. Two recent reviews in the psychological literature took stock of some of the garnered insights (Bonnefon et al., 2024; Ladak et al., 2023), and several other reviews have surveyed some of the core questions and initial answers (Bigman et al., 2019; Malle, 2016; Misselhorn,

2018; Pereira & Lopes, 2020). The range of questions is broad: how to design machines that follow norms and make moral judgments and decisions (Cervantes et al., 2020; Malle & Scheutz, 2019; Tolmeijer et al., 2021) and how humans do and will perceive such (potential) moral machines (Malle et al., 2015; Shank & DeSanti, 2018; Stuart & Kneer, 2021); legal and ethical challenges that come with robotics (Lin et al., 2011), such as challenges posed by social robots (Boada et al., 2021; Salem et al., 2015), autonomous vehicles (Bonnefon et al., 2016; Zhang et al., 2021), autonomous weapons systems (Galliott et al., 2021), and large language models (Harrer, 2023; Yan et al., 2024); deep concerns over newly developed algorithms that perpetuate sexism, racism, or ageism; and tension over the use of robots in childcare, eldercare, and health care, which is both sorely needed and highly controversial (Sharkey & Sharkey, 2010; Sio & Wynsberghe, 2015).

Artificial agents raise a number of vexing philosophical questions, such as whether they could ever have consciousness or free will, whether those properties would be required to grant them rights and moral-legal standing, whether machines could ever be genuine moral agents (or moral patients), and many more. Work on artificial agents can also inform psychological theories of moral phenomena. For example, what features do artificial agents have to have in order for people to spontaneously impose norms on them, ascribe morally relevant mental states to them (e.g., justified reasons), and exchange moral emotions with them (e.g., forgiveness to reduce guilt)? Is evidence for deep psychological complexity necessary, or might mere humanlike appearance trigger fundamental moral responses? Finally, artificial agents provide opportunities to develop more precise theoretical and computational models of moral phenomena (e.g., norms, decisions), to test them first in simulations, and eventually in actual physical implementations. But as computationally more sophisticated designs begin to show sign of moral competence (Bello & Malle, 2023; Conte et al., 2013; Pereira & Lopes, 2020), is there something lost by reducing moral judgments, decisions, and emotions to long strings of code?

1.2.2 Morality, the Self, and Identity

The nature of the self and identity has long been a central topic in social psychology (Leary et al., 2003; Suls, 2014; Wylie, 1979) and philosophy (Metzinger, 2004; Olson, 1999; Parfit, 1992). There is also a rich literature connecting the study of morality to the self and identity. We can divide this literature into two main strands. The first strand connects morality and the self; the second strand connects morality and identity.

In social psychology, the word "self" is used to mean a variety of different things (Leary & Tangney, 2012). Two of the more common meanings are the *executive self*, which regulates an agent's behavior (Baumeister & Vohs, 2003), and the *evaluative self*, which is comprised of the thoughts and feelings people have about themselves, especially in relation to others (Tesser, 1988). The executive self plays a key role in the production of moral behavior. Bandura

(1999) calls *inhibitive moral agency* the capacity to refrain from acting inhumanely toward others. This kind of behavioral self-regulation in turn depends upon the capacity for self-directed negative emotions (e.g., guilt, regret, shame), which motivate conformity to moral norms by making immoral behavior aversive to the agent (Bandura, 1999; Silver & Silver, 2021). Indeed, when self-directed negative emotions are uncoupled from immoral behavior – whether through self-serving justification of the behavior, displacement of responsibility for its consequences, or dehumanization and blaming of the victims – the results can be literally catastrophic (e.g., war, mass murder, genocide) (Bandura, 1999).

The *evaluative self* can also be a significant determinant of moral behavior, since effective self-regulation requires the capacity for monitoring one's own behavior and evaluating it in relation to moral standards. Such standards are an essential part of most people's self-concept, and people are highly motivated to think of themselves as morally upright (Steele, 1988). In fact, research has shown that a moral self-concept emerges in early childhood and is a predictor of prosocial behavior (Christner et al., 2020). Reflecting on oneself as a good person can indeed lead to more good behavior (Young et al., 2012), especially when such reflections link up to abstract values; more concrete recall of past good deeds can lead to opposite effects, called *moral licensing* (Merritt et al., 2010), whereby people become more prone to immoral behavior after engaging in behavior that boosts their moral self-esteem. Thus, the executive self and the evaluative self do not always pull in the same moral direction.

Morality is also intimately tied up with identity, in two key meanings of the term "identity." *Diachronic identity* encompasses the features that ground the perceived persistence of persons over time. Multiple studies provide support for the idea that continuity of moral traits is seen as essential to the persistence of persons over time, insofar as someone who changes their moral stripes (or loses them altogether) is no longer seen as the same person (Hitlin, 2011; Prinz & Nichols, 2010; Strohminger & Nichols, 2014). Further support for this idea comes from studies showing that dramatic improvement of a person's moral character tends to result in mitigation of blame and reduction of moral responsibility for past immoral behavior (Gomez-Lavin & Prinz, 2019) – almost as if another person had performed those behaviors.

Synchronic identity encompasses the features that determine (or seem to determine) who a person is at a given time. In particular, the "true self" refers to the features that are seen as *essential* to a person's synchronic identity (Strohminger et al., 2017), and they are often features related to morality (see Goodwin & Landy, Chapter 2 in this volume). For example, whether an emotionally driven action is seen as expressive of an agent's true self depends on whether the action is morally bad or morally good (Newman et al., 2015). Further, studies suggest that moral evaluation of a person's behavior, including blame for immoral behavior and praise for moral behavior, is sensitive to perception of whether the behavior expresses the agent's true self (Newman et al., 2015; Robbins & Alvear, 2023).

1.2.3 Free Will and Moral Responsibility

Philosophers agree on two principles: first, moral judgments apply to an action only if the agent is morally responsible for performing it; second, agents are morally responsible only for those actions that they freely choose to perform. Beyond these points, the consensus tends to break down. For example, philosophers debate what it means for an action to be freely chosen and whether human actions ever meet that condition. One major source of disagreement is the assumption of causal determinism, shared by most philosophers – the idea that every event, including every human choice, is fully determined by the causal history of the world leading up to it. According to one view, an action is freely chosen just in case the agent could have made a different choice even if every link in the causal chain of events leading up to the actual choice had been the same. On this view, known as incompatibilism, the existence of free will – and by extension, that of moral responsibility – is incompatible with causal determinism. According to an alternative view, an action is freely chosen if the choice was free of certain types of constraints, either external (e.g., coercion) or internal (e.g., compulsion). On this second view, known as compatibilism, the existence of free will – and by extension, that of moral responsibility – is compatible with causal determinism.

But those are just the views of philosophers. What do ordinary people think about these matters? How do they make sense of the ideas of free will, moral responsibility, and causal determinism? Do ordinary people find incompatibilism more intuitive than compatibilism, or the other way around? These questions have animated empirical research with the goal of identifying the psychological origins of a philosophical problem as old as philosophy itself (Nichols, 2011).

Regarding the commonsense concept of free will, the preponderance of evidence suggests that ordinary people do not think of free will in the metaphysically demanding way presupposed by incompatibilism (Monroe et al., 2017; Monroe & Malle, 2010; Nahmias et al., 2005; Vonasch et al., 2018). In commonsense thinking, free choice is a matter of freedom from the kinds of local constraints that make it difficult for an agent to express their values and commitments (Woolfolk et al., 2006). This ordinary concept of free will corresponds most closely to the compatibilist one, according to which the metaphysical issue of causal determinism is irrelevant to the reality of free will (Strawson, 1962). This is not altogether surprising, given that the concept of causal determinism is sufficiently esoteric that it may be difficult for people without philosophical training to understand what incompatibilists are worried about (Sommers, 2010).

By contrast, empirical studies of ordinary people's intuitions about moral responsibility have yielded mixed results. Alongside evidence of the intuitive appeal of compatibilism about moral responsibility, together with evidence for the idea that incompatibilist intuitions result from confusing causal determinism with fatalism (Nahmias et al., 2007; Nahmias et al. 2014), there is evidence

for the opposite view, as well as support for the idea that compatibilist intuitions about moral responsibility result from affective bias (Nichols & Knobe, 2007). Making sense of the diversity of findings in this area, much of it originating in work by experimental philosophers, is an ongoing project in moral psychology. The same applies to research on the effect of free will beliefs on moral behavior, some of which suggests that disbelief in free will (in the metaphysically robust sense presupposed by incompatibilism) is associated with a greater propensity for aggression and dishonesty (Vohs & Schooler, 2008), whereas other studies find no evidence for these claims (Open Science Collaboration, 2015). Likewise, there is some empirical support for the idea that disbelief in free will makes people more punitive (Krueger et al., 2013), but efforts to replicate such results have failed (Monroe et al., 2014). In fact, a recent meta-analysis of 145 experiments showed that manipulating free will beliefs has few, if any, downstream consequences (Genschow et al., 2023).

1.2.4 Other Topics Yet

The chapters included in this handbook touch on numerous other exciting strands of moral psychology that did not receive a dedicated chapter to review their full respective literatures. For example, chapters by Decety (Chapter 11) and by FeldmanHall and Vives (Chapter 12) engage with the affective and cognitive neuroscience of morality, chapters by Narvaez (Chapter 17) and by Baird and Matthews (Chapter 19) connect to its neurobiological underpinnings. The reader may consult additional recent work that uses insights from neuroscience to analyze long-standing philosophical issues, such as free will, consciousness, and rationalism of moral judgment (Castro-Toledo et al., 2023; May, 2023).

Likewise, methods and insights from behavioral economics appear in chapters by FeldmanHall and Vives (Chapter 12), Niemi and Nichols (Chapter 7), and Purzycki and Bendixen (Chapter 23), and the reader may want to explore additional work on the interplay between economic and moral behavior (Vila-Henninger, 2021) and on the moral impact of exposure to market processes (Bartling & Özdemir, 2023; Enke, 2023; Fike, 2023). The connection between behavioral economics paradigms and computational and cognitive neuroscience measures is another interesting recent direction (Fornari et al., 2023; Lengersdorff et al., 2020).

Evolutionary perspectives on the origins of morality are distributed over chapters by Narvaez (Chapter 17), Shweder et al. (Chapter 20), and Malle (Chapter 15), whereby the latter two focus on cultural rather than biological perspectives. Animal behavior work arises in chapters by Decety (Chapter 11) as well as FeldmanHall and Vives (Chapter 12), and the reader may benefit from integrative perspectives on phylogenetic and cultural evolution by Boehm (2018), de Waal (2014), and Tomasello (2016), and a provocative recent proposal that links genetic heritability patterns to domains of cooperative morality (Zakharin et al., 2024).

Additional topics with less representation but no less significance include the group dynamics of morality (Ellemers et al., 2023), moral learning (Cushman et al., 2017), trust (Bach et al., 2022; Malle & Ullman, 2021; Sztompka, 2019), and morality in organizations and collectives (Blomberg & Petersson, 2024; Dhillon & Nicolò, 2023; Sattler et al., 2023).

1.3 Overview of the Chapters

We now offer brief summaries of each handbook chapter, hoping that the reader will find many of these contributions enticing for further reading.

1.3.1 Part I: Building Blocks

Part I introduces some of the basic building blocks of moral psychology, topics of both core theoretical concern and major historical significance.

Geoff Goodwin and Justin Landy (Chapter 2) review empirical research on moral character, which has only recently attained a prominent role in psychology, in contrast to long traditions in ethics and education. A person's moral character comprises the dispositions to think, feel, and act morally, and these dispositions are cross-situationally and temporally fairly consistent. Against a long-standing belief in psychology that the personality disposition of warmth most strongly influences people's impressions of one another, the evidence suggests that moral character occupies this central position. Moral character exerts its influence on impressions quite independently of other personality traits, and it features prominently in people's representations of their own personality as well. Moral character is also a central element in a person's perceived identity – who the person is perceived to be "deep down" (cf. our discussion of the "true self" in the Morality, the Self, and Identity subsection [Section 1.2.2]). Finally, the authors close by charting some of the features from which people infer another's moral character, including actions but also, critically, mental states such as goals and intentions.

William Ratoff and Adina Roskies (Chapter 3) tackle the question of how first-person moral judgments and moral behavior are conceptually linked. They frame their discussion in terms of a philosophical puzzle known as "Hume's problem." The puzzle arises from the conjunction of three ideas: Humeanism, the idea that beliefs alone do not suffice to motivate action; internalism, the idea that moral judgments are intrinsically motivating; and cognitivism, the idea that moral judgments are beliefs. These three ideas are jointly inconsistent, so at least one of them must be false. But which one? The authors focus their attention on two possible solutions to the puzzle: the externalist solution, which denies that moral judgments are intrinsically motivating (rescinding internalism), and the noncognitivist solution, which denies that moral judgments are beliefs (rescinding cognitivism). The authors review empirical research to explore whether either of the solutions is supported by evidence. On the issue

of whether moral judgments are intrinsically motivating, they argue that studies of moral cognition in psychopathy and acquired sociopathy do not settle the matter, nor do studies of folk intuitions about internalism. Likewise, studies of the influence of emotion on moral judgment do not settle the dispute between cognitivism and noncognitivism, since they do not establish that emotion is constitutive of moral judgment in the way that noncognitivism requires. Thus, an empirically compelling solution to Hume's problem remains to be found.

Giulia Andrighetto and Eva Vriens (Chapter 4) examine the foundational role of norms in moral psychology, a topic that has long garnered cross-disciplinary interest from philosophy to biology, from anthropology to computer science. The authors touch briefly on the debates over potentially different types of norm (e.g., conventional, social, moral, legal) and maintain that social and moral norms, in particular, are difficult to separate unless one adopts a specific theoretical position. The authors' treatment centers on a core feature of most or all social and moral norms: that people, in complying or not complying with norms, are sensitive to other community members' norm-relevant beliefs and attitudes. By recognizing this sensitivity, scientists can, first, gain a better scientific understanding of *norm inference*, the complex processes by which people learn which norms apply to a given setting and how strong the norms are; and second, they can better diagnose whether (and how strongly) a given norm actually exists in a community. All these insights pave the way for potential interventions on people's *beliefs* about the community's norms, which are easier to change than individual moral convictions.

Joanna Demaree-Cotton and Guy Kahane (Chapter 5) introduce a frequently discussed topic in recent moral psychology: moral dilemmas. They characterize moral dilemmas as a decision-making situation that has three features: first, every available course of action has a high moral cost and therefore involves a difficult moral trade-off; second, it is morally appropriate for the agent to feel conflicted about what choice to make; and third, it is morally appropriate for the agent to feel some regret about whatever choice they made. The authors then explore different empirical accounts of why some moral trade-offs, but not others, are experienced as difficult or impossible to resolve. Among the most influential of these accounts is Greene's (2008) dual-process theory, which traces the experience of moral dilemmas to a conflict between a value backed by intuition ("System 1") and a value backed by reflection ("System 2"). The authors also review empirical research bearing on the psychological mechanisms underpinning a person's resolution of moral dilemmas and the phenomenon of "moral residue" (regret or guilt over one's resolution). They argue that further empirical work is needed to understand how people weigh competing values against one another and that such understanding requires expanding the range of moral dilemmas to include cases beyond those targeted in recent research (e.g., sacrificial dilemmas).

Samantha Abrams and Kurt Gray (Chapter 6) tackle another foundational question: What constitutes the moral domain? To answer this question, they explore three approaches to modeling moral cognition, focusing on three issues:

first, what behaviors are seen as morally wrong; second, whether moral norms are universal rather than culturally variable; and third, what psychological mechanisms underlie judgments of moral wrongness. According to Turiel's model, wrong behaviors are those seen as harmful or unfair, moral norms are universal, and wrongness judgments are largely the result of conscious reasoning from abstract principles. By contrast, in Haidt's model, wrong behaviors are not just those seen as harmful or unfair, but also those seen as disloyal, disrespectful of authority, or impure; moral norms exhibit substantial cross-cultural variation; and wrongness judgments are typically the product of intuition, rather than conscious reasoning. The model favored by the authors combines elements from both of these approaches: from Turiel, the idea that perceptions of wrongness boil down to perceptions of harm; and from Haidt, the idea that moral norms are culturally variable and the idea that wrongness judgments are more a product of intuition than reasoning.

1.3.2 Part II: Thinking and Feeling

Part II focuses on the cognitive and affective processes that make up various moral phenomena: moral decision making, moral judgment, the categorization of agents and patients, and moral emotions.

Laura Niemi and Shaun Nichols (Chapter 7) introduce some core elements of moral decision making by taking expected utility theory as a starting point. In its classic form, expected utility theory focuses on the outcomes of actions: The expected utility of a decision is the sum of the values associated with the different possible outcomes of the decision weighted by the probability of their occurrence. As such, expected utility theory is well suited to explain the moral choices recommended by utilitarianism, which characterizes right actions in terms of the maximization of aggregate utility. However, to account for more complex, nonutilitarian decisions, expected utility theory must be modeled to assign utilities to actions themselves. This action-based form of expected utility theory can readily accommodate the fact that people tend to assign low utility to actions that violate moral norms (even when the outcomes of those actions might have positive utility, such as when lying would lead to financial gain). The authors then apply this expanded action-based expected utility theory of moral decision making to questions regarding what actions count as fair, how the decision maker's actions take other people's outcomes into account, and how the value of actions changes when directed at one's own or another's group.

Jonathan Baron (Chapter 8) posits utilitarianism as a standard of rational moral judgment. He does not directly defend utilitarianism as a theory but investigates cases of apparent contradiction between people's moral decisions (sometimes grounded in nonutilitarian principles) and the consequences of those decisions that they themselves would consider worse for themselves and everybody else. For example**,** when some people use a moral principle (e.g., bodily autonomy) to assertively make a decision (e.g., to not get vaccinated), it can have negative moral consequences for others (e.g., infecting people) *and* for

themselves (risking infection). Baron asks whether such contradictions in moral reasoning can provide insights into some of the determinants of such reasoning. These insights, importantly, are valuable even for those who do not adopt utilitarianism as a normative model. From over a dozen candidate moral contradictions, Baron concludes that many deviations from utilitarian considerations in moral contexts are reflections of familiar nonmoral cognitive biases (e.g., framing effects, certain concepts of causality), but some arise from adherence to strong moral rules or principles (e.g., protected or sacred values).

Philip Robbins (Chapter 9) discusses the role of mind perception in the categorization of individuals as moral agents and moral patients. Moral agents are defined as individuals who can commit morally wrong actions and deserve to be held accountable for those actions; moral patients are defined as individuals who can be morally wronged and whose interests are worthy of moral consideration. It is generally agreed that the attribution of moral agency and moral patiency is linked to the attribution of mental capacities. Robbins surveys a variety of models of mind perception, some of which focus on the representation of mental capacities, some of which focus on the representation of mental traits. The dominant model of mind perception in moral psychology is the experience–agency model (Gray et al., 2007), which divides the space of mindedness into experiential capacities like sentience and self-awareness, and agentic capacities like deliberative reasoning and self-control. Reviewing the empirical literature on moral categorization, Robbins argues that neither the experience–agency model nor any of the major alternatives to it (i.e., the warmth–competence model, the agency–communion model, and the human nature–human uniqueness model), captures the full panoply of mental features to which everyday attributions of moral agency and moral patiency are sensitive.

Pascale Sophie Russell (Chapter 10) asks whether, and in what ways, emotions can be designated as "moral." Several emotions have been shown to be associated with moral judgments or moral behaviors. But more than association must be shown if we label some emotions characteristically moral. Russell guides the reader through a voluminous literature and applies two criteria to test the moral credentials of emotions. The first criterion is whether the emotion is significantly elicited by moral stimuli (e.g., transgressions); the second is whether it has significant community-benefiting consequences. This second criterion, less often used in past analyses, tries to capture the fact that moral norms, judgments, and decisions are all intended to benefit the community, so moral emotions should too. From this analysis, the author concludes that anger clearly meets the criteria, contempt and disgust less so. Guilt passes easily, and shame fares better than some may expect. Among the positive candidates, compassion and empathy both meet the criteria but are somewhat difficult to separate. Finally, elevation and awe have numerous prosocial consequences, but awe is rarely triggered by moral stimuli.

Jean Decety (Chapter 11) examines the complex relation between empathy and prosocial behavior and considers findings from animal behavior,

neuroscience, and psychological studies. He begins by distinguishing three components of the broader phenomenon of empathy: emotional contagion, empathic concern, and perspective taking. He reviews evidence suggesting that emotional contagion of a conspecific's pain often leads to helping behavior, but such contagion is modulated by group membership, levels of intimacy, and attitudes toward the other. Thus, contagion is not an automatic trigger for prosocial behavior. Empathic concern, too, is a powerful motivator of prosocial behaviors but is also socially modulated – extended to some people more than others and to individuals more than groups. Effortful perspective taking, finally, can provide a better understanding of other people's minds but does not always generate prosocial behavior, even when it facilitates empathic concern. In sum, various forms of empathy can motivate prosocial behaviors, but empathy is fragile and often stops short of its potential when people engage with large groups, people outside of their tribe, or anonymous strangers.

1.3.3 Part III: Behavior

Part III focuses on some of the central classes of behaviors that scholars of morality have puzzled over: prosocial behavior, antisocial behavior, conflict, and dehumanization. It also examines the primary moral sanctioning behaviors (blame and punishment) by which humans respond to moral violations, and it ends on the topic of moral communication.

Oriel FeldmanHall and Marc-Lluís Vives (Chapter 12) highlight that, for successful social living, humans' capacity to be prosocial had to surpass their capacity for selfish and harmful behavior. The authors provide an overview of the scientific study of prosocial capacities, with a focus on experimental research. Summarizing extensive work in laboratory paradigms of behavioral economics and social psychology, the authors document a strong human tendency toward behaving prosocially. They then briefly examine the phyologenetic and developmental origins of behaving prosocially and its different motives, such as reputational concerns and caring for others, as well as emotions that facilitate prosocial behavior, such as empathy or guilt. (See Decety's chapter [Chapter 11] on empathy for a more comprehensive treatment of empathy.) FeldmanHall and Vives also summarize insights from cognitive neuroscience on the brain networks that undergird prosocial behavior. They close with a call for more naturalistic experimental paradigms and the consideration of temporal dynamics of prosocial behavior.

Kean Poon and Adrian Raine (Chapter 13) provide the counterweight to FeldmanHall and Vives by inspecting the relationship between antisociality and morality from the dual perspectives of moral psychology and moral neuroscience. Their chapter provides a comprehensive overview of research on the moral cognition of different types of antisocial individuals, focusing on the interplay between cognition and emotion in psychopathic individuals. Based on their review of the research, the authors suggest that the capacity for moral reasoning in psychopathy is less defective than generally assumed. While the

propensity of psychopathic individuals to engage in immoral behavior is due largely to affective deficits (e.g., low empathy), it also stems from dysfunction in the neural circuitry underlying moral decision making. This simple narrative, however, is complicated by the fact that there is no single explanation of the immoral behavior exhibited by the full range of antisocial individuals. For example, while dysfunction in the neural circuitry of moral decision making may account for the immoral behavior of individuals with primary psychopathy and individuals prone to proactive (i.e., instrumental) aggression, it is less apt for explaining similar behavior by individuals with secondary psychopathy and a propensity for reactive aggression.

Nick Haslam (Chapter 14) introduces dehumanization as another dark side of humanity. Humanness is a central concept in moral psychology, and whereas people normally treat other humans with moral consideration, they may turn to dehumanize others as a result of moral disengagement (loosening ordinary moral inhibitions) and moral exclusion (no longer applying norms of justice, fairness, and compassion to others). Haslam reviews recent psychological accounts of dehumanization that are grounded in empirical research and highlights several common threads: Dehumanization varies from subtle to extreme (e.g., genocide), interpersonal to intergroup, and from contexts of mere perception to contexts of severe conflict. In these theoretical accounts, dehumanizing a person or group means ascribing less of certain human attributes to the target – both attributes that distinguish humans from other animals (e.g., intellect, rationality, or civility) and attributes that distinguish humans from inanimate agents (e.g., essential capacities for emotion and warmth). Haslam's analysis meshes with that of Abrams and Gray (Chapter 6, this volume) and the discussion by Robbins (Chapter 9, this volume) of mental capacities people normally ascribe to other people – thus, dehumanization is a form of dementalizing. Within this framework, Haslam reviews the empirical literature on what forms dehumanization takes and what its possible functions are. He also considers a number of critiques and debates over these findings that have recently surfaced.

Bertram F. Malle (Chapter 15) compares the two major moral sanctioning behaviors of blame and punishment from two perspectives: their cultural history and their underlying psychology. He draws a dividing line between two phases of human evolution – before and after human settlement – and proposes that, before that watershed, moral sanctions were informal, nonhierarchical, and often mild, akin to today's acts of moral blame among intimates. Soon after settlement, hierarchies emerged, in which punishment took hold as a new form of sanctioning, typically exacted by those higher up in the hierarchy, eventually by institutions of punishment. Malle reviews the empirical evidence on the cognitive and social processes underlying each of these sanctioning tools and proposes that their distinct cultural histories are reflected in their psychological properties we can observe today. Whereas blame is, on the whole, flexible, effective, and cognitively sophisticated, punishment is often more damaging, less effective, and can easily be

abused – as in past and modern forms of institutional punishment. Compare this chapter to Janice Nadler's (Chapter 21, this volume) treatment of similarities and differences between blame in ordinary life and blame within the US legal system.

Friederike Funk and Victoria McGeer (Chapter 16) close this part of the book with a discussion of moral communication, in which the topic of punishment is also center stage. Their approach to the topic is somewhat unorthodox, insofar as the term *moral communication* is typically used to refer to a class of behaviors distinct from moral sanctions (see earlier discussion in this chapter of the landscape of morality, depicted in Figure 1.1). The authors argue that moral norms are distinctive in that their transgression tends to provoke a desire in members of the community to punish the transgressor, and that such punishment has a communicative function. Indeed, on their view, punishment is best understood as a nonlinguistic form of moral communication, one that expresses sharp disapproval of the transgressor's actions and attitudes. This approach to punishment has the potential to resolve conflicting results from studies of the effect of group membership on punishment, such as the fact that in-group transgressors are sometimes treated more leniently than out-group transgressors ("in-group favoritism") and sometimes more harshly (the "black sheep effect"). The solution to the puzzle, the authors argue, is that severity of punishment depends on who the intended target of communication is and what message the punishment is intended to convey.

1.3.4 Part IV: Origins, Development, and Variation

Part IV addresses questions of variability – from the evolutionary origins of morality to its development in the earliest phases of life, all the way to cultural variability.

Darcia Narvaez (Chapter 17) discusses morality from an evolutionary-developmental, cultural, and (to a lesser extent) neurobiological perspective. The framework for her discussion is *triune ethics metatheory*, a main tenet of which is that healthy moral development requires the provision by the community of an "evolved nest" in which caregivers treat children with love and respect. Failure to receive this support can limit a person's social and emotional competence necessary for species-typical moral functioning. The natural trajectory of moral development, Narvaez suggests, tends toward an *engagement-centered* ethic oriented around the virtues of cooperation, compassion, and egalitarianism – the ethic characteristic of Indigenous cultures (and of our hunter-gatherer ancestors, as Malle's chapter [Chapter 15] suggests). This path of development is readily disrupted by practices of child-rearing in Western industrialized societies, which deprive children of the social and emotional resources needed for healthy moral development, thereby promoting the development of a *self-protection-centered* ethic oriented around competition, cold-heartedness, and dominance. Thus, understanding the role of the evolved nest in scaffolding moral development is key to understanding why antisocial behavior

is so pervasive in modern Western culture – and to designing interventions that might help to reduce it.

Kiley Hamlin and Francis Yuen (Chapter 18) present a large body of evidence suggesting that, within the first year of life, infants hold both expectations about and preferences for morally good versus bad protagonists. Across different methods, the authors show that infants distinguish between morally significant acts of helping and hindering as well as between acting fairly and unfairly; they prefer the morally good actions and the morally good protagonists; and they expect others to prefer the morally good protagonists as well. Going beyond a mere valence difference, these expectations vary systematically in response to critical factors, such as victim's state of need, in-group/out-group membership, and a character's intentions. Many of the findings appear in infants 8–12 months of age, some as early as 3 months of age. Questions remain, such as how consistent the findings are across experimenters and populations; whether the violated norm is truly moral or only a social expectation; and to what extent earliest learning guides these expectations and preferences. But overall, the evidence for budding moral distinctions in early infancy is highly compelling and provocative.

Abigail A. Baird and Margaret M. Matthews (Chapter 19) take up the issue of moral development in adolescence, focusing on the role of individual differences in shaping the emergence of a mature moral sense. Their wide-ranging discussion touches on how differences in temperament, gender, familial and peer relationships, and lived experience influence the timing and outcome of adolescent moral development. Illustrating the role of temperament, for example, *high-reactive* individuals may be more prone to impulsive behavior that violates moral norms, whereas *low-reactive* individuals may be more likely conform to moral norms because they are more sensitive to the threat of punishment. Showing the importance of interpersonal relationships, weak attachment to caregivers in adolescence is associated with impairments of empathy and a greater propensity for antisocial and immoral behavior (a major theme in Narvaez's chapter [Chapter 17]). Peer influence is another key predictor of both antisocial and prosocial behavior in adolescence. Further, moral development in adolescence critically depends on the maturation of capacities for empathy and self-conscious emotion (e.g., guilt, embarrassment, pride), a process that is shaped by the individual's lived experience. In closing, the authors suggest that the powerful effects of individual differences on adolescent moral development are best accounted for by models that explain the maturation of the moral sense at multiple levels of analysis and timescales.

Richard A. Shweder, Jacob R. Hickman, and Les Beldo (Chapter 20) ask how one can scientifically examine the moralities of different human groups without falling into ethnocentrism – without morally judging the practices of other groups as wrong or unacceptably different from one's own. The authors propose to accept (at least as a methodological orientation) "moral realism" – the view that all human communities share a small set of "moral truths." These truths are abstract and must be expressed in culturally and historically specific

ways to be workable, and their differentiated expressions across different groups can make them seem irreconcilable. But by identifying moral absolutes, the authors suggest, scientists can make sense of the great variety of cultures and moralities and still recognize their commonalities. To illustrate their points, they discuss examples of clashing moral practices, such as between Brahman Indian and Western views of a widow's obligations, and between Native American whalers' and whaling protesters' attitudes toward whaling. Each of these groups sees their own moral position as "objective" (independent of social consensus) and "absolute" (true without need for justification), but underlying their seeming differences, the authors argue, there really might be shared moral truths.

It is worth pointing out that Baron (Chapter 8, this volume), too, suggests that people may hold some absolute moral principles (protected or sacred values). But whereas the moral realist featured by Shweder and colleagues suggests that denying these intuitively and instantly grasped truths is a sign of irrationality, the utilitarian featured by Baron suggests that holding onto such truths can lead to irrationality.

1.3.5 Part V: Applications and Extensions

Part V applies some of the core concepts and theories to the domains of law, politics, and religion and closes with a discussion of how empirical work in moral psychology bears on issues in moral philosophy.

Janice Nadler (Chapter 21) examines the sanctioning doctrines within Anglo-American criminal law and explores similarities and differences between criminal blame and ordinary social blame. Nadler takes on topics of intended but incomplete transgressive conduct, the distinction between intended and unintended outcomes, as well as questions of recklessness and the role of a transgressor's character in ordinary and legal blame. Nadler shows the complexity of the legal blame process and its many parallels in ordinary blame. On the legal side, she considers both the codified principles of US criminal law and the unwritten body of less precise standards and practices that can deviate from the codified ideals. On the ordinary side of blame, Nadler highlights the importance of both causal and mental factors that people take into account for intentional and unintentional transgressions. Nadler concludes that there is a great deal of congruity between legal and ordinary blame, especially in concepts and evidence considerations, but somewhat different goals and certainly more severe outcomes on the legal side (especially when errors or biases take hold). Compare this chapter to Malle's (Chapter 15, this volume) analysis of the cultural history, social regulation, and psychological processes underlying blame and punishment.

Kate W. Guan, Gordon Heltzel, and Kristin Laurin (Chapter 22) discuss the moral dimensions of political attitudes and behavior. They argue that a person's political views – both at the level of political ideology as a whole and views on

specific matters of economic and social policy – are profoundly shaped by their beliefs about right and wrong. These political views in turn drive people's political behavior, not just at the ballot box or on the campaign trail, but in the community more generally. One downside of the way in which moral convictions fuel political attitudes and behavior is that they tend to interfere with productive communication across partisan divides, fueling a kind of animosity that stifles cooperation and compromise. Divergence in people's moral convictions, then, leads inexorably to political polarization and gridlock. To address this problem, the authors discuss a number of potentially promising interventions, some of which target individuals' attitudes (e.g., promoting empathy, reducing negative stereotypes), and others that aim at improving the quality of interpersonal relationships (e.g., increasing contact, fostering dialogue across political divides).

Benjamin Grant Purzycki and Theiss Bendixen (Chapter 23) discuss the complex, multifaceted connection between morality and religion from an evolutionary perspective. After providing some much-needed conceptual ground clearing, the authors focus on accounts of the linkage between morality and religion in terms of evolved psychological mechanisms that promote cooperation and inhibit competition. One of the better known of these accounts is the supernatural punishment hypothesis. On this view, the morality–religion link is sustained by the fact that belief in an all-knowing, all-powerful god who monitors people's behavior and punishes their moral transgressions motivates people to behave less selfishly and more cooperatively. An alternative account is that participation in religious ritual is a form of costly signaling, indicating to others that the participant can be trusted to observe the moral norms of the community, including norms of cooperation. As a result, ritual activity comes to be associated with increased cooperation and decreased competition, at least within religious groups. While there is considerable support for the idea that religion can function as a recipe for kindness and a remedy for selfishness, however, the authors caution that the psychological mechanisms underlying this function are not yet well understood.

Paul Rehren and Walter Sinnott-Armstrong (Chapter 24) suggest some lessons from moral psychology for ethics and metaethics. They note that empirical research on a wide range of topics, including moral character, happiness and well-being, free will and moral responsibility, and moral judgment, has had a profound influence on recent philosophical theorizing about the foundations of morality. In their chapter they focus on one issue of particular importance: the reliability and trustworthiness of moral judgment. They critically assess multiple lines of argument that threaten to undermine epistemic confidence in our moral judgments, including evolutionary debunking arguments, process arguments, arguments from disagreement, and arguments from irrelevant influences. Though the jury is still out on how successful these arguments are, there is little question that they have potentially profound implications both for moral epistemology (insofar as they pose a threat to moral intuitionism) and

philosophical methodology (insofar as they cast doubt on the thought-experimental method). Perhaps the most important lesson for ethics and metaethics to be drawn from moral psychology, then, may be that future progress in moral philosophy is likely to depend on philosophers and psychologists working together, rather than in isolation from one another.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, *24*(9), 694–703.

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS ONE*, *8*(6), Article e64941.

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction*, *40*(3), 1–16.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209.

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 478–515). John Wiley & Sons, Ltd.

Bartling, B., & Özdemir, Y. (2023). The limits to moral erosion in markets: Social norms and the replacement excuse. *Games and Economic Behavior*, *138*, 143–160.

Baumeister, R. F., & Vohs, K. D. (2003). Self-regulation and the executive function of the self. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 197–217). Guilford Press.

Bello, P., & Malle, B. F. (2023). Computational approaches to morality. In R. Sun (Ed.), *Cambridge handbook of computational cognitive sciences* (pp. 1037–1063). Cambridge University Press.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365–368.

Blomberg, O., & Petersson, B. (2024). Team reasoning and collective moral obligation. *Social Theory and Practice* 50(3), 483–516.

Boada, J. P., Maestre, B. R., & Genís, C. T. (2021). The ethical issues of social assistive robotics: A critical literature review. *Technology in Society*, *67*, Article 101726.

Boehm, C. (2018). Collective intentionality: A basic and early component of moral evolution. *Philosophical Psychology*, *31*(5), 680–702.

Bonnefon, J.-F., Rahwan, I., & Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology*, *75*(1), 653–675.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Bostyn, D. H., & Roets, A. (2016). The morality of action: The asymmetry between judgments of praise and blame in the action–omission effect. *Journal of Experimental Social Psychology*, *63*, 19–25.

Cameron, C. D., Payne, B. K., Sinnott-Armstrong, W., Scheffer, J. A., & Inzlicht, M. (2017). Implicit moral evaluations: A multinomial modeling approach. *Cognition*, *158*, 224–241.

Castro-Toledo, F. J., Cerezo, P., & Gómez-Bellvís, A. B. (2023). Scratching the structure of moral agency: Insights from philosophy applied to neuroscience. *Frontiers in Neuroscience*, *17* Article 1198001.

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501–532.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264.

Christner, N., Pletti, C., & Paulus, M. (2020). Emotion understanding and the moral self-concept as motivators of prosocial behavior in middle childhood. *Cognitive Development*, *55* Article 100893.

Coates, D. J., & Tognazzini, N. A. (Eds.). (2012). *Blame: Its nature and norms*. Oxford University Press.

Conte, R., Andrighetto, G., & Campenni, M. (2013). *Minding norms: Mechanisms and dynamics of social order in agent societies*. Oxford University Press.

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

Cushman, F. (2013). The role of learning in punishment, prosociality, and human uniqueness. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its evolution.* (pp. 333–372). MIT Press.

Cushman, F., Kumar, V., & Railton, P. (Eds.). (2017). Moral learning [Special issue]. *Cognition*, *167*, 1–282.

Cusimano, C., Thapa, S., & Malle, B. F. (2017). Judgment before emotion: People access moral evaluations faster than affective states. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1848–1853). Cognitive Science Society.

de Waal, F. B. M. (2014). Natural normativity: The "is" and "ought" of animal behavior. *Behaviour, 151*(2–3), 185–204.

Dhillon, A., & Nicolò, A. (2023). Moral costs of corruption: A review of the literature. In K. Basu & A. Mishra (Eds.), *Law and economic development: Behavioral and moral foundations of a changing world* (pp. 93–129). Springer International Publishing.

Drew, P. (1998). Complaints about transgressions and misconduct. *Research on Language & Social Interaction*, *31*(3–4), 295–325.

Ellemers, N., Pagliaro, S., & Nunspeet, F. van (Eds.). (2023). *The Routledge international handbook of the psychology of morality*. Taylor & Francis.

Enke, B. (2023). Market exposure and human morality. *Nature Human Behaviour*, *7*(1), 134–141.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.

Fike, R. (2023). Do disruptions to the market process corrupt our morals? *Review of Austrian Economics*, *36*(1), 99–106.

Fornari, L., Ioumpa, K., Nostro, A. D., Evans, N. J., De Angelis, L., Speer, S. P. H., Paracampo, R., Gallo, S., Spezio, M., Keysers, C., & Gazzola, V. (2023). Neuro-computational mechanisms and individual biases in action-outcome learning under moral conflict. *Nature Communications*, *14*(1), Article 1.

Friedman, M. (2013). How to blame people responsibly. *Journal of Value Inquiry*, *47*(3), 271–284.

Galliott, J., Macintosh, D., & Ohlin, J. D. (Eds.). (2021). *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare*. Oxford University Press.

Genschow, O., Cracco, E., Schneider, J., Protzko, J., Wisniewski, D., Brass, M., & Schooler, J. W. (2023). Manipulating belief in free will and its downstream consequences: A meta-analysis. *Personality and Social Psychology Review*, *27*(1), 52–82.

Gobodo-Madikizela, P. (2002). Remorse, forgiveness, and rehumanization: Stories from South Africa. *Journal of Humanistic Psychology*, *42*(1), 7–32.

Gollan, T., & Witte, E. H. (2008). "It was right to do it, because . . ." *Social Psychology*, *39*(3), 189–196.

Gomez-Lavin, J., & Prinz, J. (2019). Parole and the moral self: Moral change mitigates responsibility. *Journal of Moral Education*, *48*(1), 65–83.

Goodwin, G. P., & Gromet, D. M. (2014). Punishment. *Cognitive Science*, *5*(5), 561–572.

Grappi, S., Romani, S., & Bagozzi, R. P. (2013). Consumer response to corporate irresponsible behavior: Moral emotions and virtues. *Journal of Business Research*, *66*(10), 1814–1821.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol 3: The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). MIT Press.

Guglielmo, S. (2015). Moral judgment as information processing: An integrative review. *Frontiers in Psychology*, *6*, Article 1637.

Guglielmo, S., & Malle, B. F. (2017). Information-acquisition processes in moral judgments of blame. *Personality and Social Psychology Bulletin*, *43*(7), 957–971.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLoS ONE*, *14*(3), Article e0213544.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford University Press.

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, *90*, Article 104512.

Hindriks, F. (2008). Intentional action and the praise-blame asymmetry. *Philosophical Quarterly*, *58*(233), 630–641.

Hitlin, S. (2011). Values, personal identity, and the moral self. In S. J. Schwartz, K. Luyckx, & V. L. Vignoles (Eds.), *Handbook of identity theory and research* (pp. 515–529). Springer.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6.

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. J. Gregor, Ed. & Trans.). Cambridge University Press. (Original work published 1785)

Kohlberg, L. (1981). *Essays on moral development*. Harper & Row.

Krueger, F., Hoffman, M., Walter, H., & Grafman, J. (2013). An fMRI investigation of the effects of belief in free will on third-party punishment. *Social Cognitive and Affective Neuroscience*, *9*(8), 1143–1149.

Kupfer, T. R., & Giner-Sorolla, R. (2017). Communicating moral motives: The social signaling function of disgust. *Social Psychological and Personality Science*, *8* (6), 632–640.

Ladak, A., Loughnan, S., & Wilks, M. (2023). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*(1), 27–34.

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2016). Unintended, but still blameworthy: The roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cognition & Emotion*, *30*(7), 1271–1288.

Leary, M. R., & Tangney, J. P. (Eds.). (2012). *Handbook of self and identity* (2nd ed.). Guilford Press.

Leary, M. R., Tangney, J. P., & Leary, M. (Eds.). (2003). *Handbook of self and identity* (Paperback ed). Guilford Press.

Lengersdorff, L. L., Wagner, I. C., Lockwood, P. L., & Lamm, C. (2020). When implicit prosociality trumps selfishness: The neural valuation system underpins more optimal choices when learning to avoid harm to others than to oneself. *Journal of Neuroscience*, *40*(38), 7286–7299.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2011). *Robot ethics: The ethical and social implications of robotics*. MIT Press.

Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243–256.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Malle, B. F., Guglielmo, S., Voiklis, J., & Monroe, A. E. (2022). Cognitive blame is socially shaped. *Current Directions in Psychological Science*, *31*(2), 169–176.

Malle, B. F., & Scheutz, M. (2019). Learning how to behave: Moral competence for social robots. In O. Bendel (Ed.), *Handbuch Maschinenethik [Handbook of machine ethics]*. Springer.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI'15* (pp. 117–124). ACM.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3–25). Academic Press.

Marazziti, D., Baroni, S., Landi, P., Ceresoli, D., & Dell'Osso, L. (2013). The neurobiology of moral sense: Facts or hypotheses? *Annals of General Psychiatry*, *12*(1), Article 6.

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133–143.

May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.

May, J. (2023). Moral rationalism on the brain. *Mind & Language*, *38*(1), 237–255.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass*, *4*(5), 344–357.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. MIT Press.

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 3. The neuroscience of morality* (pp. 81–92). MIT Press.

Mill, J. S. (1998). *Utilitarianism*. Oxford University Press.

Misselhorn, C. (2018). Artificial morality: Concepts, issues and challenges. *Society*, *55*(2), 161–169.

Moisuc, A., & Brauer, M. (2019). Social norms are enforced by friends: The effect of relationship closeness on bystanders' tendency to confront perpetrators of uncivil, immoral, and discriminatory behaviors. *European Journal of Social Psychology*, *49*(4), 824–830.

Molho, C., Tybur, J. M., Van Lange, P. A. M., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, *11*, Article 34.

Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology*, *11*(2), 99–111.

Monroe, A. E., Brady, G. L., & Malle, B. F. (2017). This isn't the free will worth looking for: General free will beliefs do not influence moral judgments, agent-specific choice ascriptions do. *Social Psychological and Personality Science*, *8*(2), 191–199.

Monroe, A. E., Dillon, K. D., Guglielmo, S., & Baumeister, R. F. (2018). It's not what you do, but what everyone else does: On the role of descriptive norms and subjectivism in moral judgment. *Journal of Experimental Social Psychology*, *77*, 1–10.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, *27*, 100–108.

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, *1*(2), 211–224.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, *116*(2), 215–236.

Murray, S., O'Neill, K., Bridges, J., Sytsma, J., & Irving, Z. (2024). Blame for Hum(e)an beings: The role of character information in judgments of blame. *Social Psychological and Personality Science*. https://doi.org/10.1177/19485506241233708

Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy*, *31*(1), 214–242.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561–584.

Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if 'my brain made me do it': People's intuitions about free will and neuroscientific prediction. *Cognition*, *133*(2), 502–516.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*(1), 96–125.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science*, *331*(6023), 1401–1403.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, *41*(4), 663–685.

Olson, E. T. (1999). *The human animal: Personal identity without psychology*. Oxford University Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716.

Panagopoulou-Koutnatzi, F. (2014). The practice of naming and shaming through the publicizing of "culprit" lists. In C. M. Akrivopoulou & N. Garipidis (Eds.), *Human rights and the impact of ICT in the public sphere: Participation, democracy, and political autonomy* (pp. 145–155). Information Science Reference/IGI Global.

Parfit, D. (1992). *Reasons and persons*. Clarendon Press.

Pereira, L. M., & Lopes, A. B. (2020). *Machine ethics: From machine morals to the machinery of morality*. Springer Nature.

Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise. *Psychological Science*, *14*(3), 267–272.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9*(1), 29–43.

Prinz, J., & Nichols, S. (2010). Moral emotions. In J. M. Doris (Ed.), *The moral psychology handbook* (pp. 111–146). Oxford University Press.

Przepiorka, W., & Berger, J. (2016). The sanctioning dilemma: A quasi-experiment on social norm enforcement in the train. *European Sociological Review*, *32*(3), 439–451.

Riordan, C. A., Marlin, N. A., & Kellogg, R. T. (1983). The effectiveness of accounts following transgression. *Social Psychology Quarterly*, *46*(3), 213–219.

Robbins, P., & Alvear, F. (2023). Deformative experience: Explaining the effects of adversity on moral evaluation. *Social Cognition*, *41*(5), 415–446.

Royzman, E. B., Goodwin, G. P., & Leeman, R. F. (2011). When sentimental rules collide: "Norms with feelings" in the dilemmatic context. *Cognition*, *121*(1), 101–114.

Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral-conventional distinction. *Cognition*, *112*(1), 159–174.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn. (2015). Towards safe and trustworthy social robots: Ethical challenges and practical issues. In A. Tapus,

E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social Robotics: 7th International Conference, ICSR 2015, Proceedings* (pp. 584–593). Springer International Publishing.

Sattler, S., Dubljević, V., & Racine, E. (2023). Cooperative behavior in the workplace: Empirical evidence from the agent-deed-consequences model of moral judgment. *Frontiers in Psychology*, *13*, Article 1064442.

Sauer, H. (2011). Social intuitionism and the psychology of moral reasoning. *Philosophy Compass*, *6*(10), 708–721.

Sauer, H. (2012). Morally irrelevant factors: What's left of the dual process-model of moral cognition? *Philosophical Psychology*, *25*(6), 783–811.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*(7), 1307–1351.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). Academic Press.

Semin, G. R., & Manstead, A. S. R. (1983). *The accountability of conduct: A social psychological analysis*. Academic Press.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411.

Shank, D. B., Kashima, Y., Peters, K., Li, Y., Robins, G., & Kirley, M. (2019). Norm talk and human cooperation: Can we talk ourselves into cooperation? *Journal of Personality and Social Psychology*, *117*(1), 99–123.

Sharkey, A., & Sharkey, N. (2010). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, *14*(1), 27–40.

Silver, E., & Silver, J. R. (2021). Morality and self-control: The role of binding and individualizing moral motives. *Deviant Behavior*, *42*(3), 366–385.

Sio, F. S. de, & Wynsberghe, A. van. (2016). When should we use care robots? The nature-of-activities approach. *Science and Engineering Ethics*, 22, 1745–1760.

Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass*, *5*(2), 199–212.

Sorial, S. (2016). Performing anger to signal injustice: The expression of anger in victim impact statements. In C. Abell, J. Smith, C. Abell, & J. Smith (Eds.), *The expression of emotion: Philosophical, psychological and legal perspectives* (pp. 287–310). Cambridge University Press.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261–302). Academic Press.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 1–25.

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, *12*(4), 551–560.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), Article 363, 1–27.

Suls, J. (Ed.). (2014). *Psychological perspectives on the self: Vol. 4. The self in social perspective*. Psychology Press.

Sztompka, P. (2019). Trust in the moral space. In M. Sasaki (Ed.), *Trust in contemporary society* (Vol. 42, pp. 31–40). Koninklijke Brill NV.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372.

Tedeschi, J. T., & Reiss, M. (1981). Verbal strategies as impression management. In C. Antaki (Ed.), *The psychology of ordinary social behaviour* (pp. 271–309). Academic Press.

Tesser, A. (1988). Towards a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). Academic Press.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Tiberius, V. (2015). *Moral psychology: A contemporary introduction*. Routledge/ Taylor & Francis Group.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2021). Implementations in machine ethics: A survey. *ACM Computing Surveys*, *53*(6), Article 132, 1–38.

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

Vila-Henninger, L. A. (2021). A dual-process model of economic behavior: Using culture and cognition, economic sociology, behavioral economics, and neuroscience to reconcile moral and self-interested economic action. *Sociological Forum*, *36*(Suppl. 1), 1271–1296.

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49–54.

Vonasch, A. J., Baumeister, R. F., & Mele, A. R. (2018). Ordinary people think free will is a lack of constraint, not the presence of a soul. *Consciousness and Cognition*, *60*, 133–151.

Wilson, J. Q. (1993). The moral sense. *American Political Science Review*, *87*(1), 1–11.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301.

Wylie, R. C. (1979). *The self-concept* (Vol. 2). University of Nebraska Press.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112.

Yang, Q., Yan, L., Luo, J., Li, A., Zhang, Y., Tian, X., & Zhang, D. (2013). Temporal dynamics of disgust and morality: An event-related potential study. *PLoS ONE*, *8*(5), Article e65094.

Yoder, K. J., & Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: A high-density ERP study. *Neuropsychologia*, *60*, 39–45.

Young, L., Chakroff, A., & Tom, J. (2012). Doing good leads to more good: The reinforcing power of a moral self-concept. *Review of Philosophy and Psychology*, *3*(3), 325–334.

Yucel, M., & Vaish, A. (2021). Eliciting forgiveness. *WIREs Cognitive Science*, *12*(6), Article e1572.

Zakharin, M., Curry, O. S., Lewis, G., & Bates, T. C. (2024). Modular morals: The genetic architecture of morality as cooperation. *PsyArXiv*. https://doi.org/10.31234/osf.io/srjyq

Zeelenberg, M., Breugelmans, S. M., & de Hooge, I. E. (2012). Moral sentiments: A behavioral economics approach. In A. Innocenti & A. Sirigu (Eds.), *Neuroscience and the economics of decision making* (pp. 73–85). Routledge/Taylor & Francis Group.

Zhang, Q., Wallbridge, C. D., Jones, D. M., & Morgan, P. (2021). The blame game: Double standards apply to autonomous vehicle accidents. In N. Stanton (Ed.), *Advances in human aspects of transportation* (pp. 308–314). Springer International Publishing.