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.

31

Computational Approaches to Morality

Paul Bello and Bertram F. Malle

31.1 Introduction

Morality regulates individual behavior so that it complies with community interests (Curry et al., 2019; Haidt, 2001; Hechter & Opp, 2001). Humans achieve this regulation by motivating and deterring certain behaviors through the imposition of norms – instructions of how one should or should not act in a particular context (Fehr & Fischbacher, 2004; Sripada & Stich, 2006) – and, if a norm is violated, by levying sanctions (Alexander, 1987; Bicchieri, 2006). This chapter examines the mental and behavioral processes that facilitate human living in moral communities and how these processes might be represented computationally and ultimately engineered in embodied agents.

Computational work on morality arises from two major sources. One is empirical moral science, which accumulates knowledge about a variety of phenomena of human morality, such as moral decision making, judgment, and emotions. Resulting computational work tries to model and explain these human phenomena. The second source is philosophical ethics, which has for millennia discussed moral principles by which humans *should* live. Resulting computational work is often labeled *machine ethics*, which is the attempt to create artificial agents with moral capacities reflecting one or more of the ethical theories. A brief discussion of these two sources will ground the subsequent discussion of computational morality.

31.2 A Map of Moral Phenomena

A variety of moral phenomena have been studied in moral science, and Figure 31.1 provides a map to distinguish them.

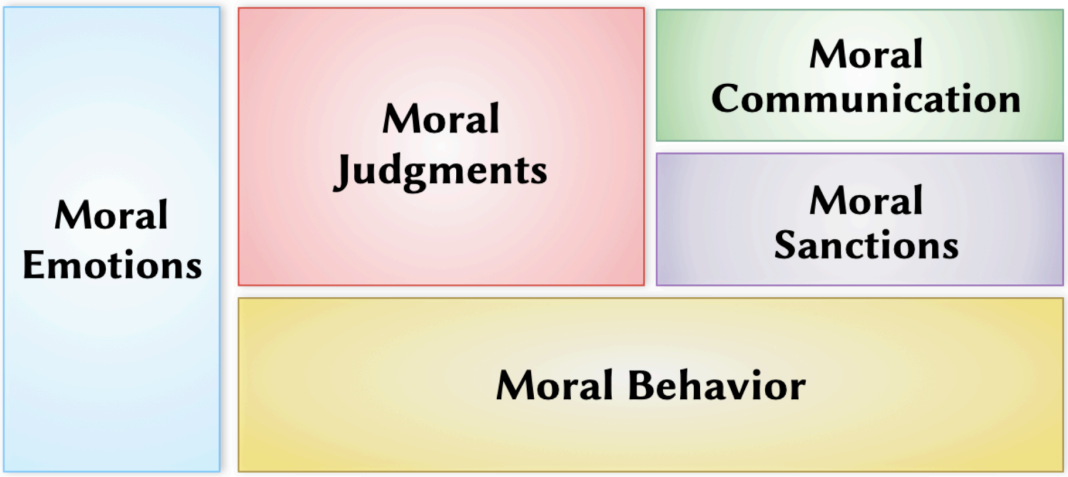

Figure 31.1 Five major moral phenomena: moral behavior (including moral decision making), moral judgments, moral emotions, moral sanctions, and moral communication.

31.2.1 Five Moral Phenomena

Moral behavior includes, first, intentional actions that conform to, violate, or exceed moral standards. These actions rely on *moral decisions* – understood as conscious choices among paths of action to comply with moral standards (Kohlberg, 1984; Turiel, 2002). Second, many morally significant behaviors are unintentional, such as recklessness, preventable accidents, or unintended side effects (Laurent et al., 2016; Weiner, 2001). Computational models, however, have focused strongly on moral decision making and action.

In contrast to the agent perspective of moral behavior, *moral judgments* are made from an observer perspective: people appraise an event, behavior, or person in light of moral standards. Research has identified at least four classes of moral judgment, which differ primarily in what they judge and what information they process (Cushman, 2008; Malle, 2021) – distinctions that have important implications for computational modeling. In brief, *evaluations* can be made about intentional and unintentional behavior, persons, events – anything that could be compared to a normative standard; and they broadly assess how good or bad the judged object is. *Norm judgments* focus on actions and declare whether a given action falls under a norm – whether it is prohibited, obligated, or permitted. *Moral wrongness judgments* declare that an action violated a relevant norm, but they are also sensitive to the person's reasons for acting, and some reasons can provide a justification, making the action no longer wrong. Finally, *blame judgments* criticize a person for an intention, action, or unintentional outcome. Of all moral judgments, blame integrates the most information – about the norms that were violated, the agent's causal involvement, whether the agent acted intentionally or not; if so, what the agent's reasons were for acting; and if not, whether the agent could have and should have prevented the unintentional event (Alicke, 2000; Malle et al., 2014; Shaver, 1985). In contrast to computational models of moral decision making, models of moral judgment are quite rare, but a rich set of problems is ready to be tackled.

A third prominent moral phenomenon is *moral emotions*. After a long period in which morality was predominantly treated as a cognitive phenomenon, the early twenty-first century saw a rise of interest in morality's emotional aspects – moral emotions as results of moral judgment (e.g., sympathy and anger; Weiner, 2001), as capacities of regulation (e.g., guilt, shame, and empathy; Eisenberg, 2000; Tangney & Dearing, 2002), as causes of moral judgment (Alicke, 2000; Prinz, 2006), or as competing with moral reasoning (Cushman et al., 2010). Computational work on emotions has only begun to emerge in the last decade.

Whereas moral judgments and emotions are typically in the observer's head, *moral sanctions* are social acts that express a judgment or emotion and attempt to regulate the violator's future behavior, most prominently in the form of punishment. *Moral communication* encompasses a variety of social acts, including moral praise and criticism, justification and apology, statements of remorse and forgiveness. Both sanctions and communication, as social behaviors, have been hardly modeled computationally, but a rich array of opportunities awaits.

31.2.2 Norms as a Foundation

A growing consensus in empirical moral science is that the above phenomena are *moral* not by virtue of a special brain circuit or unique cognitive mechanism, but by virtue of applying fundamental processes of human decision making, judgment, emotions, and so on to moral matters – which are matters of human behavior governed by moral norms (e.g., Bartels et al., 2015; Bicchieri, 2006). Without knowledge of the relevant norms of a community, people could not judge how bad a certain outcome is; decide what is the right or wrong decision; or even know whether to feel sad or resentful. Thus, any computational model of human morality must incorporate an analysis of norms.

What are norms? The following working definition of norms integrates a number of previous proposals (Bicchieri, 2006; Brennan et al., 2013; Cialdini et al., 1991; Gibbs, 1965; Malle et al., 2017):

> *A norm is an instruction, in a given community, to (not) perform a behavior in a given context, provided that a sufficient number of individuals in the community*
> *(i) demand, to a certain degree, of each other to follow the instruction and*
> *(ii) do in fact follow this instruction.*

This definition has five components, $< S, C, A, D_f, P >$. A norm N always exists relative to a social community S in which that norm holds (Hechter & Opp, 2001; Sachdeva et al., 2011). It operates on a particular behavior B (thus being more specific than, say, abstract values like freedom or justice), and it operates in a given context C (Aarts & Dijksterhuis, 2003; Bartels et al., 2015). Further, the norm comes with a deontic modality D (e.g., prescription, prohibition), which has a force parameter f, expressing the strength of the norm (Heider, 1958, Chapter 8; Malle, 2020). Finally, a norm has a prevalence P, which indicates how consistently community members adhere to the norm (Bicchieri, 2006; Cialdini et al., 1991). These properties of norms will re-emerge in a number of the subsequent sections.

31.3 Philosophical Ethics

Artificial moral agents must be capable of acting, and more importantly, of choosing the "right" action to perform in a given situation. This will often involve assessing the moral goodness or badness of actions or outcomes: an activity associated with consequentialist and deontological ideas in philosophical ethics. What follows are brief overviews of two popular schools of ethical thought: deontology and consequentialism, along with some of their challenging implications. Where appropriate, computational models or analyses will be mentioned. Alternative ethical theories such as virtue ethics are only beginning to be explored from a computational perspective (Govindarajulu et al., 2019; Howard & Muntean, 2017), but aspects of its core idea of moral habit learning have recently resurfaced in reinforcement learning frameworks of moral decision making, discussed in more detail in Section 31.4.2.2.

31.3.1 Deontological Ethics

Broadly, deontological ethics concerns itself with the moral status and motivation for individual acts. Most famously, this is illustrated in the moral philosophy of Immanuel Kant. Kant's two primary imperatives evaluate the status of a moral rule by whether (1) the reasoner would want every other agent to abide by it, and (2) whether or not it uses other autonomous, rational agents as pure means to achieve a desired end. Kant's philosophy is framed against the background of the will acting out of duty to the moral law. This may be understood as an agent who desires to keep the moral law, and recognizes the rational benefit in doing so as opposed to giving in to other irrational or nonrational inclinations. A discussion of whether machines might ever be Kantian in this sense can be found in Powers (2006). Recent empirical results lend support to the idea that universalization, the principle stated in the first of Kant's two primary imperatives, is consistent with spontaneous moral judgments made by adults (Levine et al., 2020).

Deontology is often associated solely with Kant's moral philosophy, but it also finds expression in the contractualism of John Rawls that takes moral acts to be the ones that we would all agree ought to be done if we were ignorant of our place in the social hierarchy when performing them (Scanlon, 1998). A computational example of Rawlsian ethical decision making can be found in Leben (2017), with some preliminary empirical evidence for contractarian judgments in humans to be found in Levine et al. (2018). Also in this group is the theory of W. D. Ross (1930) that considers maximizing the good as one of a plurality of prima facie duties, each of which can outweigh the others in different situational contexts. A well-known implementation of prima facie duties is the MedEthEx system (Anderson & Anderson, 2006), which uses expert bioethical analysis of dilemma cases to seed a case-based reasoning system that can offer advice on novel ethical cases. From the cases and the expert decisions, MedEthEx attempts to learn how to order priorities for a set of duties: Respect for Autonomy, Nonmaleficence, Beneficence, and Justice, which are then used in searching for the best action to perform in novel cases that are deemed similar to those in the database. Both natural law and Divine command theories are also deontological in nature, grounding right action in conformance to God-given moral imperatives (Quinn, 1978) such as the Judeo-Christian Ten Commandments, the latter having been given an initial logical formalization and computational treatment (Bringsjord & Taylor, 2012).

31.3.2 Consequentialist Ethics

Consequentialism or teleological ethics, broadly speaking, is the idea that actions are to be evaluated solely in terms of their outcomes or their "goods." This is at odds with deontological theories that evaluate action in terms of what is right to do. The most well-known variety of consequentialism is act utilitarianism, due to J. S. Mill, who fixes value on pleasure, equating utility with the amount of pleasure less the amount of pain experienced by individuals. Applied to ethics, this takes moral decision making to be the process of determining the action that results in the most utility for the greatest number of people.

Other conceptions of value can be found in rule utilitarianism and preference utilitarianism. Rule utilitarianism takes right action to be conformant to rules that lead to the greatest good. Rule utilitarianism is exceptionless, and inherits many of the same counterexamples that plague deontological frameworks. Exception-tolerant versions of rule utilitarianism have been developed, but they have been criticized on the grounds of collapsing into act utilitarianism when the number of exceptions becomes large. A discussion of rule-utilitarianism and its advantages for building artificial moral agents can be found in Bauer (2020). Preference utilitarianism takes actions to be morally right that best fulfill the preferences (i.e., interests, desires) of others. Naturally, questions arise as to how to weigh the preferences of those involved in a moral decision if they should conflict, introducing a new set of ethical challenges. Recently, preference utilitarianism has been promoted in AI ethics research under the banner of value alignment (Russell, 2019) as a way to prevent threats to humans from superintelligent machines, should we ever be successful at engineering them. Preference utilitarianism is not without its problems. If the vast majority of a group desire that members of another group die, and this wins the competition of preference satisfaction, preference utilitarianism would recommend a machine to engage in extermination.

31.3.3 Computational Challenges

Each of the ethical theories has difficulties that have been explored by philosophers, legal scholars, and others (Brundage, 2014). All theories share the serious implementation issue of how to frame a moral decision problem. For example, how many agents should a Kantian, Rawlsian, or utilitarian algorithm take into consideration while computing aggregate welfare? The normative frameworks themselves are silent, leaving the modeler to introduce extra-normative constraints, possibly from psychology, to help guide computation. Taken at face value, utilitarian theories impose an enormous epistemic burden in the form of thinking about vast numbers of agents and the factors that impact their collective welfare, under enormous uncertainty over virtually infinite time horizons. Low-probability, high-value states are washed out in utility computations. A simplified analysis of the scaling difficulties for both utilitarian and (some) deontological theories is given in Brundage (2014). Apart from scaling difficulties, both deontological and (some) utilitarian theories face the possibility of impasses in inference. For example, preference-based utilitarian algorithms may encounter a situation where agents under consideration have incompatible preferences that must somehow be resolved. However, the type of impasse most thoroughly explored in the literature is that of conflict between norms, which has become something of a research area unto itself among deontologists (see Section 31.4.1.4).

31.4 Moral Decision Making

Computational models of moral decision making have been inspired by philosophical ethics to build general-purpose algorithms for selecting ethical actions and by descriptive work in the cognitive and social sciences of normative behavior. Many of the resulting efforts have taken rule-based approaches, often grounded in formal logic. These will be reviewed first, followed by brief discussions of case-based reasoning, recent reinforcement learning frameworks, and the cognitive science of moral dilemmas.

31.4.1 Rule-Based Approaches

Formal logic has had an outsized influence on the development of rule-based approaches to moral decision making. In a very early paper, Shoham and Tennenholtz (1995) describe well-known problems with distributed collections of robotic agents trying to co-exist in an environment. Prior approaches to handling situations where collective behavior led to poor outcomes (e.g., collisions between moving robots) relied upon agent-to-agent communication and negotiation techniques. Doing so incurs a large computational burden on each agent, which can be reduced if all agents follow social laws, leading to a proposal for a language of social constraints to be used by multi-agent systems that is a precursor to richer and more explicit specification of norms. A start on such explicit specification was outlined by several authors who insisted (1) that norms were not to be modeled as hard constraints and (2) that agents are to be "autonomous" they should have the opportunity to learn, reason over, negotiate, accept, reject, abide by, and violate norms in order to have "some degree of control" over their actions (Castelfranchi et al., 2000).

31.4.1.1 Deontic Logic

These needs found partial satisfaction in the theoretical development and computational treatment of various modal logics of belief, desire, intention, and obligation. In particular, the development of deontic logic (Von Wright, 1951), which captures the relationships between obligation, forbiddance, and permission, has been an inspiration to researchers in the multi-agent systems communities who seek to build norm-governed, agent-based simulations. Deontic logics are differentiated in two ways: first, by different sets of axioms that provide inference rules to transform premises into justified conclusions; second, by the syntax and semantics of deontic terms such as obligation, forbiddance, and permission, resulting in different semantic machinery for evaluating deontic inferences. For example, obligations, permissions, and forbiddances are typically represented as $\mathbf{O}(\phi)$, $\mathbf{P}(\phi)$, and $\mathbf{F}(\phi)$, where ϕ is a well-formed formula. However, the exploration of various well-known deontic paradoxes (Carmo & Jones, 2002; van der Torre & Tan, 1997) has led to the development of dyadic deontic logics (Prakken & Sergot, 1997), where basic syntax for deontic terms looks like $\mathbf{O}(\phi \mid \alpha)$, meaning that if α then ϕ is obligated. Thus, certain deontic logics capture the context specificity of norms that has been established empirically (Aarts & Dijksterhuis, 2003). Semantic differences between standard and dyadic deontic logics are beyond the scope of this chapter, but interested readers are directed to the

more thorough explanations found in Goble (2003). For a discussion of a highly expressive family of multi-operator deontic logics and an automated reasoning technology in which they have been encoded, see Govindarajulu et al. (2019).

31.4.1.2 Belief-Desire-Intention Frameworks

Agents are not only sets of obligations, but rather have beliefs, desires, and intentions (BDI) that guide their practical reasoning and facilitate action. Much of the foundational work on normative multi-agent systems leans heavily on the view of rational agency or practical reasoning promoted by philosophers such as Bratman (1987), and formalized by Rao and Georgeff (1991). For a review, see Meyer et al. (2015). A typical BDI agent maintains a set of beliefs, desires, and intentions that are reasoned over, along with a plan library. Practical reasoning begins with perception that updates the current set of the agent's beliefs, examines the current deliberation, and looks at the top of the stack of active intentions. It searches its plan library for an action plan with a post-condition (outcome) that matches the content of the intention. Candidate plans are then winnowed down by matching the set of necessary pre-conditions for each against the agent's current set of beliefs about the state of the world and how it meets the preconditions. The contents of plans that survive the matching process become intended actions, which are then executed (Bordini et al., 2007), while unfulfilled intentions are kept in a stack. The original BDI framework was developed as a logic of rational agency, but a number of BDI logic-conformant implementations have been successfully used to solve real-world problems (Dastani, 2008; D'Inverno et al., 2004; Rao, 1996). Normative extensions to the BDI framework have been used in multi-agent simulations; however, robust implementations in real-world agents (e.g., swarm robots) have not been attempted.

Logical representations of norms and accompanying BDI-style agents come with the very obvious advantage that their computations are inspectable, facilitating attempts at building artificial moral agents capable of explaining their decisions. In principle, computational logics offer certain attractive guarantees regarding the correctness of the conclusions that they draw. On the other hand, they have a number of disadvantages as well. BDI and deontic logics are modal logics as opposed to the more well-known first-order logic that has been a staple of AI research since the inception of the field. Modal logics are substantially more difficult to automate, with only a handful of very recent attempts offering a path forward (Benzmüller, 2019). Effectively, all of the work in the BDI tradition discussed here uses encoding schemes that capture at most a fragment of BDI logic in first-order logic in order to keep computation tractable. Importantly, these fragments do not typically allow for nesting of modal operators, thus leaving beliefs about beliefs (for example) out of play. Nested expressions are critical for theory of mind, or the ability for one agent to reason about the mental states of others – a central ability for complex moral judgment.

31.4.1.3 Focus on Norms: The EMIL-I-A Architecture

As a matter of psychological reality, norms are central to human moral decision making. One of the most elaborate models of such norm-based moral decision making is the EMIL family of architectures, which consists of implemented computational architectures that have been used to explore how self-interested agents might achieve significant degrees of co-operation in a social

community. These moral decisions are modeled as being deeply guided by social and moral norms. One central tenet of the EMIL-I-A architecture, which is a member of the EMIL family, is that norms are not just external forces in the community but can become internalized in an agent (Andrighetto et al., 2010a). In this process, the cognitive maintenance of a norm becomes detached from external rewards and punishment, eventually resulting in often automatic behavioral responses that are norm-conforming while still allowing the agent deliberation and control if necessary. Simulation studies of iterated Prisoner's dilemma games have shown that agents capable of internalizing norms, in contrast with traditional strategic (decision-theoretic) agents, maintain co-operation even when punishment is rare or unlikely (Realpe-Gómez et al., 2018).

Enabling such internalization of norms, the EMIL-I-A architecture (where the I-A stands for Internalizing Agent) embeds norm representations within a BDI framework, with intimate interactions between norm recognition, normative beliefs, and normative goals Andrighetto et al. (2010) . The underlying cognitive model of norms draws on the definition of a norm as being a prescription that members of a society generally comply with (Ullmann-Margalit, 1977); but Andrighetto and colleagues added the proviso that when a prescription spreads within a society, it gives raise to shared normative beliefs and goals among its members. "Normative beliefs" are mental representations that a given action has a normative status (i.e., being obligated, prohibited, etc.) for a given set of agents in a particular context. The authors complement these normative beliefs with "normative goals," defined as "the will to perform an action because and to the extent that this is believed to be prescribed by a norm" (Andrighetto et al., 2010a, p. 327). Thus, EMIL-I-A addresses three features of the earlier presented working definition of human norms (see Section 31.2.2): deontic modality/status (*D*), context-specificity (*C*), and community-relativity (*S*).

Another concept that connects Andrighetto et al.'s work with the psychological reality of humans norms is their notion of norm "salience" (Andrighetto, Brandts, et al., 2013; Conte et al., 2013). In earlier work, salience was defined as a norm's "degree of activation," in close affinity to social psychological work that showed norms guide behavior when they are, at that moment, on the agent's mind (Aarts & Dijksterhuis, 2003; Lindenberg, 2013). Then salience expanded to "the degree of activity and importance of a norm" (Andrighetto et al., 2010b, p. 329). And most recently, it became "the perceived degree of importance and strength of a norm" (Andrighetto, Castelfranchi, et al., 2013, p. 145). These components of salience seem to map onto the deontic force parameter $D_f$and the prevalence parameter *P*, respectively, in Section 31.2.2's working definition of norms, although combined into a single EMIL-I-A parameter.

In sum, few models of norm-conforming decision making are as well aligned with concepts of human moral psychology as EMIL-I-A, and the type of multi-agent simulations used to explore EMIL-I-A's capabilities are valuable (e.g., Realpe-Gómez et al., 2018). However, they do fall short of what would be required of a robotic system interacting with people in the real world. .

31.4.1.4 Norm Conflict Resolution

Considerable efforts have gone into computational solutions to one of the core features of moral decision making: that norms can conflict and such conflicts must be resolved. An extensive survey by Santos et al. (2017) catalogues over fifty approaches to detecting and/or resolving norm conflicts in multi-agent systems (MAS). Outside the MAS literature, several other

approaches to norm conflict resolution exist. Thagard (1998) proposed multiple-constraint satisfaction processes ("coherence") among competing normative propositions. Guarini (2007) critiques this approach at multiple levels, including its lack of psychological realism and difficulties of using coherence criteria for the justification of moral claims. Numerous argumentation frameworks (Dung, 1995) have been developed to resolve conflicts between plans, when each plan favors different goals and norms. Competing plans are evaluated by aggregating arguments for (e.g., norms adhered to) and arguments against them (e.g., norms violated), heeding the counted number of fulfilled goals and norms, as well as preference orderings among them (Shams et al., 2020). A strength of this framework is that it delivers justifications for the reasoner's moral decisions, something that is increasingly recognized as essential in moral communication. A weakness is that its resolution criteria – counts of fulfilled goal/norms and preference orderings among them – can contradict one another, demanding yet new conflict resolution.

Conflicts among resolution criteria may be avoidable with a continuous deontic force parameter for norms (akin to D_f), such as used by Kasenberg & Scheutz (2018). The model relies on linear temporal logic and Markov decision processes to represent acts and consequences probabilistically, and it minimizes a cost function in which violating more important norms accrues higher costs. Strengths of the proposal include the ability to handle uncertainty about consequences and the ability to provide justifications for the decisions. A weakness, which the authors admit, is that the mathematical machinery does not scale well to even moderate numbers of norms. Scalability is also a challenge for approaches that use formal verification methods to select least norm-violating plans among multiple conflicting ones (e.g., (Dennis et al., 2016)).

There is surprisingly little empirical research on human norm conflict resolution (Broeders et al., 2011; Holyoak & Powell, 2016). The considerable computational work on this topic might inspire new experiments and psychological theories, which in turn may help refine the computational models.

31.4.2 Other Approaches to Moral Decision Making

While logical reasoning and traditional approaches to planning and acting have been central to computational modeling of moral decision making in AI, they have not been the only avenues explored. Much of applied ethics and the law focuses on the analysis of cases, and how to apply judgments produced in the past to a current case. More recently, learning-based approaches to moral decision making have been explored in cognitive science primarily using reinforcement learning as a unifying framework. Both of these approaches are briefly explored next.

31.4.2.1 Moral Decision Making Using Cases

The MedEthEx system uses inductive logic programming to first extract principles from cases previously judged by expert ethicists and then test them on yet further cases until a set of principles are generated that best cover expert judgment across the widest number of cases (Anderson et al., 2006). Being an implementation of prima facie duties, the cases were marked up with tuples, consisting of each duty and an integer representing how violated or satisfied the duty was judged to be by the expert analysis. This is a fascinating mix of learning with more traditional rule-based reasoning, but the approach suffers from reduction to integers of abstract

duties such as “beneficence,” which are rather richly textured and difficult to apply to concrete actions in context.

Other case-based approaches eschew generalizing over larger numbers of cases and specifically acknowledge that the abstract and complex nature of moral principles are still beyond comprehensibility for machines. The combination of Truth-Teller and SIROCCO, developed specifically to be ethical decision-aides, retrieve past cases or newly generated hypothetical cases to compare against a current problem of interest (McLaren, 2006). Truth-Teller focuses primarily on case comparison, where each case details a dilemma in which a choice between performing an action or not is supported by respective sets of reasons. Comparison is performed at the level of reason content, meaning that reasons can be stronger or weaker than other reasons, or not comparable at all. SIROCCO was developed to draw cases out of memory to feed the case-comparison process described previously. Interestingly, the retrieval process is two-step, with surface-level comparisons computed first before deep structure mapping is applied to more promising candidates. This two-stage process is reminiscent of the MAC/FAC model of analogical retrieval pioneered by Forbus et al. (1995), which has substantial empirical support. A combination of Truth-Teller and SIROCCO adjudicate conflicting reasons before providing an analysis to support the human decision maker.

Finally, cases have been employed to examine a fundamental question in the study of philosophical ethics: whether there are general ethical principles at all. One can imagine that, in the limit, every morally charged situation or case has an analysis all of its own. This is a highly oversimplified description of moral particularism, which assumes, in contrast to moral generalism, that there are no abstract principles that exist across cases (Dancy, 2009). One corollary of particularism is that moral case classification ought to be impossible, as should generalization to new cases, since both classification and generalization rely on the notion of learned regularities. Guarini (2010) employed connectionist modeling techniques to explore exactly these issues. Interestingly, case classification could be achieved in relatively simple feedforward and recurrent neural architectures. However, Guarini points out that re-classification of a decided case, upon being given an objection or further information, almost certainly requires general rules.

31.4.2.2 Reinforcement Learning

A recent addition to the computational toolbox of moral decision making are reinforcement learning (RL) approaches (Abel et al., 2016; Crockett, 2013; Cushman, 2013). In short, these approaches conceptualize decision making as a sequence of actions that are transitions from one state of the environment to the next. The algorithms need feedback from the environment (“rewards”) on the state transitions, generating a “reward function.” The system can then find an optimal sequence of actions (“policy”) that maximizes rewards over some time horizon. Two attractive features characterize these approaches. The first is that systems choose among possible actions using a unified valuation (reward) function, which some suggest is compatible with cognitive and neural evidence about human decision making generally, not just moral decision

making (Haas, 2020).[1] A second advantage is that RL models are by nature capable of learning – both bottom-up learning from observation or exploration (Hadfield-Menell et al., 2016) and dynamic updating of initial top-down settings (e.g., a starting set of rules; Malle et al., 2020). Additional features worth mentioning are that RL models are highly suitable for context-specific norm activation (because all actions are individuated relative to a situation or "state"), that reward functions may be able to represent graded deontic forces of norms (Rosen et al., 2022) and they are able to internalize norm-guided actions and execute them reflexively, in line with the popular two-systems view of moral cognition (Cushman, 2013).

RL approaches to moral decision making also have disadvantages. First, they are conceptually lean, lacking important concepts such as intentionality, reasons, or justification, so the agent does not in any way understand *why* it acts as it does and cannot explain its decisions to others (Arnold et al., 2017). Such concepts and processes may be grafted onto the RL algorithms (e.g., actions with certain beliefs and desires are rewarded differently from actions with other beliefs and desires). Indeed, Arnold et al. suggest one such hybrid model, in which the representational format is a modal logic but learning occurs within an RL framework.

A second limitation of RL models is their complete reliance on external feedback. This feedback, and therefore the system's reward function, may be the reflection of a teacher's personal preferences, not the reflection of a community's norm system, and the RL agent would not know the difference. Further, because an RL agent's actions "are strictly determined by the reward signal or signals in the environment" (Haas, 2020, p. 238), the system is unable to maintain previously learned norms in light of novel input from "bad actors." Without significant filtering of external feedback (e.g., by assessing soure reliability or community agreement), a pure RL agent would quickly adopt the worst behaviors of those it learns from.

31.4.2.3 Decision Making in Moral Dilemmas

Philosophers have used moral dilemmas to pit ethical theories against each other, such as in the well-known trolley dilemma, which contrasts utilitarian with deontological reasoning (Foot, 1967): A train has lost control and is destined to kill five people. Is it permissible to switch it onto another track where the five people can be saved but one person is killed? And if one had to push a heavy person off a footbridge to stop the train, would that be permissible? Utilitarians would say yes; deontologists would say no.

Results of numerous studies (Christensen & Gomila, 2012) suggest that people are neither utilitarians nor deontologists, but in the course of this research it became clear that people's moral decisions are deeply influenced by the distinction between intended and unintended (merely foreseen) consequences. Most people find it morally acceptable to cause a person to die if it saves five people and the death is not intended, merely an unavoidable side effect of the decision. By contrast, intentionally using the person as a means to stop the train and save the people is not acceptable. People's moral preferences are in line here with the Principle of Double Effect (Aquinas, 2003). Double-effect reasoning depends critically on a fairly sophisticated capacity for causal and counterfactual reasoning, but formal representation and

[1] This does not imply an automatic commitment to utilitarianism as a classic ethical theory, as the optimal policy can minimize rule violations, maximize utility calculations, or both.

computation of such reasoning has recently seen significant progress (Govindarajulu & Bringsjord, 2017; Pereira & Saptawijaya, 2017).

Psychological and neural scientists have adopted trolley dilemmas to draw a contrast between two psychological processes believed to underlie moral decision making: *reason* vs. *emotion*. Greene (2007) proposed a competition model according to which people have immediate aversive emotional reactions to certain violations (e.g., pushing and killing a person) but also engage in conscious reasoning (e.g., deliberating about the number of people saved), which can temper their emotional reaction. Initial brain imaging evidence and reaction time data seemed to support this dual-process theory (Greene et al., 2001, 2008), but it has faced numerous challenges more recently (e.g., Gürçay & Baron, 2017; Royzman et al., 2011; Sauer, 2012).

On the computational side, Bretz and Sun (2018) used the computational cognitive architecture Clarion to model moral decision making in variants of the trolley dilemma. Integrating implicit and explicit cognitive processes with motivational processes, rather than a simple emotion–reason duality, they offered a compelling account of empirical studies by Greene et al. (2009). However, those studies measured moral judgments (e.g., "Is it morally acceptable for [agent] to…"), not moral decisions. This is common in the moral dilemma literature, though direct comparisons suggest that judgment and decision measures sometimes do not lead to the same results (Francis et al., 2016; Gold et al., 2015; Schaich Borg et al., 2006).

The way to overcome the confound is to model what is common in decisions and judgments. Mikhail (2008) suggests that a fundamental conceptual structure of human action underlies moral and legal judgment and decision making. This structure relates acts, means, ends, and side effects to each other in ways that, according to Mikhail, form the top-level computations of moral reasoning. There is recent evidence that humans do represent moral behavior in such structures (Levine et al., 2018), but the structures operate over concrete actions governed by context-specific norms, leaving powerful processes still unaccounted for.

31.5 Moral Judgment

When making moral judgments, people appraise events, behaviors, or persons in light of moral standards, with the canonical case being an observer's judgment of another person's behavior. Section 31.2.1 distinguished between four kinds of moral judgment: evaluations, norm judgments, wrongness judgments, and blame judgments. How can they be captured computationally?

31.5.1 Evaluations

Evaluations are the appraisal of events or behaviors as good or bad and form a building block of many cognitive architectures (e.g., ACT-R, Clarion). They also lie at the core of RL models of decision making, as the continuous "value" that actions acquire by virtue of the rewards they elicit. However, such action values are typically grounded in the agent's subjective perspective, tied to personal preferences that agents develop in response to environmental feedback for their actions. This makes RL a candidate model for decision making, but *moral evaluation* demands a community perspective – assessing what counts as morally good or bad in this community, relative to its norms and values, not merely relative to the observer's (or individual other

people's) personal preferences. Although RL algorithms can acquire value representations for actions that others perform (Cushman, 2013), it is not as obvious how they can distinguish between moral (community-based) and nonmoral (personal goal-based) value functions. Recently, RL models have emerged that try to integrate the community perspective into an agent's value function (Abel et al., 2016), but the model does not yet distinguish between collective preferences (e.g., most people want coffee) and actual norms (e.g., one must stand in line at the coffee counter).

31.5.2 Norm Judgments

Norm judgments assess an action as permissible, obligatory, or forbidden, which requires retrieving the deontic modality and, likely, the deontic force of the norms that govern a given action. NorMAS systems (such as EMIL-I-A, discussed in Section 31.4.1.3) are able to model these judgments and often take into account the context specificity of norms, but less often their community specificity or community prevalence, and rarely the graded nature of norms. A challenge that all extant models face is that, in order to assess whether a particular action falls under a particular norm, the action must be identified under the description presupposed by the norm (e.g., "shake hands"). In many everyday settings, this identification requires segmenting, representing, and interpreting perceived behavior in terms of agency, causality, and mind – serious obstacles in state-of-the-art machine learning (Marcus & Davis, 2019; Pearl & Mackenzie, 2018). Most computational approaches therefore feed preprocessed information to their algorithms, with all of the interpretational work already done.

One exception comes from Kleiman-Weiner et al. (2015), who focus on the moral perceiver's inferences of another agent's beliefs and intentions en route to permissibility judgments. In the context of trolley dilemmas, they model this process of third-person social-moral cognition, thus speaking to moral judgments in dilemmas, not decisions (where trolley dilemmas are often located). Their account is based on an influence diagram (acyclic graph) representation of the causal dependencies among an agent's decision options, the states each decision option causes, and their resulting utilities. Under the assumption that the agent maximizes utility given beliefs (and that utilities are based on desires and norms), the moral perceiver can infer which states the agent intended and which ones were side effects. Essentially conducting counterfactual double-effect reasoning, the model treats nodes as mere side effects just in case removing them from the causal structure (and the optimal policy) would not change the action taken by the agent.

31.5.3 Judgments of Wrongness

Wrongness judgments build on norm judgments but not only take intentionality into account but the agent's specific reasons and justifications for the action (Cushman, 2008; Malle, 2021). Cushman (2013) proposed that wrongness judgments can be captured by model-free RL models, but the role of justification is not part of such a model, and Ayars (2016) maintained that a model-free RL agent cannot distinguish between morally wrong actions and simply dispreferred actions. Conitzer et al. (2017) propose that an AI system can learn to make moral wrongness judgments via a machine-learning approach: collect a large number of action-in-context stimuli,

along with all their morally relevant features and labels that declare them to be morally wrong or not; then train a deep neural network to infer wrongness from features and generalize to new stimuli. If the initial stimulus collection is representative, such a network will be a practically useful prediction machine, but it is unlikely to be a model of the human cognitive process of wrongness judgments.

31.5.4 Judgments of Blame

Blame judgments go several steps beyond wrongness judgments, as they apply to both intentional and unintentional behavior (or outcomes) and process information about norms, causality, intentionality, justifications, and counterfactuals. Cognitive information processing models of blame have existed for many decades (see Guglielmo, 2015). The first computational model was developed by Shultz (1987). It required describing a violation as an input vector of binary information about harm, foreseeability, intention, and so on. The model then used thirty-nine production rules to infer other judgments as output, primarily responsibility and blame. In effect, the system executed formalized representations of inferences such as "If harm was caused by A, was foreseen by A, … A is responsible." This model redescribed, in more precise language, the best psychological theory of blame at the time, making it potentially amenable to automated reasoning in artificial agents. From the perspective of cognitive theory, however, such redescriptions do not substantially exceed insights gained from existing linear regression models of experimental data.

Mao and Gratch (2012) offered a more elaborate model. The system performs dialogue analysis of short narratives that describe agents' main actions, consequences, and speech acts, and it builds hierarchical plan representations using up to twenty-seven predicates, fifteen functions, and twenty-six inference rules, which capture concepts such as *intends, believes, coercion*, and *causal responsibility.* This expressiveness to represent the complex concepts and processes involved in blame judgments is a strength of the model, yet it still captures only a portion of this complexity, omitting elements such as justification and counterfactuals (e.g., obligations to prevent violations), and it culminates in only a qualitative assignment of who is to blame, rather than a graded judgment of how much blame the person deserves.

Sileno et al. (2017) used simplicity theory (a variant of information theory) to represent concepts such as *causal contribution*, *foreseeability*, and *intention* in terms of the conditional expectedness of situations, given actions or other situations, along with a concept of *emotion* (akin to perceived value). An agent's moral responsibility (in effect, blame) for an action is defined as a function of the resulting situation's (dis)value, how much the action caused the situation, how much the agent foresaw it, and the complexity of description (which is not further clarified). A strength of the proposal is to consider uncertainty, continuous variables, and distinct points of view (e.g., what the agent knew vs. what an observer knows), thus hinting at a theory of mind capacity. However, it is unclear how some of the terms could be measured (e.g., the objective complexity of events) and, as with other models, some central concepts are omitted, such as moral norms (beyond personal desirability), the agent's reasons for acting, or obligations to prevent violations.

More technically refined but conceptually narrower is the formal treatment of blameworthiness by Halpern and Kleiman-Weiner (2018). Their concept of blame is, roughly, counterfactual causal responsibility for negative outcomes. This formalism handles the notion of

preventability (blame increases if the person could have taken an alternative action that would have prevented the negative outcome) but would need to be augmented by a concept of norm to handle degrees of obligation to prevent. They also define intention within a utility maximizing framework, but they do not integrate intention and blame or handle reasons and their justification.

These preliminary models of moral judgments provide promising starting points, and it now becomes imperative to account for the full range of information that humans process and the full range of moral judgments they form. Future models may aggregate separate components into a processing hierarchy (e.g., RL for evaluation, extended BOID for wrongness, all the way to a hybrid for blame) and connect them to the complex nonmoral capacities of theory of mind and causal-counterfactual reasoning.

31.6 Other Moral Phenomena

Compared to moral decision making and moral judgment, computational work on the remaining moral phenomena outlined in Figure 31.1 has been sparse. Nonetheless, some promising starting points may accelerate development in the near future.

31.6.1 Moral Emotions

Affect and emotion can relate to morality in at least two ways. First, they can causally interact with moral phenomena. Here, computational work is sparse. Arkin & Ulam (2009) incorporated the emotion of guilt as a corrective process in a lethal autonomous weapon's responses to the unintended harm it causes, improving its future decisions. Cervantes, Rodríguez, López, Ramos, & Robles (2016) proposed an ambitious model in which ethical decision making is strongly influenced by emotions, moods, and evaluative experiences. Though inspired by brain science, the model has not yet been shown to model any psychological data, and as a framework for artificial agents' decision making, the computation of over two dozen parameters (assessed for each of many potential actions) appears daunting.

Second, some emotions can themselves be moral. Such "moral emotions" include guilt and remorse as the clearest cases, but also disgust, anger, or sympathy. Though several computational models of emotions have been offered (for reviews see Kowalczuk & Czubenko, 2016; Rosales, Rodríguez, & Ramos, 2019), models of specifically moral emotions are rare. Ferreira et al. (2013) coded moral emotions such as shame or reproach as reactions to norm violations. More extensively, Battaglino, Damiano, & Lombardo (2014) equipped a BDI architecture to assess not only whether the agent's goals are achieved but also whether its values (e.g., honesty, loyalty, justice) are maintained. Emotions are constituted by combinations of the agent's "appraisals" of behaviors or events as desirable, causal, or blameworthy (following the cognitive theory of emotions; Ortony et al., 1988). For example, failure to achieve a goal (appraised as undesirable) leads to "distress," whereas threats to values (appraised as blameworthy) lead to "shame" if appraised as self-caused or "reproach" if appraised as caused by another agent. The intensity of the resulting emotion is proportional to the importance of the goals and/or values at stake. One advantage of such a system is its transparency, as it allows the agent to explain exactly why it "feels" a certain emotion ("because I didn't achieve this very

important goal and also went against one of my values…"). A disadvantage lies in its summative functions (see formulas in Battaglino et al., 2013), which permit disconcerting trade-offs, such as that fulfilling two goals can make up for one violated value. A general question, applicable to this and related models is what effective work the "emotion states" actually do, if they are merely linear functions of a number of nonemotional appraisals. One response is that, at the level of action guidance, they may be dispensable, but if the interwoven appraisals result in *expressed* emotions, such as remorse or gratitude, then humans interacting with such computational agents may better understand the agent and be more accepting of it. Whether a robot that expresses human-like emotions constitutes deceptive design is an important concern (Danaher, 2020).

31.6.2 Moral Sanctions

Moral sanctions include social blame, acts of shaming, and interpersonal or institutional punishment. There is some amount of empirical research on social blame (Balafoutas et al., 2014) and shaming (e.g., Coricelli et al., 2014), whereas punishment responses have been studied somewhat more extensively and systematically, primarily using economic games (Zinchenko, 2019). Sometimes they are computationally modeled as linear functions of the stimulus and role conditions – for example, degree of punishment = *f*(how much money a player takes away from another player); see Stallen et al. (2018). It is unclear how generalizable economic games with strangers are to the broad range of moral situations in ordinary life (Guala, 2012), so an expansion of research in this domain is needed. Studies do suggest that neural processes underlying sanctioning behavior are distinct from those underlying moral judgments (Buckholtz et al., 2015; Zinchenko, 2019). These neural models may be amenable to computational treatment, perhaps relatable to computational models of moral evaluation (Cervantes et al., 2016).

31.6.3 Moral Communication

With increasing interactions between humans and artificial agents, the need to communicate about moral matters is increasing. Computational work in this domain, however, is sparse. Some authors have begun to model justifications as explanations of decisions that refer to norms (Kasenberg et al., 2019), and models based on argumentation logic are able to explain their resolutions to norm conflicts (Shams et al., 2020). Many other forms of moral communication have been left untouched, such as expressed moral criticism (which, computationally, would require both full-fledged moral judgment capacities and sophisticated communication and theory of mind skills), or apologies. Given the continued error-proneness of artificial agents, implementing capacities for effective apology would seem particularly useful. Psychological research has only recently begun to identify the decisive components of such effectiveness (Cerulo & Ruane, 2014; Slocum et al., 2011). Successful apologies must certainly build on theory of mind skills (simulating what would soften the other's blame judgments) and discourse skills (e.g., foregrounding the victim). There is also evidence that apologies are most successful when the apologizing offender incurs a cost, such as through atoning actions (Ohtsubo et al., 2018; Watanabe & Laurent, 2020). It is an intriguing question how an artificial agent might convince humans that it incurred such a cost.

31.7 Conclusion

The wide diversity of moral phenomena poses significant challenges for empirical research and computational modeling. No single brain area or psychological mechanism exists that represents norms, selects moral actions, makes moral judgments, instantiates moral emotions, and conducts moral communication. In addition, moral processes build on almost the entire suite of human mental capacities – from attention to memory, from evaluation to causal perception, from counterfactual analysis to theory of mind. As a result, no one computational model, tool, or approach will be able to formalize and elucidate these diverse phenomena. This challenging situation, however, offers the opportunity to build the best computational tools for specific functions and phenomena and enable a fruitful confluence of many different schools of thought – logic, connectionism, probabilistic inference, reinforcement learning, and many more. The question should not be which model is correct but what an integrative model will look like. Such a model must pay close attention to the rapidly growing empirical science of morality and capture the distinctions and patterns that characterize human moral phenomena. Such a model will have significant innovative impact on moral science – by pointing to undiscovered relations and developing novel predictions, demanding new experiments and revisions to theory. And with such an integrative model, the goal of building artificial moral agents, for those who pursue it, will be more feasible, safer, and better attuned to human social reality.

Acknowledgments

This work was supported by ONR Awards N0001419WX00020 and N00014–14-1-0144 to PB and BFM, respectively, and NSF grant 1717701, awarded to BFM.

References

Aarts, H., & Dijksterhuis, A. (2003). The silence of the library: environment, situational norm, and social behavior. Journal of Personality and Social Psychology, 84(1), 18–28. https://doi.org/10.1037/0022-3514.84.1.18

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. AAAI Workshop: AI, Ethics, and Society, Volume WS-16-02 of 13th AAAI Workshops.

Alexander, J. C. (1987). The Micro-Macro Link. Oakland, CA: University of California Press.

Alicke, M. D. (2000). Culpable control and the psychology of blame. Psychological Bulletin, 126(4), 556–574. https://doi.org/10.1037//0033-2909.126.4.556

Anderson, M., & Anderson, S. L. (2006). MedEthEx: a prototype medical ethics advisor. Paper Presented at the 18th Conference on Innovative Applications of Artificial Intelligence.

Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. IEEE Intelligent Systems, 21(4), 56–63. https://doi.org/10.1109/MIS.2006.64

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). https://doi.org/10.1371/journal.pone.0064941

Andrighetto, G., Castelfranchi, C., Mayor, E., McBreen, J., Lopez-Sanchez, M., & Parsons, S. (2013). (Social) norm dynamics. In G. Andrighetto, G. Governatori, P. Noriega, & L. W. N. van der Torre (Eds.), Normative Multi-Agent Systems (Vol. 4, pp. 135–170). Wadern: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/DFU.Vol4.12111.135

Andrighetto, G., Villatoro, D., & Conte, R. (2010a). Norm internalization in artificial societies. AI Communications, 23(4), 325–339.

Andrighetto, G., Villatoro, D., & Conte, R. (2010b). Norm internalization in artificial societies. AI Communications, 23(4), 325–339.

Aquinas, T. (2003). On Law, Morality and Politics (W. P. Baumgarth, Ed.; R. J. Regan, Trans.; 2nd ed.). Indianapolis, IN: Hackett Publishing Company, Inc.

Arkin, R. C., & Ulam, P. (2009). An ethical adaptor: behavioral modification derived from moral emotions. Proceedings of the 2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation - (CIRA), 381–387. https://doi.org/10.1109/CIRA.2009.5423177

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment – What will keep systems accountable? In The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports, WS-17-02: AI, Ethics, and Society (pp. 81–88). Palo Alto, CA: The AAAI Press.

Ayars, A. (2016). Can model-free reinforcement learning explain deontological moral judgments? Cognition, 150, 232–242. https://doi.org/10.1016/j.cognition.2016.02.002

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. Proceedings of the National Academy of Sciences, 111(45), 15924–15927. https://doi.org/10.1073/pnas.1413170111

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), The Wiley Blackwell Handbook of Judgment and Decision Making (pp. 478–515). Oxford: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333.ch17

Battaglino, C., Damiano, R., & Lesmo, L. (2013). Emotional range in value-sensitive deliberation. Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems, 769–776.

Battaglino, C., Damiano, R., & Lombardo, V. (2014). Moral values in narrative characters: an experiment in the generation of moral emotions. In A. Mitchell, C. Fernández-Vara, & D. Thue (Eds.), Interactive Storytelling (pp. 212–215). Cham: Springer International Publishing.

Bauer, W. A. (2020). Virtuous vs. Utilitarian artificial moral agents. AI & SOCIETY, 35(1), 263–271. https://doi.org/10.1007/s00146-018-0871-3

Benzmüller, C. (2019). Universal (meta-)logical reasoning: recent successes. Science of Computer Programming, 172, 48–62. https://doi.org/10.1016/j.scico.2018.10.008

Bicchieri, C. (2006). The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge: Cambridge University Press.

Bordini, R. H., Hübner, J. F., & Wooldridge, M. (2007). Programming Multi-Agent Systems In Agentspeak Using Jason. Oxford: John Wiley & Sons, Inc.

Bratman, M. E. (1987). Intention, Plans, and Practical Reason. Cambridge, MA: Harvard University Press.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). Explaining Norms. Oxford: Oxford University Press.

Bretz, S., & Sun, R. (2018). Two models of moral judgment. Cognitive Science, 42, 4–37. https://doi.org/10.1111/cogs.12517

Bringsjord, S., & Taylor, J. (2012). The divine-command approach to robot ethics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), Robot Ethics: The Ethical and Social Implications of Robotics (pp. 85–108). Cambridge, MA: MIT Press.

Broeders, R., van den Bos, K., Müller, P. A., & Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. Journal of Experimental Social Psychology, 47(5), 923–934. https://doi.org/10.1016/j.jesp.2011.03.018

Brundage, M. (2014). Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial Intelligence, 26(3), 355–372.

Buckholtz, J. W., Martin, J. W., Treadway, M. T., et al. (2015). From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron, 87(6), 1369–1380. https://doi.org/10.1016/j.neuron.2015.08.023

Carmo, J., & Jones, A. J. I. (2002). Deontic logic and contrary-to-duties. In D. M. Gabbay & F. Guenthner (Eds.), Handbook of Philosophical Logic (Vol. 8, pp. 265–343). Cham: Springer. https://doi.org/10.1007/978-94-010-0387-2_4

Castelfranchi, C., Dignum, F., Jonker, C. M., & Treur, J. (2000). Deliberative normative agents: principles and architecture. In N. R. Jennings & Y. Lespérance (Eds.), Intelligent Agents VI. Agent Theories, Architectures, and Languages (pp. 364–378). Cham: Springer. https://doi.org/10.1007/10719619_27

Cerulo, K. A., & Ruane, J. M. (2014). Apologies of the rich and famous: cultural, cognitive, and social explanations of why we care and why we forgive. Social Psychology Quarterly, 77(2), 123–149.

Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. Cognitive Computation, 8(2), 278–296. https://doi.org/10.1007/s12559-015-9362-8

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. https://doi.org/10.1016/j.neubiorev.2012.02.008

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. In M. P. Zanna (Ed.), Advances in Experimental Social Psychology (Vol. 24, pp. 201–234). New York, NY: Academic Press.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. Proceedings of the 31st AAAI Conference on Artificial Intelligence, 4831–4835.

Conte, R., Andrighetto, G., & Campenni, M. (2013). Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies. New York, NY: Oxford University Press.

Coricelli, G., Rusconi, E., & Villeval, M. C. (2014). Tax evasion and emotions: an empirical test of re-integrative shaming theory. Journal of Economic Psychology, 40, 49–61. https://doi.org/10.1016/j.joep.2012.12.002

Crockett, M. J. (2013). Models of morality. Trends in Cognitive Sciences, 17(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. Current Anthropology, 60(1), 47–69. https://doi.org/10.1086/701478

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. Cognition, 108(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. Personality and Social Psychology Review, 17(3), 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris (Ed.), The Moral Psychology Handbook (pp. 47–71). Oxford: Oxford University Press.

Danaher, J. (2020). Robot betrayal: a guide to the ethics of robotic deception. Ethics and Information Technology, 22(2), 117–128. https://doi.org/10.1007/s10676-019-09520-3

Dancy, J. (2009). Moral particularism. In E. N. Zalta (Ed.), Stanford Encyclopedia of Philosophy. Center for the Study of Language and Information, Stanford University. https://plato.stanford.edu/entries/moral-particularism/ [last accessed July 27, 2022].

Dastani, M. (2008). 2APL: a practical agent programming language. Autonomous Agents and Multi-Agent Systems, 16(3), 214–248. https://doi.org/10.1007/s10458-008-9036-y

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems, 77, 1–14. https://doi.org/10.1016/j.robot.2015.11.012

D'Inverno, M., Luck, M., Georgeff, M., Kinny, D., & Wooldridge, M. (2004). The dMARS architecture: a specification of the distributed multi-agent reasoning system. Autonomous Agents and Multi-Agent Systems, 9(1), 5–53. https://doi.org/10.1023/B:AGNT.0000019688.11109.19

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 77(2), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Eisenberg, N. (2000). Emotion, regulation, and moral development. Annual Review of Psychology, 51, 665–697.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. Trends in Cognitive Sciences, 8(4), 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Ferreira, N., Mascarenhas, S., Paiva, A., et al. (2013). An agent model for the appraisal of normative events based in in-group and out-group relations. AAAI Conference on Artificial Intelligence.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. Oxford Review, 5, 5–15.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: a model of similarity-based retrieval. Cognitive Science, 19(2), 141–205. https://doi.org/10.1207/s15516709cog1902_1

Francis, K. B., Howard, C., Howard, I. S., et al. (2016). Virtual morality: transitioning from moral judgment to moral action? PLoS One, 11(10), e0164374. https://doi.org/10.1371/journal.pone.0164374

Gibbs, J. P. (1965). Norms: the problem of definition and classification. American Journal of Sociology, 70(5), 586–594. https://doi.org/10.1086/223933

Goble, L. (2003). Preference semantics for deontic logic: Part I — Simple models. Logique et Analyse, 46(183/184), 383–418.

Gold, N., Pulford, B. D., & Colman, A. M. (2015). Do as I Say, Don't Do as I Do: differences in moral judgments do not translate into differences in decisions in real-life trolley problems. Journal of Economic Psychology, 47, 50–61. https://doi.org/10.1016/j.joep.2015.01.001

Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. Proceedings of the International Joint Conference on AI (IJCAI 2017), 4722–4730.

Govindarajulu, N. S., Bringsjord, S., & Peveler, M. (2019). On quantified modal theorem proving for modeling ethics. Electronic Proceedings in Theoretical Computer Science, 311, 43–49. https://doi.org/10.4204/EPTCS.311.7

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. Trends in Cognitive Sciences, 11(8), 322–323. https://doi.org/10.1016/j.tics.2007.06.004

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. Cognition, 111(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Guarini, M. (2007). Computation, coherence, and ethical reasoning. Minds and Machines, 17(1), 27–46. https://doi.org/10.1007/s11023-007-9056-4

Guarini, M. (2010). Particularism, analogy, and moral cognition. Minds and Machines, 20(3), 385–422. https://doi.org/10.1007/s11023-010-9200-4

Guglielmo, S. (2015). Moral judgment as information processing: an integrative review. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.01637

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. Thinking & Reasoning, 23(1), 49–80. https://doi.org/10.1080/13546783.2016.1216011

Haas, J. (2020). Moral gridworlds: a theoretical proposal for modeling artificial moral cognition. Minds and Machines, 30(2), 219–246. https://doi.org/10.1007/s11023-020-09524-9

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 29 (pp. 3909–3917). Red Hook, NY: Curran Associates, Inc.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological Review, 108(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.

Hechter, M., & Opp, K.-D. (Eds.). (2001). Social Norms. New York, NY: Russell Sage Foundation.

Heider, F. (1958). The Psychology of Interpersonal Relations. Oxford: Wiley.

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: a framework for commonsense moral reasoning. Psychological Bulletin, 142(11), 1179–1203. https://doi.org/10.1037/bul0000075

Howard, D., & Muntean, I. (2017). Artificial moral cognition: moral functionalism and autonomous moral agency. In T. Powers (Ed.), Philosophy and Computing (pp. 121–159). Cham: Springer. https://doi.org/10.1007/978-3-319-61043-6_7

Kasenberg, D., Roque, A., Thielstrom, R., Chita-Tegmark, M., & Scheutz, M. (2019). Generating justifications for norm-related agent decisions. 12th International Conference on Natural Language Generation (INLG), Tokyo, Japan.

Kasenberg, D., & Scheutz, M. (2018). Norm conflict resolution in stochastic domains. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 85–92.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In Proceedings of the 37th Annual Conference of the Cognitive Science Society (pp. 1123–1128). Cognitive Science Society.

Kohlberg, L. (1984). The Psychology of Moral Development: The Nature and Validity of Moral Stages. New York, NY: Harper & Row.

Kowalczuk, Z., & Czubenko, M. (2016). Computational approaches to modeling artificial emotion – An overview of the proposed solutions. Frontiers in Robotics and AI, 3. https://doi.org/10.3389/frobt.2016.00021

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. https://doi.org/10.1080/02699931.2015.1058242

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. Ethics and Information Technology, 19(2), 107–115.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J. B., & Cushman, F. A. (2020). The logic of universalization guides moral judgment [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/p7e6h

Levine, S., Leslie, A. M., & Mikhail, J. (2018). The mental representation of human action. Cognitive Science, 42(4), 1229–1264. https://doi.org/10.1111/cogs.12608

Lindenberg, S. (2013). How cues in the environment affect normative behaviour. In L. Steg, A. E. van den Berg, & J. I. M. de Groot (Eds.), Environmental Psychology: An Introduction (pp. 119–128). Oxford: BPS/ Blackwell.

Malle, B. F. (2020). Graded representations of norm strength. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (pp. 3342–3348). Cognitive Science Society.

Malle, B. F. (2021). Moral judgments. Annual Review of Psychology, 72. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., Rosen, E., Chi, V. B., Berg, M., & Haas, P. (2020). A general methodology for teaching norms to social robots. Proceedings of the 29th International Conference on Robot & Human Interactive Communication (RO-MAN 2020).

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, & G. S. Virk (Eds.), A World with Robots: International Conference on Robot Ethics: ICRE 2015 (pp. 3–17). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-46667-5_1

Mao, W., & Gratch, J. (2012). Modeling social causality and responsibility judgment in multi-agent interactions. Journal of Artificial Intelligence Research, 44, 223–273.

Marcus, G., & Davis, E. (2019). Rebooting AI: Building Artificial Intelligence We Can Trust. New York, NY: Pantheon.

McLaren, B. M. (2006). Computational models of ethical reasoning: challenges, initial steps, and future directions. IEEE Intelligent Systems, 21, 29–37.

Meyer, J. J. Ch., Broersen, J. M., & Herzig, A. (2015). BDI Logics. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, & B. Kooi (Eds.), Handbook of Logics of Knowledge and Belief (pp. 453–498). Rickmansworth: College Publications. https://dspace.library.uu.nl/handle/1874/315954

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), Moral Psychology, Vol. 3: The Neuroscience of Morality (pp. 81–92). Cambridge, MA: MIT Press.

Ohtsubo, Y., Matsunaga, M., Tanaka, H., et al. (2018). Costly apologies communicate conciliatory intention: an fMRI study on forgiveness in response to costly apologies. Evolution and Human Behavior, 39(2), 249–256. https://doi.org/10.1016/j.evolhumbehav.2018.01.004

Ortony, A., Clore, G. L., & Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect (1st ed.). New York, NY: Basic Books.

Pereira, L. M., & Saptawijaya, A. (2017). Counterfactuals, logic programming and agent morality. In R. Urbaniak & G. Payette (Eds.), Applications of Formal Philosophy: The Road Less Travelled (pp. 25–53). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58507-9_3

Powers, T. M. (2006). Prospects for a Kantian machine. IEEE Intelligent Systems, 21(4), 46–51. https://doi.org/10.1109/MIS.2006.77

Prakken, H., & Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations. In D. Nute (Ed.), Defeasible Deontic Logic (pp. 223–262). Cham: Springer. https://doi.org/10.1007/978-94-015-8851-5_10

Prinz, J. (2006). The emotional basis of moral judgments. Philosophical Explorations, 9(1), 29–43. https://doi.org/10.1080/13869790500492466

Quinn, P. L. (1978). Divine Commands and Moral Requirements. Oxford: Clarendon Press.

Rao, A. S. (1996). AgentSpeak(L): BDI agents speak out in a logical computable language. In W. Van de Velde & J. W. Perram (Eds.), Agents Breaking Away (pp. 42–55). Cham: Springer.

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, 473–484. http://dl.acm.org/citation.cfm?id=3087158.3087205

Realpe-Gómez, J., Andrighetto, G., Nardin, L. G., & Montoya, J. A. (2018). Balancing selfishness and norm conformity can explain human behavior in large-scale prisoner's dilemma games and can poise human groups near criticality. Physical Review E, 97(4), 042321. https://doi.org/10.1103/PhysRevE.97.042321

Rosales, J.-H., Rodríguez, L.-F., & Ramos, F. (2019). A general theoretical framework for the design of artificial emotion systems in Autonomous Agents. Cognitive Systems Research, 58, 324–341. https://doi.org/10.1016/j.cogsys.2019.08.003

Ross, W. D. (1930). The Right and the Good. Oxford: Oxford University Press.

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. https://doi.org/10.1016/j.cognition.2011.06.006

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. New York, NY: Viking.

Sachdeva, S., Singh, P., & Medin, D. (2011). Culture and the quest for universal principles in moral reasoning. International Journal of Psychology, 46(3), 161–176. https://doi.org/10.1080/00207594.2011.568486

Santos, J. S., Zahn, J. O., Silvestre, E. A., Silva, V. T., & Vasconcelos, W. W. (2017). Detection and resolution of normative conflicts in multi-agent systems: a literature survey. Autonomous Agents and Multi-Agent Systems, 31(6), 1236–1282. https://doi.org/10.1007/s10458-017-9362-z

Sauer, H. (2012). Morally irrelevant factors: what's left of the dual process-model of moral cognition? Philosophical Psychology, 25(6), 783–811. https://doi.org/10.1080/09515089.2011.631997

Scanlon, T. (1998). What We Owe to Each Other (Issue 1, pp. 169–175). Cambridge, MA: Harvard University Press.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. Journal of Cognitive Neuroscience, 18(5), 803–817. https://doi.org/10.1162/jocn.2006.18.5.803

Shams, Z., Vos, M. D., Oren, N., & Padget, J. (2020). Argumentation-based reasoning about plans, maintenance goals, and norms. ACM Transactions on Autonomous and Adaptive Systems, 14(3), 9:1–9:39. https://doi.org/10.1145/3364220

Shaver, K. G. (1985). The Attribution of Blame: Causality, Responsibility, and Blameworthiness. New York, NY: Springer Verlag.

Shoham, Y., & Tennenholtz, M. (1995). On social laws for artificial agent societies: off-line design. Artificial Intelligence, 73(1–2), 231–252. https://doi.org/10.1016/0004-3702(94)00007-N

Shultz, T. R. (1987). A computational model of causation, responsibility, blame, and punishment. Meeting of the Society for Research in Child Development, Baltimore, MD.

Sileno, G., Saillenfest, A., & Dessalles, J.-L. (2017). A computational model of moral and legal responsibility via simplicity theory. In A. Wyner & G. Casini (Eds.), Legal Knowledge and Information Systems (pp. 171–176). Clifton, VA: IOS Press. http://ebooks.iospress.nl/publication/48059

Slocum, D., Allan, A., & Allan, M. M. (2011). An emerging theory of apology. Australian Journal of Psychology, 63(2), 83–92. https://doi.org/10.1111/j.1742-9536.2011.00013.x

Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), The Innate Mind (Vol. 2: Culture and Cognition) (pp. 280–301). Oxford: Oxford University Press.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K. W., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. The Journal of Neuroscience, 38(12), 2944–2954. https://doi.org/10.1523/JNEUROSCI.1242-17.2018

Tangney, J. P., & Dearing, R. L. (2002). Shame and Guilt. New York, NY: Guilford Press.

Thagard, P. (1998). Ethical coherence. Philosophical Psychology, 11(4), 405–422. https://doi.org/10.1080/09515089808573270

Turiel, E. (2002). The Culture of Morality: Social Development, Context, and Conflict. Cambridge: Cambridge University Press.

Ullmann-Margalit, E. (1977). The Emergence of Norms. Oxford: Clarendon Press.

van der Torre, L. W. N., & Tan, Y.-H. (1997). The many faces of defeasibility in defeasible deontic logic. In D. Nute (Ed.), Defeasible Deontic Logic (pp. 79–121). Cham: Springer. https://doi.org/10.1007/978-94-015-8851-5_5

Von Wright, G. H. (1951). Deontic logic. Mind, LX(237), 1–15. https://doi.org/10.1093/mind/LX.237.1

Watanabe, S., & Laurent, S. M. (2020). Feeling bad and doing good: forgivability through the lens of uninvolved third parties. Social Psychology, 51(1), 35–49. https://doi.org/10.1027/1864-9335/a000390

Weiner, B. (2001). Responsibility for social transgressions: an attributional analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), Intentions and Intentionality: Foundations of Social Cognition (pp. 331–344). Cambridge, MA: MIT Press.

Zinchenko, O. (2019). Brain responses to social punishment: a meta-analysis. Scientific Reports, 9. https://doi.org/10.1038/s41598-019-49239-1