CellPress
OPEN ACCESS

Trends in
Cognitive Sciences

Feature Review

Computational ethics

Edmond Awad,[1,2,3,34,*] Sydney Levine,[4,5,34,*] Michael Anderson,[6] Susan Leigh Anderson,[7] Vincent Conitzer,[8,9,10,11] M.J. Crockett,[12] Jim A.C. Everett,[13] Theodoros Evgeniou,[14,32] Alison Gopnik,[15] Julian C. Jamison,[1,16] Tae Wan Kim,[17] S. Matthew Liao,[18] Michelle N. Meyer,[19,20,21] John Mikhail,[22] Kweku Opoku-Agyemang,[23,24,33] Jana Schaich Borg,[25,26] Juliana Schroeder,[27] Walter Sinnott-Armstrong,[10,26,28] Marija Slavkovik,[29] and Josh B. Tenenbaum[4,30,31]

Technological advances are enabling roles for machines that present novel ethical challenges. The study of 'AI ethics' has emerged to confront these challenges, and connects perspectives from philosophy, computer science, law, and economics. Less represented in these interdisciplinary efforts is the perspective of cognitive science. We propose a framework – computational ethics – that specifies how the ethical challenges of AI can be partially addressed by incorporating the study of human moral decision-making. The driver of this framework is a computational version of reflective equilibrium (RE), an approach that seeks coherence between considered judgments and governing principles. The framework has two goals: (i) to inform the engineering of ethical AI systems, and (ii) to characterize human moral judgment and decision-making in computational terms. Working jointly towards these two goals will create the opportunity to integrate diverse research questions, bring together multiple academic communities, uncover new interdisciplinary research topics, and shed light on centuries-old philosophical questions.

Highlights

The past 15 years have seen an increased interest in developing ethical machines; manifested in various interdisciplinary research communities (under the umbrella term 'AI ethics'). Less represented in these interdisciplinary efforts is the perspective of cognitive science.

We propose a framework – computational ethics – that specifies how the ethical challenges of AI can be addressed better by incorporating the study of how humans make moral decisions.

As the driver of this framework, we propose a computational version of reflective equilibrium.

The goal of this framework is twofold: (i) to inform the engineering of ethical AI systems, and (ii) to characterize human moral judgment and decision-making in computational terms.

Working jointly towards these two goals may prove to be beneficial in making progress on both fronts.

A computational approach to ethics

David Marr set out to describe vision in computational terms by integrating insights and methods from psychology, neuroscience, and engineering [1]. His revolutionary approach to the study of vision offered a model for the field of cognitive science. The key to Marr's innovation was his emphasis on explaining visual perception as an **algorithmic** (see Glossary) process – a process that transforms one type of information (an input) into another type of information (the output). The goal was to understand the input–output transformation with a sufficiently high degree of precision that it could be captured in mathematical terms. The result of this algorithm-focused pursuit was an account of visual perception that characterized the richness of the human mind in a way that could be programmed into a machine.

This approach had two important consequences. The first was that it has become increasingly possible to build machines with a human-like capacity for visual perception. For example, convolutional neural networks (CNNs), the engine underlying most of the recent progress in computer vision, learn internal multi-level representations analogous to the human visual hierarchy [2]. Given these advances, we now have machines that can detect whether a skin cancer is malignant or benign [3], can detect street signs in naturalistic settings [4], and can classify objects into a thousand categories at better than human performance levels [5]. The second was that the mechanisms of human vision were studied and understood in more precise terms than ever before. For example, various aspects of visual perception and decoding (specifically, inference of selected objects) can be understood as a Bayesian inference [6,7]. Moreover, there was a positive feedback loop between the machine-centric and human-centric research lines. The

[1]Department of Economics, University of Exeter, Exeter, UK
[2]Institute for Data Science and AI, University of Exeter, Exeter, UK
[3]Center for Humans and Machines, Max-Planck Institute for Human Development, Berlin, Germany
[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA
[5]Department of Psychology, Harvard University, Cambridge, MA, USA
[6]Department of Computer Science, University of Hartford, West Hartford, CT, USA
[7]Department of Philosophy, University of Connecticut, Storrs, CT, USA
[8]Department of Computer Science, Duke University, Durham, NC, USA

 https://doi.org/10.1016/j.tics.2022.02.009

algorithms developed in studying the cognitive science of human vision were used to program machines that both matched and extended what the human mind is capable of. Conversely, the challenge of trying to program machines with the capacity for vision generated new hypotheses for how vision might work in the mind (and the brain). The key to this success was thinking about vision computationally – that is, in algorithmic terms.

Inspired by Marr's success, we propose that a computationally grounded approach could be similarly valuable for the study of ethics [8]. By analogy to Marr's 'computational vision', we characterize this approach as 'computational ethics'. As we conceive it, computational ethics includes scholarly work that aims to formalize **descriptive ethics** and **normative ethics** in algorithmic terms, as well as work that uses this formalization to help to both (i) engineer ethical AI systems, and (ii) better understand human moral decisions and judgments (the relationship between our proposed framework and other interdisciplinary efforts that tackle the challenges of AI ethics is discussed in Box 1).

We first consider how formalizing our normative views and theories of moral cognition can enable progress in engineering ethical AI systems that behave in ways we find morally acceptable [9,10]. Such considerations will yield valuable lessons for **machine ethics** (Box 2 discusses whether humans should delegate ethics to machines). The following example illustrates the process of developing machine ethics in kidney exchange.

Example 1 [kidney exchange]. *Thousands of patients are in need of kidney transplants, and thousands of individuals are willing to donate kidneys (sometimes on the condition that kidneys are allocated a certain way). However, kidneys can only be allocated to compatible patients, and there are always more people in need of kidneys than willing donors. How should kidneys be allocated?*

Box 1. Relationship to current interdisciplinary efforts

Fifteen years ago the fields of machine ethics (implementing ethical decision-making in machines) [56,132] and roboethics (how humans design, use, and treat robots) [133,134] emerged to bring the perspective of ethics to AI development. Since then 'AI ethics' has emerged as an umbrella term to describe work concerning both AI and ethics. New research directions for AI ethics include algorithmic accountability (the obligation to be able to explain and/or justify algorithmic decisions) [135], algorithmic transparency (openness about the purpose, structure, and actions of algorithms) [136], algorithmic fairness/bias (attempts to design algorithms that make fair/unbiased decisions) [137], and AI for (social) good (ensuring that AI algorithms have a positive impact) [138]. Similarly, new multidisciplinary fields of research have been initiated, including responsible AI (RAI; the development of guidelines, regulations, laws, and certifications regarding how AI should be researched, developed, and used) [139], explainable AI (XAI; the development and study of automatically generated explanations for algorithmic decisions) [140,141], and machine behavior (the study of machines as a new class of actors with their unique behavioral patterns and ecology) [31].

These fields and communities have already begun to communicate via academic conferences including AIES (Artificial Intelligence, Ethics, and Society), supported by the Association for the Advancement of AI (AAAI) and the Association for Computing Machinery (ACM), and FAT/ML (Fairness, Accountability, and Transparency in Machine Learning), as well as workshops such as FATES (Fairness, Accountability, Transparency, Ethics, and Society on the Web), FACTS-IR (Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval), and HWB (Handling Web Bias). There are also governmental and global initiatives such as the AI for Good Global Summit, AI for Good initiative, and Partnership on AI. Other organizations – such as the Organization for Economic Co-operation and Development (OECD) AI Policy Observatory, UNESCO, the World Economic Forum, and the Institute of Electrical and Electronics Engineers (IEEE) – have convened a wide range of stakeholders to lay out ethical principles for the development and implementation of AI.

Often missing from these pursuits is the perspective of cognitive science that studies how humans (as individuals or as groups) think about, learn, and make moral decisions. The aim of the computational ethics framework is to complement and supplement the work being done in these communities by reviewing the ongoing research and providing a new structure that helps to focus work toward both building ethical machines and better understanding human ethics.

[9]Department of Economics, Duke University, Durham, NC, USA
[10]Department of Philosophy, Duke University, Durham, NC, USA
[11]Institute for Ethics in AI, University of Oxford, Oxford, UK
[12]Department of Psychology, Yale University, New Haven, CT, USA
[13]School of Psychology, University of Kent, Canterbury, UK
[14]INSEAD, Fontainebleau, France
[15]Department of Psychology, University of California, Berkeley, CA, USA
[16]Global Priorities Institute, Oxford University, Oxford, UK
[17]Ethics Group, Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA
[18]Center for Bioethics, New York University, New York, NY, USA
[19]Center for Translational Bioethics and Health Care Policy, Geisinger Health System, Danville, PA, USA
[20]Steele Institute for Health Innovation, Geisinger Health System, Danville, PA, USA
[21]Geisinger Commonwealth School of Medicine, Scranton, PA, USA
[22]Georgetown University Law Center, Washington, DC, USA
[23]International Growth Centre, London School of Economics, London, UK
[24]Machine Learning X Doing, Toronto, ON, Canada
[25]Social Science Research Institute, Duke University, Durham, NC, USA
[26]Duke Institute for Brain Sciences, Duke University, Durham, NC, USA
[27]Haas School of Business, University of California, Berkeley, CA, USA
[28]Kenan Institute for Ethics, Duke University, Durham, NC, USA
[29]Department of Information Science and Media Studies, University of Bergen, Bergen, Norway
[30]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[31]Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA
[32]Tremau, Paris, France
[33]Development Economics X, Toronto, ON, Canada
[34]Equal contributions

*Correspondence:
e.awad@exeter.ac.uk (E. Awad) and
smlevine@mit.edu (S. Levine).

Box 2. On delegating ethics to machines

Should humans delegate ethics to machines? In opposition to this idea, van Wynsberghe and Robbins propose 'a moratorium on the commercialization of robots claiming to have ethical reasoning skills' [142]. In support of the idea, others have cited several reasons for deploying moral AI. The following considerations are relevant.

Inevitability

Admittedly, moral AI could have unwanted side effects, including abuses and misuses, and these dangers lead critics to oppose the development of moral AI. However, some argue that moral AI is inevitable. Nevertheless, the fact that something is inevitable – like death and taxes – does not make it good. The lesson instead is that people will inevitably develop solutions as needs arise. For example, the global shortage of caregivers in hospitals and nursing homes will lead to more robotic caregivers. Such robots will face moral tradeoffs. If they do less harm and better protect patient autonomy if they are able to reason morally, then there is a reason to try to design robotic caregivers to be moral reasoners [143].

Trust

AI cannot do much good if the public does not use it, and use requires trust. When AI uses black box methods to instill ethics, this itself undermines trust, which can lead to diminished use and benefits. However, a responsible and transparent development can increase the public's trust in AI, especially if they know that it is sensitive to their rights and other moral concerns [144].

Complexity

Moral judgment is complex. It is not simply about safety or harm minimization, and includes other factors – including fairness, honesty, autonomy, merit, and roles – that affect what is morally right or wrong. Humans often overlook relevant factors or become confused by complex interactions between conflicting factors. They are also sometimes overcome by emotions, such as dislike of particular groups or fear during military conflicts [145]. Some researchers hope that sophisticated machines can avoid these problems and then make better moral judgments and decisions than humans. To achieve this goal, robots need to be equipped with broad moral competence for unpredictable problems, through proper and responsible design. However, a potential downside is that over-reliance on moral AI could make humans less likely to develop their own moral reasoning skills [118].

Of course, all these issues deserve much more careful consideration [146,147]. It is also crucial to discuss how to govern and regulate moral AI [148,149].

Glossary

Algorithm/algorithmic: these terms are used here in their most comprehensive sense, including but not limited to the connotation of a set of well-defined rules, formulas, formal representations, mathematical or computational models, automated reasoning, or processes that automatically adapt or evolve through learning, typically using data.
Descriptive ethics: the study of patterns of moral beliefs, judgments, and behaviors that actually exist or are produced in the world (often by humans). Frequently pursued by cognitive scientists and psychologists, and often contrasted with normative ethics.
Human ethics: this term is used here to include both descriptive and normative pursuits, in contrast to machine ethics.
Machine ethics: the study of how to design, implement, and generate ethical decision-making in computers, robots, or other automated machines.
Normative ethics: the study of determining what is actually right and wrong. Frequently pursued by moral philosophers, and often contrasted with descriptive ethics.

Can an algorithm help to solve this problem? If so, what is the optimal solution? An initial answer might be: maximize the number of recipients (i.e., matches). However, there are multiple solutions that achieve the maximum number of matches but result in different individuals receiving kidneys. How should we decide among these solutions [11–13]? There are many ways to determine what a fair or justified allocation is and to determine who deserves to get a kidney and who does not. One path forward is to interface with normative ethics and moral psychology and to take inspiration from the factors that have been used by ethicists and by ordinary people when making similar judgments (the question of when and how to integrate the input of these two groups is taken up in the section on normative–descriptive alignment). For the answers to be useful to designing the algorithm, they must be formalized in computational terms. Only following that formalization can algorithmic kidney exchanges be adapted to reflect insights from normative and descriptive **human ethics** (Box 3 for further discussion).

Second, we consider how we can learn about human moral cognition through the quest to develop ethical AI. Such consideration will yield valuable lessons for (normative and descriptive) human ethics. The following example illustrates how developing AI can lend insight into the ethics of medical resource allocation.

Example 2 [medical resource allocation]. *There has been an outbreak of a new, life-threatening disease. Critical care resources – beds, ventilators, and staff – are all in short supply during this surge. To which patients should these resources be allocated?*

Box 3. Extended example – kidney exchange

This box explores how computational ethics can be used to make progress in a particular ethical challenge: matching kidney donors to compatible patients (example 1).

Work relevant to the 'formalize' phase could contribute in several ways. 'Formalizing normative ethics' focuses on representing (perhaps using first-order logic) a set of abstract principles that together form a sound (systematic and consistent) and complete (covering all cases) algorithm. 'Formalizing descriptive ethics' focuses on characterizing (perhaps with computational models) the case-based intuitions of laypersons about which features (e.g., age, critical condition) matter for people when considering different solutions. 'Balancing conflicting values' would formulate this problem as, for example, a combinatorial optimization problem (to maximize the number of recipients, following normative utilitarian views) while adapting weights to reflect the importance of different features (following descriptive preferences for tie breaking), and then applying a computational technique (e.g., an integer program formulation) to solve it.

Suppose, as a result, that an algorithm is developed to prioritize patients based on their general health. Although this may seem reasonable at the outset, work relevant to the 'evaluate' phase will help to study the influence of these decisions at the statistical level, and uncover second-order effects on society. For example, the algorithm may end up by discriminating against poorer participants who are likely to have more comorbidities as a result of their economic disadvantage. Work in 'evaluating machine ethical behavior' could use data about patients to evaluate this possibility.

Suppose that, upon probing this algorithm with different inputs, we find that patients from poorer socioeconomic backgrounds are indeed being significantly disadvantaged by this algorithm. Moreover, work on 'evaluating human ethical behavior' may uncover how this disadvantage may spill over to human ethical decisions in other domains such as employment (e.g., by disadvantaging job candidates experiencing hindrances in their mental capacity as a consequence of kidney failure). Work under the 'formalize' phase (e.g., 'balancing conflicting values') may then develop technical adaptations to mitigate such bias.

Insights about comorbidities may help 'formal descriptive/normative ethics' to realize the considerations implicit in our considered judgments that have not been explicitly articulated. Newly formalized moral principles are then evaluated again, and so on, until formalized moral principles are in coherence with quantified representative moral intuitions.

During a surge caused by a pandemic or mass trauma event, many more patients are likely to benefit from care than can receive that care as a result of scarce human and other resources. In these circumstances, state or local health departments or individual health care institutions may shift to so-called 'crisis standards of care' in which populations, rather than individual patients, become the focus of decision-making, and care is inevitably rationed. Various criteria for the ethical allocation of scarce medical resources have been adopted by different state and private actors [14–17]. To avoid one source of bias, most frameworks recommend that 'triage officers' – clinicians with the relevant expertise (e.g., critical care medicine) but who are not the treating physician of any of the patients in question – are appointed to make allocation decisions. Nevertheless, even in the hands of triage officers, some allocation policies are informal or underspecified, and even the most formal and granular policies may still leave room for individual discretion. As a consequence, actual allocation decisions are often guided by implicit principles that clinicians may not even be aware of, may vary substantially from clinician to clinician, and may be internally inconsistent. As a result, the psychological mechanisms behind much of the moral decision-making of healthcare professionals remains poorly understood.

Can algorithms help to solve this problem? At one level, a machine-learning algorithm could be applied to a dataset of allocation decisions to surface variables that are not formally included in triage frameworks but nevertheless appear to play a role in allocation decisions. These criteria might be ethically sound criteria that serve the goal of transparency, or they may be ethically unfounded biases and heuristics. Human ethicists and clinicians – with public input – would need to decide. At a deeper level, formalizing these factors could enhance our understanding of them. The process of programming triage robots would itself likely entail further clarification of allocation principles and rules. It would require that the principles of moral decision-making used by clinicians are studied, understood, specified, and formalized, and would force

researchers and policymakers to articulate and commit to specific consistent resolutions about what should be done in these crisis situations.

Examples 1 and 2 both show (i) how human ethics can inform machine ethics, and conversely (ii) how machine ethics can inform human ethics. However, for this exchange to work, our theories of ethics need to be stated in computational terms – that is, as an algorithm that translates a specific input (or set of inputs) into a precise output (such as a specific moral judgment or decision).

Lessons from computational cognitive science

Currently, many theories of human cognition and moral decision-making are not formulated in algorithmic terms, and they fail to make fine-grained predictions. Instead, they are stated verbally at a high level of abstraction and make qualitative predictions (e.g., subjects choose X in Y circumstances more often than they do in Z circumstances). Putting our (normative and descriptive) theories of human morality in computational terms allows channels of communication to open with theories of machine ethics; translating our ethical theories into computational terms puts all the ideas in a common language.

Example 1 provides a case for which we may be able to make machine ethical models more human-like by simply collecting data of human (expert) decisions or preferences and training an algorithm on those data (a type of 'bottom-up' approach) [13,18]. The resulting AI system will, in some senses, be 'aligned' with human values. Nevertheless, on other occasions, this approach of simply matching patterns of human judgments will be insufficient, as is the case in example 2 when a novel situation arises for which there are no closely analogous human judgments. In cases like this and others, we may need to incorporate moral reasoning processes and principles into AI systems (in a 'top-down' manner) to be able to generate human-like judgments. The primary way to uncover these principles and processes, and to build computational models of them that are truly generative and explanatory, is through the tools of cognitive science.

Some work of this nature has already begun in earnest [19,20]. For example, recent work has begun to formalize the ability of the mind to flexibly make moral judgments in novel situations when there are no pre-established rules that fit the case [6]. One cognitive mechanism to determine if a novel action is morally permitted is known as 'universalization'. Its precise formulation can be expressed computationally, broadly following the following algorithm – (i) consider how many people are interested and appropriately situated to do the action in question, (ii) estimate the utility consequences of all those people doing the proposed action, and (iii) use a mathematical transformation to estimate moral acceptability [21]. Moreover, computational cognitive scientists have developed formal accounts of moral learning by developing models of how individuals learn the abstract moral principles of their communities through observation and interaction with others around them [22,23], and by describing how children apply rational learning mechanisms to infer moral rules from scant evidence [24]. Other work has focused on how moral judgment operates in mature decision-makers by formally describing how different individuals use different moral strategies or principles in the same situations [25], by developing novel formalisms for intention inference and its role in moral permissibility judgments [26,27], or by formalizing select moral principles and describing how and when they are used by human subjects [21]. Computational accounts of the general properties of human norms [28–30] are also under development. This work is a small but significant step towards developing algorithmic accounts of the human moral mind that, in principle, could be incorporated into AI systems.

We can also use the tools and insights from cognitive science to address newly emerging challenges of using AI systems to make morally charged decisions. For instance, many current

AI tools are opaque, complex, and nonlinear, and it is difficult to understand the workings of the algorithm from looking at its code**.** Likewise, there is no way to peer inside a human mind (or brain) and directly read answers to the questions of how the mind works, why we make decisions the way we do, and what we are likely to do in a novel situation in the future. Moreover, when we provide explanations of our own decisions and behavior, they are only sometimes accurate. To understand how and why humans act and decide the ways they do, cognitive and behavioral scientists have developed sophisticated experimental methods that characterize the workings of the mind by observing how it behaves in carefully controlled environments (e.g., experimental interventions with résumés that are identical except for the applicant's gender), and some have started to apply such methods to study machines (e.g., probing hiring algorithms, price discrimination, or price steering) [31–36]. This may be a way forward to understand the black box algorithms of some AI systems, thus opening new and rich research questions.

Computational reflective equilibrium (CRE)

A key requirement of formalizing ethics in algorithmic terms is consistency – cases that share the same relevant characteristics should result in similar decisions. Consistency is often considered to be a central requirement for normative ethics [37]. Moral principles are expected to give the same answer when confronted with acts and circumstances that have the same set of morally relevant features.

However, consistency can be an elusive requirement, especially when considering a complex system such as the human mind or an AI system that has been developed to assume instrumental roles in society with high-stake responsibilities. First, in such complex systems, multiple parts are interrelated and several units function interdependently. Adjusting some parts to achieve consistency can introduce new inconsistencies in other parts of the system. Second, in such complex systems, the 'input' is high-dimensional, making the set of all possible inputs intractably large. This makes uncovering inconsistencies extremely difficult. Finally, these systems (the human mind and the AI system) are constantly learning, and this fluidity adds complications to meeting this requirement.

To tackle the problem of inconsistency, we propose that researchers think about their work as being governed by a computational version of reflective equilibrium (RE) [8]. RE, that is widely used in the field of moral philosophy [38,39], involves bringing moral principles (commitments to abstract, general moral rules) and moral intuitions (moral judgments of particular cases) into alignment with one another and with scientific knowledge (i.e., removing inconsistencies and contradictions). Crucial to reaching this alignment is the use of examples or cases to which moral principles are applied, and against which moral intuitions are tested (Figure 1). Upon discovering that an intuition and a principle conflict in a given 'use case', achieving RE demands that one or the other is abandoned or modified to resolve the conflict.

In the computational version of reflective equilibrium (CRE) that we propose, alignment is sought between formalized moral principles about general types of actions (principles that are represented/implemented using logic, mathematical formulas, code, and computational models) and quantified representative moral intuitions about particular actions (quantified via the use of statistical tools, collected from representative samples of experts and the public). As Figure 2 shows, alignment between these principles and intuitions is sought based on carefully chosen evaluation metrics (e.g., accuracy, bias, type I/II errors, categorical acceptability) in continuously identified test cases (including toy examples, edge cases, real-life cases, and simulations). Moving from conflicts between formalized principles and quantified intuitions towards a state of CRE would require that one or the other is abandoned or modified (upon collective reflection and cooperative scientific work).

Reflective equilibrium

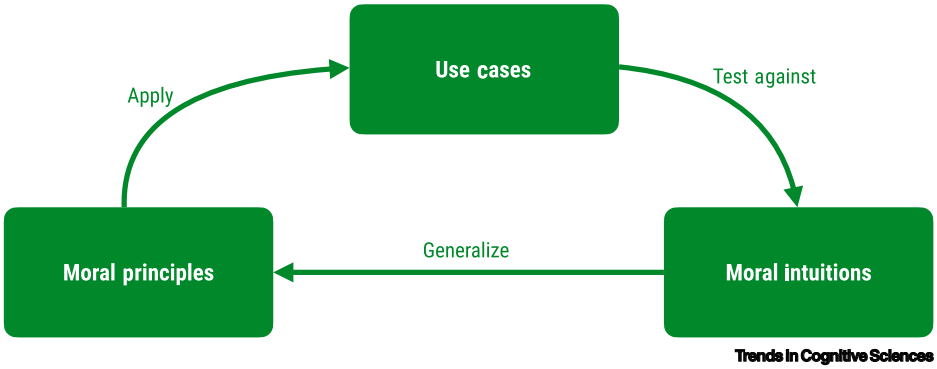

Figure 1. Reflective equilibrium framework. This framework involves bringing moral principles and moral intuitions into alignment with one another through the use of examples or cases to which the moral principles are applied, and against which the moral intuitions are tested.

The aim of computational ethics is to use CRE to resolve contradictions and reach the most coherent, justified, and widely acceptable ethical system for AI (within the bounds of the contributing disciplines), as well as a precise understanding of human ethics. In doing so, the framework proposes that existing and emerging lines of research can contribute to each aspect of CRE.

The computational ethics framework

In this section we suggest a framework to conceptually organize the emerging algorithmically oriented study of ethics and further suggest how this pursuit can move forward in a coordinated fashion. We hope that, by conceptualizing the literature in this logically structured manner, we can help researchers in diverse disciplines to better appreciate how their work fits into the broader framework.

One primary objective of this framework is to understand human ethics computationally such that it can be applied to machine ethics, ultimately allowing for a dynamic exchange between these two pursuits. The key to this endeavor is framing ethics in algorithmic terms. Doing so creates the opportunity to integrate diverse research questions, bring together multiple academic communities, uncover new interdisciplinary topics, and shed light on centuries-old philosophical questions. A long-term outcome of this coordinated effort will be the creation of AI systems that behave in ways that are consistent with our moral values [10,40].

In the following we organize and categorize these research lines into two stages: 'formalize' and 'evaluate'. We also describe how research in these areas is already underway, although it is often siloed in disparate academic domains. We then explain how the research questions in each of the stages can, and should, inform each other. Using this framework encourages CRE – the process of bringing our theoretical commitments (in the 'formalize' stage) in line with our considered judgments (in the 'evaluate' stage) [38].

Formalize

The 'formalize' stage calls for algorithmic characterizations of normative ethics as well as of descriptive ethics. It also addresses the technical and engineering problems of implementing (or deploying) moral principles and theories in artificial systems. Implementation forces designers and engineers to make hard choices [41] – demanding concreteness to replace hedging or

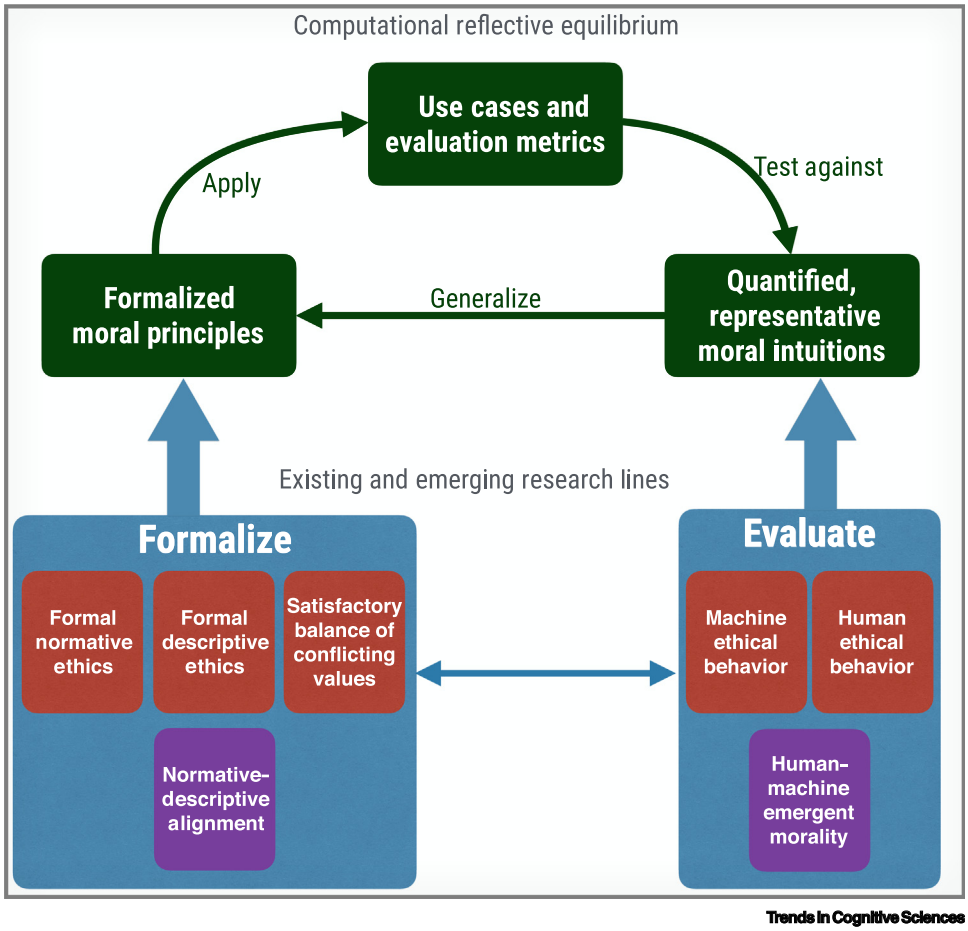

Figure 2. Computational ethics framework. This proposed framework consists of computational reflective equilibrium (a computational counterpart to reflective equilibrium) as the framework backbone, and existing and emerging research lines organized under the two main themes of formalization and evaluation which aim to supplement the content of the framework. Red boxes indicate lines of research that are already in progress. Purple boxes indicate emerging lines of research that are generated from the iterative process described here.

vagueness, navigating discrepancies between normative and descriptive ethics, and accounting for divergent values of different stakeholders (from diverse backgrounds, locations, and demographics, especially those who have been historically under-represented). Moreover, this step is important for stress-testing our models of moral cognition as they are implemented in AI systems because this process requires that the models handle a large range of novel cases.

Formalizing normative ethics

Normative ethics is the enterprise of determining what is morally right and wrong, good and bad – and what one should do on the basis of this information. This research line focuses on representing and formalizing normative moral principles using rules, formal logic, and other knowledge representation techniques or formalisms (e.g., conditional preference networks, CP-Nets) together with automated reasoning (including reasoning under uncertainty and non-monotonic reasoning) [42–48] and formalized conceptions of virtue or moral character [49].

Formalizing descriptive ethics

Descriptive ethics, on the other hand, is the process of determining what people think is morally right or wrong, and how they go about making those judgments. Theories of moral psychology posit that moral cognition is driven by affect [50,51], rules [8,52,53], utility calculations [26,51], agreement-based mechanisms [21,54], or some combination of these factors. This research line focuses on formalizing these views by describing the functions of the system in computational terms or on formalizing specific mechanisms or algorithms used by the mind [19,21–26,28,55].

Balancing conflicting values

Different stakeholders can have different values; moreover, normative values themselves can conflict. Therefore, implementation must resolve moral tradeoffs [13,23,56–63] by using a top-down, bottom-up, or hybrid approach [58,64]. Different types of technology have been proposed, including rule-based approaches [42,43], optimization techniques [13], probabilistic graphical models [22,23,26,65], inductive logic programming [57,66], aggregation rules [59,63,67], and value alignment through inverse reinforcement learning [68–70]. A comprehensive list of technical implementations is given in [71].

Evaluate

This phase involves evaluating machine and human decisions and behavior on normative and descriptive grounds. Those interested in the descriptive side of evaluation may focus on whether AI behavior is consistent with human behavior or whether a model of human moral judgment accurately predicts human judgments in novel cases. Those interested in normative evaluation will instead ask whether an AI performs as it should or whether human choices are biased in some way. This phase also involves evaluating the behavior that results when humans and AI systems interact; although human response to the behavior of AI is often unpredictable before launching the system, human response often determines whether an AI system has an overall positive or negative impact.

Evaluation should be an ongoing process that continually checks the implemented algorithms [72]. As the environment changes (or the algorithm changes as it learns, in some cases), algorithms that were once bringing about positive outcomes or behaving like humans might shift to being morally problematic or exhibiting non-human behavior.

Evaluating machine ethical decisions/behavior

This process seeks to study the machine's decisions/behavior (by using simulations before real-world deployment [13,73], controlled naturalistic environments, or real-world contexts) to ensure that the algorithm performs as intended (either by meeting normative standards or matching descriptive human behavior). This includes evaluating the moral and societal impact of machines and algorithms [74–76], and quantifying the degree to which commercialized algorithms have a negative impact on society and individuals [74,77–80]. It also includes addressing questions such as: how can we use the tools of behavioral and cognitive sciences to study AI (opaque, complex, and nonlinear) algorithms and identify their potential bias [31]? How can we detect when a machine is dishonest, and how can we evaluate situations of acceptable dishonesty by machines (if any) [81]? Finally, machine behavior may be compared to human behavior on comparable tasks to detect where the machine makes non-human-like decisions.

Evaluating human ethical decisions/behavior

This process uses computational tools to evaluate human moral decisions/behavior. For instance, AI systems might be used to improve human moral decision-making by correcting for ignorance, confusion, or bias [18]. In other cases, researchers may use methods of experimental

psychology, data science, and cognitive science to analyze moral behavior of individuals or groups that interact with AI systems. For example, some work has considered whether the news feeds algorithms contribute to spread of misinformation [82–84], political polarization [85,86], and moral outrage [87,88], and to what extent they create social and political bubbles. Other work has begun to describe the cognition behind people's aversion to getting advice from AI in some cases [89] whereas they prefer the advice of AI to humans in others [90]. Such attitudes (aversion or appreciation) influence the ethical behavior of individuals towards machines or other humans. Moreover, people may start to modify their behavior in anticipation of what machines will do and/or be blamed for [91,92].

A key question in the 'evaluate' phase is to determine who decides whether an observed behavior is acceptable. People may sometimes disagree. When there is no clear answer from ethical theory (in the case of true moral dilemmas, for instance, or disagreement among ethical theories), we may draw on the tools of political philosophy (e.g., social choice theory and deliberative democracy) to help to navigate societal and global differences in values [93–97]. Novel methods may prove useful [98,99], such as the direct incorporation of citizen preferences [100]. Although this pursuit is largely beyond the scope of the present article, we see it as an open research question whether computational and AI methods can improve these tools of democratic decision-making.

Emerging research lines

The iterative process of formalization and evaluation will result in increased attention to new lines of research that address questions brought to light by the RE process. We list in the following text two possible examples that are already emerging as prime areas for rapid growth as computational ethics develops.

Normative–descriptive alignment

In some cases, what laypersons judge to be correct is at odds with what experts tend to think is normatively correct or what existing policies require. Recently, organizations such as the IEEE (Institute of Electrical and Electronics Engineers) have begun to develop ethical standards for AI-enabled products which combine normative and descriptive concerns [101,102]. This is also happening within individual companies and regulatory bodies (such as the European Commission [103]). However, these policies and regulations proposed by experts do not always align perfectly with public views. Which values should be implemented into an AI system when normative and descriptive ethics collide?

For example, a German commission was established in 2016 to determine which ethical principles should govern autonomous vehicles [104]. One question it considered was how vehicles should handle cases where one person would be saved at the expense of another. The commission prohibited the use of the victims' ages to make these decisions. However, research conducted to gauge public opinion showed that the public were typically willing to spare the young over the old [105]. Similar disagreements between the public and policymakers have occurred in the spheres of COVID-19 vaccine prioritization [106,107] and how to allocate scarce ECMO (extracorporeal membrane oxygenation) [108] and ventilators [109,110] to hospitalized patients during the COVID-19 pandemic. When developing algorithms to execute or assist these processes, how should these disagreements be resolved?

Research on 'normative–descriptive alignment' may focus on understanding points of overlap and disagreement between normative and descriptive ethics, or on developing computational tools and data-driven approaches that help to bring the two closer, if desirable (of course,

there are disagreements among various normative views themselves; for this reason, it is perhaps useful for computational ethics to start off by focusing on points of normative consensus [111]).

For example, is it possible for descriptive ethics to inform normative ethics [66,112,113], for instance by breaking ties between opposing normative views [114]? Alternatively, there may be a strong normative case for a particular ethical approach, but if descriptive data show that the public will reject that approach wholesale, such a policy might not be feasible or might have undesirable side effects [9,115].

On the other hand, it will sometimes be important to bring lay beliefs into line with normative views. This can sometimes be done simply by informing the public [18]. For example, most people have no idea of how much pain and inconvenience is involved in kidney dialysis while waiting for a transplant or how long someone can live on dialysis. Providing this information might bring the public closer to experts. In other cases, more intensive educational tools may be needed to shift public opinion. Work in this area may focus on ways to build public engagement tools [116], creating educational curricula (e.g., https://exploreaiethics.com), or using a range of techniques to inform the public on which positions are defensible from a normative perspective.

Another possibility is that AI systems should be programmed not in accordance with the actual beliefs of the public (in so far as there are consistent public beliefs, discussed in the section on balancing conflicting values, above) but with what the public would believe under ideal conditions, such as the absence of ignorance, forgetfulness, inattention, confusion, excessive emotion, or bias [18,117]. One promising recent approach to doing this for kidney allocation begins by gathering data on what the public thinks are the morally relevant (e.g., urgency) and morally irrelevant (e.g., race) features of kidney recipients [13,18,117]. The researchers can then gather a large dataset to test the relative weights that people give to each relevant feature (by asking people to choose one recipient out of two possible patients with different profiles). An AI system can be trained to learn these weights, and can then modified in the light of further experimental evidence concerning what unbiased participants would choose. In this way, kidney donation algorithms can reflect the preferences of the public, with some idealization.

Finally, although human moral decision-making may not be ideal, one path forward towards making morally competent machines may be to formalize human moral decision-making with plans of modifying and improving it to meet normative standards. This is especially plausible given that many normative systems are at least 'human-like' even if they do not precisely replicate human judgment in all cases – they often are based on human intuition, follow patterns of human moral reasoning, and broadly capture many judgments that are widely shared. Moreover, because the most advanced AI technologies at this point tend to be data-driven models, practically speaking, we may be able to create AI systems that behave more ethically by starting with models trained on human data.

Human–machine emergent morality

In the 'evaluate' phase it will be crucial to consider the performance of moral principles in multi-agent settings (via modeling and simulation). In many situations, humans and machines will be expected to cooperate on tasks (or at least operate next to each other in a shared environment), and the moral behavior of humans will be influenced and shaped by that of the machines, and vice versa [118]. Consider the case of a human judge who uses a machine to help in making judgments about bail [119]. It may be tempting for the judge to incorporate the input of the machine in a way that would direct the blame to the machine should the decision subsequently become controversial. In turn, machine judges may be designed to make decisions to minimize

their own liability. Either way, lawyers may adapt to such settings by forming strategies around the roles and decisions by human and machine judges, and which were influenced by the 'blame game' in the first place.

Also consider autonomous vehicles: before they fully dominate the roads, autonomous vehicles are expected to drive next to human-driven vehicles, in which case new behavior may arise. For example, human drivers may learn to exploit the cautious behavior of autonomous vehicles by making late and unsafe lane changes or by not yielding the right of way to autonomous vehicles, which may lead to an increase in dangerous maneuvers or aggressive driving. In turn, autonomous vehicles could learn to abuse or threaten human drivers in various ways [120]. Further, how might we characterize situations where it is more beneficial for machines and humans to be dishonest to each other, and how does that impact on their cooperation [81]? These situations and behaviors can be modeled using algorithms, tools, and frameworks from game theory and multi-agent systems [121–123].

Part of this work would focus on the intuitive, analytical, and computational models that are created by humans and machines of each other. Such work would be inspired by theory of mind, a topic that concerns how we infer the mental states of others (e.g. intentions, beliefs, expectations, and desires [124,125]). For an AI to interact and cooperate with humans and to understand and anticipate human moral behavior, we need to program AI with a model of the human moral and social mind (or develop an AI that can learn such a model) [126,127]. Conversely, the emerging topic of 'theory of machine' investigates lay theories of how machines work [90]. People's 'theory of machine' is likely to influence how they behave around machines, how they treat machines, and the extent to which they are willing to trust machines [128]. The dynamic exchange between a machine's theory of the human mind and a human's theory of machine mind is particularly important for the challenge of formalizing the 'ethics of care' [129,130]. Human caretakers extend their goals and values to include the goals and values of another agent, which may be very different from their own.

Concluding remarks

Computational ethics shows how machine ethics and human ethics can be developed in dynamic exchange with one another. The key to the dynamic exchange is to characterize ethics in algorithmic terms. Doing so allows the methods and insights of cognitive science to play a key role in the collaborative and iterative process outlined by this novel interdisciplinary framework.

The proposed framework of computational ethics outlines a research program that focuses on the development and use of technical tools for formalizing and evaluating ethical decision-making, including the implementation of ethical decision-making, the use of mathematics and computational models to represent moral principles, and the use of experimentation and simulations to evaluate moral behavior. Other interdisciplinary efforts (under the umbrella term 'AI ethics') have arisen with different focal points (Box 1), and computational ethics strives to contribute to and draw on insights from these efforts. That being said, the proposed research program is not currently well equipped to address questions that fall outside the intersection of ethics and computation, and it is focused on work that uses computational tools to study ethics.

Moreover, work in many adjacent fields (such as anthropology, sociology, law, psychology, and political science) that is relevant to ethics can inform computational ethics. After all, ethics is shaped by customs, social constructs, institutions, organizations, and recorded cultural knowledge and

Outstanding questions

How can we coherently represent and formalize major normative ethical theories?

How can we develop a model that precisely captures human moral judgments and decision-making?

How can we build a machine that can balance the values of different stakeholders?

How can we evaluate whether a commercialized algorithm reproduces human biases (such as those around race and gender)?

How can we identify whether a commercialized algorithm leads to a concerning change in the behavior of humans (as individuals or groups)?

How can we mitigate and resolve misalignment between normative and descriptive ethics?

How can we build machines that coexist with humans in shared environments without converging on unethical behavior?

Is it ethical for humans to delegate explicitly ethical decisions to machines?

If ethical machines were used in the real world, how would we govern them?

How might the emphasis on algorithms result in understating, ignoring, suppressing, or abandoning current ethical concepts or standards that prove difficult or impossible to formalize? How can we guard against this?

How can we ensure that the researchers whose viewpoints dominate computational ethics are themselves sufficiently diverse that the field's preferred solutions to problems identified will be representative of the global population?

Is thinking about ethics as a cognitive process limiting? How can the perspectives of other fields augment the work of computational ethics?

How can we ensure that the engineering of ethical AI systems would lead to a positive impact on

history. Cognition and computation play only a part in this complex landscape (Boxes 4 and 5 discuss how computational ethics intersects with some of these other disciplines).

Nevertheless, many challenges remain. The framework we have outlined does not guarantee a future with unbiased AI systems. For example, the researchers whose viewpoints are likely to dominate computational ethics (and the CRE process) are themselves mostly WEIRD (Western, educated, industrialized, rich, democratic) [131]. This means there is a risk that the field's preferred solutions to problems identified will not be representative of the global population. In addition, the initial choice of stakeholders and conflicts could shape initial and intermediate computational tools in ways that might be unattractive, inappropriate, or even dangerous.

The emphasis on algorithms and formulating ethics computationally may result in understating, ignoring, suppressing, or abandoning current ethical concepts or standards that prove difficult or impossible to formalize. However, we note that our proposed framework does not aim to replace existing research programs, but instead aims to work together with them in addressing sources of bias or other deficiencies.

Computational ethics is described here in an abstract and preliminary manner. The practical steps this framework will use to achieve RE and to establish a new type of interaction among disciplines remain to be determined. We think of this manuscript as a guide for an ambitious and long-term multi-stakeholder project. Much work remains to be done.

As this paradigm develops further, we are optimistic that a novel field of computational ethics will emerge. Such a field will accelerate the need to develop environments for training new scholars who are computationally skilled, empirically informed, theoretically sophisticated, and ethically

society, and how can we identify cases and situations when it does not?

How can a computational framework deal with situations of unresolvable tradeoffs between incommensurate but genuine different ethical values (e.g., universal utilitarian considerations of the greatest good versus personal commitments and obligations to close relatives)?

Box 4. Intersection with related research fields – law, regulation, and political science

How should we audit, certify, regulate, and manage – that is, govern – ethical machines? We lay out here some ideas on how computational ethics might interact with the fields of law, regulation, and political science [150].

Law and regulation

Codification of ethics and implementation of ethical algorithms must be responsive to law and other regulatory constraints [148,150]. It also must be sensitive to the many complexities of algorithmic decision-making in the modern administrative state [151]. This creates multiple challenges. For example: how can regulation cope with the pace of technological advances in AI and the scalable impact of highly automated systems? To address this question, some work has focused on creating scalable, adaptive, and automated frameworks and systems for AI governance [149]. Much about these regulations – which are just, which are effective – is outside the scope of computational ethics, although directly adjacent. The most relevant scholarship may be work that focuses on computational tools for governing AI. Examples include 'adaptive regulation' – regulatory policies that are designed to be updated automatically or periodically, planned or unplanned [152], and 'oversight programs' – algorithms that can monitor, audit, and hold other algorithms accountable [153]. Such algorithms and systems themselves need to be guided by ethical principles. Other work relevant to computational ethics includes building computational models that capture policy-relevant preferences of humans. This falls into two categories: (i) normative, which requires input from ethicists, legal scholars, and domain experts on AI policy; and (ii) descriptive, which requires input from stakeholders (including public opinion, nationally and globally). Sample topics of inquiry include: what legal constraints and limits should be placed on AI [154]? How should the legal system change to reflect the involvement of AI in decision making (if at all) [155]? How can values such as fairness and empathy be promoted in the emerging digital administrative state [156]?

Political science

When there is no clear answer from ethical theory or philosophy (in the case of true moral dilemmas, for instance, or disagreement among ethical theories), we might also draw on the tools of political science (e.g., social choice theory and deliberative democracy) to help to navigate our differences [93–96]. Novel methods may prove useful [98,99], such as the direct incorporation of citizen preferences [100]. Work under computational ethics intersects well with a variety of formal methods that are used to answer these questions, including research on computational social choice [96] and computational politics [157].

Box 5. Intersection with related research fields – social sciences

Ethics is shaped by customs, social constructs, institutions, organizations, and recorded cultural knowledge and history. We describe here how computational ethics might interact with the perspectives and methods of the social sciences.

Cultural anthropology

Studying human morality from a cultural lens is crucial for theory development, an essential piece of computational ethics. Findings based on data from a small number of culturally similar countries (e.g., the USA, UK, and Canada) have resulted in a biased picture of human morality [131,158]. However, identifying cultural traits and biases alone is only relevant to computational ethics when such factors are embedded in mathematical models, including those of cultural evolution or gene–culture coevolution that would mathematically show how, when, and where some cultural biases of moral behavior can evolve and lead to exhibited robust differences between cultural societies [159–162]. This also implicates the use of simulations and experiments to evaluate, challenge, and validate the assumptions and predictions of such models [159,163,164].

Economics

Because computational ethics ultimately involves tools for ethical decision-making, it intersects with the field of economics in terms of both methodology and substance. Regarding the former, experimental and behavioral economics (e.g., on cooperation, trust, and punishment) [165,166] as well as game theory (both fully rational and evolutionary simulations) can help to model interactions between individuals or groups of humans and/or algorithms [167–172]. Regarding the latter, traditional welfare economics [94,95] complements more recent work on eliciting normative judgments in difficult but policy-relevant settings [173,174].

Sociology and social psychology

Understanding human ethics is incomplete without considering human social behavior, and the success or failure of building ethical machines will be largely manifested through their influence on society and their behavior in social groups. It is thus worth paying close attention to social dynamics that involve humans and machines [175,176]. Accordingly, research lines in the 'evaluate' category would benefit by learning from computational methods that have been developed and used in sociology, especially within its branches of computational social science [177,178], computational sociology [179], complex systems [180], and network science [181,182], and which can contribute to the study social interactions between humans, machines, and a mix of both.

We also acknowledge that many other disciplines will be relevant to the agenda of computational ethics, including philosophy, other humanities and social sciences (e.g., linguistics, education), engineering (e.g., robotics, biological engineering, applied engineering), and business management (e.g., business ethics, IT management, organizational behavior).

sensitive; for awarding grants to the scholars pursuing this agenda; and for organizing workshops and conferences that attract scholars from various disciplines who are working on each of the contributing questions (see Outstanding questions).

Acknowledgments

The authors would like to thank Ariel Procaccia and Iyad Rahwan for their helpful comments on the manuscript. V.C., J.S.B., and W.S-A. are grateful for support from the Templeton World Charity Foundation (grant TWCF0321).

Declaration of interests

The authors declare no conflicts of interest.

References

1. Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman
2. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446
3. Esteva, A. *et al.* (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118
4. Zhu, Z. *et al.* (2016) Traffic-sign detection and classification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2110–2117, IEEE
5. Krizhevsky, A. *et al.* (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90
6. Zhaoping, L. and Li, Z. (2014) *Understanding Vision: Theory, Models, and Data*, Oxford University Press
7. Weiss, Y. *et al.* (2002) Motion illusions as optimal percepts. *Nat. Neurosci.* 5, 598–604
8. Mikhail, J. (2011) *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge University Press
9. Bonnefon, J.-F. *et al.* (2020) The moral psychology of AI and the ethical opt-out problem. In *Ethics of Artificial Intelligence* (Liao, S.M., ed.), pp. 109–126, Oxford University Press
10. Russell, S. (2019) *Human Compatible: AI and the Problem of Control*, Penguin, UK
11. Roth, A.E. *et al.* (2004) Kidney exchange. *Q. J. Econ.* 119, 457–488

12. Bertsimas, D. *et al.* (2013) Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Oper. Res.* 61, 73–87
13. Freedman, R. *et al.* (2020) Adapting a kidney exchange algorithm to align with human values. *Artif. Intell.* 283, 103261
14. White, D.B. *et al.* (2020) *Allocation of Scarce Critical Care Resources during a Public Health Emergency – Executive Summary*, University of Pittsburgh School of Medicine
15. White, D.B. and Lo, B. (2020) A framework for rationing ventilators and critical care beds during the COVID-19 pandemic. *JAMA* 323, 1773–1774
16. Hanfling, D. *et al.* (2020) *Rapid Expert Consultation on Crisis Standards of Care for the COVID-19 Pandemic*, The National Academies Press
17. New York State Task Force on Life and the Law (2015) *Ventilator Allocation Guidelines*, New York State Department of Health
18. Sinnott-Armstrong, W. and Skorburg, J.A. (2021) How AI can AID bioethics. *Journal of Practical. Ethics* 9, jpe1175
19. Crockett, M.J. (2016) How formal models can illuminate mechanisms of moral judgment and decision making. *Curr. Dir. Psychol. Sci.* 25, 85–90
20. Mikhail, J. (2009) Moral grammar and intuitive jurisprudence: a formal model of unconscious moral and legal knowledge. In *Psychology of Learning and Motivation* (50) (Ross, B.H., ed.), pp. 27–100, Academic Press
21. Levine, S. *et al.* (2020) The logic of universalization guides moral judgment. *Proc. Natl. Acad. Sci. U. S. A.* 117, 26158–26169
22. Kleiman-Weiner, M. *et al.* (2017) Learning a commonsense moral theory. *Cognition* 167, 107–123
23. Kim, R. *et al.* (2018) A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Furman, J. *et al.*, eds), pp. 197–203, Association for Computing Machinery
24. Nichols, S. *et al.* (2016) Rational learners and moral rules. *Mind Lang.* 31, 530–554
25. van Baar, J.M. *et al.* (2019) The computational and neural substrates of moral strategies in social decision-making. *Nat. Commun.* 10, 1483
26. Kleiman-Weiner, M. *et al.* (2015) Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (Noelle, D.C. *et al.*, eds), pp. 1123–1128, Cognitive Science Society
27. Levine, S. *et al.* (2018) The mental representation of human action. *Cogn. Sci.* 42, 1229–1264
28. Malle, B.F. *et al.* (2019) Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 21–27, Association for Computing Machinery
29. Malle, B.F. *et al.* (2021) Cognitive properties of norm representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43), pp. 819–825, Cognitive Science Society
30. Malle, B. (2020) Graded representations of norm strength. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (Denison, S. *et al.*, eds), pp. 3342–3348, Cognitive Science Society
31. Rahwan, I. *et al.* (2019) Machine behaviour. *Nature* 568, 477–486
32. Wang, D. *et al.* (2019) Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Brewster, S. *et al.*, eds), pp. 1–15, Association for Computing Machinery
33. Correll, S.J. *et al.* (2007) Getting a job: is there a motherhood penalty? *Am. J. Sociol.* 112, 1297–1339
34. Kübler, D. *et al.* (2018) Gender discrimination in hiring across occupations: a nationally-representative vignette study. *Labour Econ.* 55, 215–229
35. Hannak, A. *et al.* (2014) Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference* (Williamson, C. *et al.*, eds), pp. 305–318, Association for Computing Machinery
36. Chen, L. *et al.* (2016) An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th International Conference on World Wide Web* (Bourdeau, J. *et al.*, eds), pp. 1339–1349, International World Wide Web Conferences Steering Committee
37. Hare, R.M. (1981) *Moral Thinking: Its Levels, Method, and Point*, Oxford University Press
38. Rawls, J. (1971) *A Theory of Justice*, Harvard University Press
39. Nichols, P. (2012) Wide reflective equilibrium as a method of justification in bioethics. *Theor. Med. Bioeth.* 33, 325–341
40. Christian, B. (2021) *The Alignment Problem: How Can Machines Learn Human Values?* Atlantic Books
41. Wilson, R.C. and Collins, A.G. (2019) Ten simple rules for the computational modeling of behavioral data. *Elife* 8, e49547
42. Van Den Hoven, J. and Lokhorst, G. (2002) Deontic logic and computer-supported computer ethics. *Metaphilosophy* 33, 376–386
43. Hooker, J.N. and Kim, T.W.N. (2018) Toward non-intuition-based machine and artificial intelligence ethics. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Furman, J. *et al.*, eds), pp. 130–136, Association for Computing Machinery
44. Leben, D. (2020) Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Markham, A. *et al.*, eds), pp. 86–92, Association for Computing Machinery
45. Awad, E. *et al.* (2020) When is it morally acceptable to break the rules? A Preference-Based Approach. In *12th Multidisciplinary Workshop on Advances in Preference Handling* (Endres, M. *et al.*, eds), IOS Press
46. Loreggia, A. *et al.* (2018) Preferences and ethical principles in decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Furman, J. *et al.*, eds), pp. 222, Association for Computing Machinery
47. Limarga, R. *et al.* (2020) Non-monotonic reasoning for machine ethics with situation calculus. In *AI 2020: Advances in Artificial Intelligence* (Gallagher, M. *et al.*, eds), pp. 203–215, Springer
48. Pagnucco, M. *et al.* (2021) Epistemic Reasoning for Machine Ethics with Situation Calculus. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Fourcade, M. *et al.*, eds), pp. 814–821, Association for Computing Machinery
49. Wallach, W. and Vallor, S. (2020) Moral machines: from value alignment to embodied virtue. In *Ethics of Artificial Intelligence* (Liao, S.M., ed.), pp. 383–412, Oxford University Press
50. Haidt, J. (2012) *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*, Cambridge University Press
51. Greene, J. (2014) *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*, Atlantic Books
52. Nichols, S. and Mallon, R. (2006) Moral dilemmas and moral rules. *Cognition* 100, 530–542
53. Levine, S. and Leslie, A. (2021) Preschoolers use the means-ends structure of intention to make moral judgments. *PsyArXiv* Published online September 16, 2020. http://dx.doi.org/10.31234/osf.io/np9a5
54. Baumard, N. *et al.* (2013) A mutualistic approach to morality: the evolution of fairness by partner choice. *Behav. Brain Sci.* 36, 59–78
55. Crockett, M.J. (2013) Models of morality. *Trends Cogn. Sci.* 17, 363–366
56. Anderson, M. and Anderson, S.L. (2011) *Machine Ethics*, Cambridge University Press
57. Anderson, M. and Anderson, S.L. (2018) GenEth: a general ethical dilemma analyzer. *Paladyn, J. Behav. Robot.* 9, 337–357
58. Wallach, W. and Allen, C. (2009) *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press
59. Noothigattu, R. *et al.* (2018) A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence* (McIlraith, S. and Weinberger, K., eds), pp. 1587–1594, AAAI Press
60. Thornton, S.M. *et al.* (2017) Incorporating ethical considerations into automated vehicle control. *IEEE Trans. Intell. Transp. Syst.* 18, 1429–1439

61. Thornton, S.M. *et al.* (2018) Value sensitive design for autonomous vehicle motion planning. In *2018 IEEE Intelligent Vehicles Symposium* (Wang, F.-Y. *et al.*, eds), pp. 1157–1162, IEEE
62. Kramer, M.F. *et al.* (2018) When do people want AI to make decisions? In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Furman, J. *et al.*, eds), pp. 204–209, Association for Computing Machinery
63. Conitzer, V. *et al.* (2017) Moral decision making frameworks for artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Singh, S. and Markovitch, S., eds), pp. 4831–4835, AAAI Press
64. Petersen, S. (2020) Machines learning values. In *Ethics of Artificial Intelligence* (Liao, S.M., ed.), pp. 413–435, Oxford University Press
65. Kleiman-Weiner, M. *et al.* (2017) Constructing social preferences from anticipated judgments: when impartial inequity is fair and why? In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (Gunzelmann, G. *et al.*, eds), pp. 676–681, Cognitive Science Society
66. Awad, E. *et al.* (2020) An approach for combining ethical principles with public opinion to guide public policy. *Artif. Intell.* 287, 103349
67. Lee, M.K. *et al.* (2019) WeBuildAI: participatory framework for algorithmic governance. *Proc. ACM Hum. Comput. Interact.* 3, 1–35
68. Russell, S. (1998) Learning agents for uncertain environments. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (Bartlett, P. and Mansour, Y., eds), pp. 101–103, Association for Computing Machinery
69. Hadfield-Menell, D. *et al.* (2016) Cooperative inverse reinforcement learning. *Adv. Neural Inf. Proces. Syst.* 29, 3909–3917
70. Noothigattu, R. *et al.* (2019) Teaching AI agents ethical values using reinforcement learning and policy orchestration. *IBM J. Res. Dev.* 63, 2:1–2:9
71. Tolmeijer, S. *et al.* (2020) Implementations in machine ethics. *ACM Comput. Surv.* 53, 1–38
72. Babic, B. *et al.* (2019) Algorithms on regulatory lockdown in medicine. *Science* 366, 1202–1204
73. Choi, H. *et al.* (2021) On the use of simulation in robotics: opportunities, challenges, and suggestions for moving forward. *Proc. Natl. Acad. Sci. U. S. A.* 118
74. O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books
75. Wachter-Boettcher, S. (2017) *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, W.W. Norton & Company
76. Thieme, A. *et al.* (2020) Machine learning in mental health. *ACM Trans. Comput. Human Interact.* 27, 1–53
77. Buolamwini, J. and Gebru, T. (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (81) (Friedler, S.A. and Wilson, C., eds), pp. 77–91, PMLR
78. Obermeyer, Z. *et al.* (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453
79. Dressel, J. and Farid, H. (2018) The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* 4, eaao5580
80. Rambachan, A. *et al.* (2020) An economic perspective on algorithmic fairness. *AEA Papers Proc.* 110, 91–95
81. Ishowo-Oloko, F. *et al.* (2019) Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Mach. Intell.* 1, 517–521
82. Vosoughi, S. *et al.* (2018) The spread of true and false news online. *Science* 359, 1146–1151
83. Aral, S. (2020) *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy and Our Health – And How We Must Adapt*, Harper Collins
84. Pennycook, G. and Rand, D.G. (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2521–2526
85. Bakshy, E. *et al.* (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 1130–1132
86. Lee, J.K. *et al.* (2014) Social media, network heterogeneity, and opinion polarization. *J. Commun.* 64, 702–722
87. Crockett, M.J. (2017) Moral outrage in the digital age. *Nat. Hum. Behav.* 1, 769–771
88. Brady, W.J. *et al.* (2021) How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* 7, eabe5641
89. Dietvorst, B.J. *et al.* (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126
90. Logg, J.M. *et al.* (2019) Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151, 90–103
91. Awad, E. *et al.* (2020) Drivers are blamed more than their automated cars when both make mistakes. *Nat. Hum. Behav.* 4, 134–143
92. Kleinberg, J. *et al.* (2020) Algorithms as discrimination detectors. *Proc. Natl. Acad. Sci. U. S. A.* 117, 30096–30100
93. Arrow, K.J. (1950) A difficulty in the concept of social welfare. *J. Polit. Econ.* 58, 328–346
94. Sen, A. (1999) The possibility of social choice. *Am. Econ. Rev.* 89, 349–378
95. Arrow, K.J. *et al.* (2010) *Handbook of Social Choice and Welfare*, Elsevier
96. Brandt, F. *et al.* (2016) *Handbook of Computational Social Choice*, Cambridge University Press
97. Kahng, A. *et al.* (2019) Statistical foundations of virtual democracy. In *Proceedings of the 36th International Conference on Machine Learning* (97) (Chaudhuri, K. and Salakhutdinov, R. *et al.*, eds), pp. 3173–3182, PMLR
98. Guerrero, A.A. (2014) Against elections: the lottocratic alternative. *Philos Public Aff* 42, 135–178
99. Munn, N. (2019) Democracy without voting. In *2019 American Philosophical Association Pacific Division Conference*, University of Waikato Research Commons
100. Weernink, M.G.M. *et al.* (2014) A systematic review to identify the use of preference elicitation methods in healthcare decision making. *Pharmaceut. Med.* 28, 175–185
101. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2018) *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, IEEE
102. Winfield, A. (2019) Ethical standards in robotics and AI. *Nat. Electron.* 2, 46–48
103. European Commission (2020) *On Artificial Intelligence – A European Approach to Excellence and Trust (White Paper)*, European Commission
104. Luetge, C. (2017) The German ethics code for automated and connected rriving. *Philos. Technol.* 30, 547–558
105. Awad, E. *et al.* (2018) The moral machine experiment. *Nature* 563, 59–64
106. Persad, G. *et al.* (2021) Public perspectives on COVID-19 vaccine prioritization. *JAMA Netw. Open* 4, e217943
107. Duch, R. *et al.* (2021) Citizens from 13 countries share similar preferences for COVID-19 vaccine allocation priorities. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2026382118
108. Dao, B. *et al.* (2021) Ethical factors determining ECMO allocation during the COVID-19 pandemic. *BMC Med. Ethics* 22, 70
109. Asghari, F. *et al.* (2021) Priority setting of ventilators in the COVID-19 pandemic from the public's perspective. *AJOB Empir. Bioeth.* 12, 155–163
110. Wilkinson, D. *et al.* (2020) Which factors should be included in triage? An online survey of the attitudes of the UK general public to pandemic triage dilemmas. *BMJ Open* 10, e045593
111. Liao, S.M. (2020) A short introduction to the ethics of artificial intelligence. In *Ethics of Artificial Intelligence* (Liao, S.M., ed.), pp. 1–42, Oxford University Press
112. Kim, T.W. and Donaldson, T. (2018) Rethinking right: moral epistemology in management research. *J. Bus. Ethics* 148, 5–20
113. Weaver, G.R. and Trevino, L.K. (1994) Normative and empirical business ethics: separation, marriage of convenience, or marriage of necessity? *Bus. Ethics Q.* 4, 129–143
114. Savulescu, J. *et al.* (2019) From public preferences to ethical policy. *Nat. Hum. Behav.* 3, 1241–1243

115. Everett, J.A.C. *et al.* (2016) Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* 145, 772–787
116. Forum for Ethical AI (2019) *Democratising Decisions about Technology: A Toolkit*, Forum for Ethical AI
117. Skorburg, J.A. *et al.* (2020) AI methods in bioethics. *AJOB Empir. Bioeth.* 11, 37–39
118. Vallor, S. (2016) *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press
119. Berk, R. (2012) *Criminal Justice Forecasts of Risk: A Machine Learning Approach*, Springer Science & Business Media
120. Sadigh, D. *et al.* (2016) Planning for autonomous cars that leverage effects on human actions. In *Proceedings of Robotics: Science and Systems XII* (Hsu, D. *et al.*, eds), MIT Press
121. Crandall, J.W. *et al.* (2018) Cooperating with machines. *Nat. Commun.* 9, 233
122. Roughgarden, T. (2005) *Selfish Routing and the Price of Anarchy*, MIT Press
123. Papadimitriou, C.H. (2001) Algorithms, games, and the internet. In *Automata, Languages and Programming* (Orejas, F. *et al.*, eds), pp. 1–3, Springer
124. Dennett, D.C. (1989) *The Intentional Stance*, MIT Press
125. Saxe, R. and Young, L. (2013) Theory of mind: how brains think about thoughts. In *The Oxford Handbook of Cognitive Neuroscience* (Vol. 2) (Ochsner, K.N. and Kosslyn, S., eds), pp. 204–213, Oxford University Press
126. Breazeal, C.L. (2002) *Designing Sociable Robots*, MIT Press
127. Breazeal, C. (2003) Emotion and sociable humanoid robots. *Int. J. Human-Comput. Stud.* 59, 119–155
128. Bigman, Y.E. *et al.* (2019) Holding robots responsible: the elements of machine morality. *Trends Cogn. Sci.* 23, 365–368
129. Gopnik, A. (2016) *The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children*, Random House
130. Vallor, S. (2011) Carebots and caregivers: sustaining the ethical ideal of care in the twenty-first century. *Philos. Technol.* 24, 251–268
131. Henrich, J. *et al.* (2010) The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83
132. Anderson, M. and Anderson, S.L. (2006) Guest editors' introduction: machine ethics. *IEEE Intell. Syst.* 21, 10–11
133. Veruggio, G. (2004) A proposal for a roboethics. In *1st International Symposium on Roboethics: The Ethics, Social, Humanitarian, and Ecological Aspects of Robotics*, Roboethics.org
134. Tzafestas, S.G. (2015) *Roboethics: A Navigating Overview*, Springer
135. Wieringa, M. (2020) What to account for when accounting for algorithms. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Hildebrandt, M. *et al.*, eds), pp. 1–18, Association for Computing Machinery
136. Weller, A. (2019) Transparency: motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Samek, W. *et al.*, eds), pp. 23–40, Springer
137. Mehrabi, N. *et al.* (2019) A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 1–35
138. Tomašev, N. *et al.* (2020) AI for social good: unlocking the opportunity for positive impact. *Nat. Commun.* 11, 2468
139. Dignum, V. (2019) *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer
140. Wachter, S. *et al.* (2017) Transparent, explainable, and accountable AI for robotics. *Science. Robotics* 2, eaan6080
141. Wachter, S. *et al.* (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. J. Law Technol.* 31, 841–887
142. van Wynsberghe, A. and Robbins, S. (2019) Critiquing the reasons for making artificial moral agents. *Sci. Eng. Ethics* 25, 719–735
143. Poulsen, A. *et al.* (2019) Responses to a critique of artificial moral agents. *ArXiv* Published online March 17, 2019. https://arxiv.org/abs/1903.07021
144. Shin, D. (2020) User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. *J. Broadcast. Electron. Media* 64, 541–565
145. Arkin, R. (2009) *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall/CRC
146. Vanderelst, D. and Winfield, A. (2018) The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (Furman, J. *et al.*, eds), pp. 317–322, Association for Computing Machinery
147. Cave, S. *et al.* (2019) Motivations and risks of machine ethics. *Proc. IEEE* 107, 562–574
148. Winfield, A.F. *et al.* (2019) Machine ethics: the design and governance of ethical ai and autonomous systems. *Proc. IEEE* 107, 509–517
149. Falco, G. *et al.* (2021) Governing AI safety through independent audits. *Nat. Mach. Intell.* 3, 566–571
150. Winfield, A.F.T. and Jirotka, M. (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philos. Trans. A Math. Phys. Eng. Sci.* 376, 20180085
151. Coglianese, C. and Lehr, D. (2016) Regulating by robot: administrative decision making in the machine-learning era. *Geo. LJ* 105, 1147
152. Bennear, L.S. and Wiener, J.B. (2019) Adaptive regulation: instrument choice for policy learning over time (Draft working paper). Published online February 12, 2019. https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/Regulation%20-%20adaptive%20reg%20-%20Bennear%20Wiener%20on%20Adaptive%20Reg%20Instrum%20Choice%202019%2002%2012%20clean.pdf
153. Etzioni, A. and Etzioni, O. (2016) AI assisted ethics. *Ethics Inf. Technol.* 18, 149–156
154. Organisation for Economic Co-operation and Development (2019) *Artificial Intelligence in Society*, OECD
155. Price 2nd, W.N. *et al.* (2019) Potential liability for physicians using artificial intelligence. *JAMA* 322, 1765–1766
156. Ranchordas, S. (2021) Empathy in the digital administrative state. *Duke Law J.* 71, 1341–1389
157. Tufekci, Z. (2014) Engineering the public: big data, surveillance and computational politics. *First Monday* 19, fm.v19i7.4901
158. Muthukrishna, M. *et al.* (2020) Beyond Western, educated, industrial, rich, and democratic (WEIRD) psychology: measuring and mapping scales of cultural and psychological distance. *Psychol. Sci.* 31, 678–701
159. Henrich, J. and Boyd, R. (1998) The evolution of conformist transmission and the emergence of between-group differences. *Evol. Hum. Behav.* 19, 215–241
160. McElreath, R. and Henrich, J. (2007) Modeling cultural evolution. In *Oxford Handbook of Evolutionary Psychology* (Barrett, L. and Dunbar, R., eds), pp. 571–586, Oxford University Press
161. Schaller, M. and Muthukrishna, M. (2021) Modeling cultural change: Computational models of interpersonal influence dynamics can yield new insights about how cultures change, which cultures change more rapidly than others, and why. *Am. Psychol.* 76, 1027–1038
162. Muthukrishna, M. and Schaller, M. (2020) Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change. *Personal. Soc. Psychol. Rev.* 24, 103–120
163. Wakano, J.Y. and Aoki, K. (2007) Do social learning and conformist bias coevolve? Henrich and Boyd revisited. *Theor. Popul. Biol.* 72, 504–512
164. Eriksson, K. *et al.* (2007) Critical points in current theory of conformist social learning. *J. Evol. Psychol.* 5, 67–87
165. Rand, D.G. *et al.* (2009) Positive interactions promote public cooperation. *Science* 325, 1272–1275
166. Jordan, J.J. *et al.* (2016) Third-party punishment as a costly signal of trustworthiness. *Nature* 530, 473–476
167. Sigmund, K. and Nowak, M.A. (1999) Evolutionary game theory. *Curr. Biol.* 9, R503–R505
168. Axelrod, R. and Hamilton, W.D. (1981) The evolution of cooperation. *Science* 211, 1390–1396
169. Letchford, J. *et al.* (2008) An 'ethical' game-theoretic solution concept for two-player perfect-information games. In *International Workshop on Internet and Network Economics (WINE)* (Papadimitriou, C. and Zhang, S., eds), pp. 696–707, Springer
170. Davoust, A. and Rovatsos, M. (2020) Social contracts for non-cooperative games. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Markham, A. *et al.*, eds), pp. 43–49, Association for Computing Machinery

171. Schillo, M. *et al.* (2000) Using trust for detecting deceitful agents in artificial societies. *Appl. Artif. Intell.* 14, 825–848
172. Wolpert, D. *et al.* (2011) Strategic choice of preferences: the persona model. *BE J. Theor. Econom.* 11, 00001022021935170415 93
173. Jamison, J.C. (2016) Perceptions regarding the value of life before and after birth. *Reprod. Syst. Sex. Disord.* 4, 100195
174. Alesina, A. *et al.* (2017) *Intergenerational Mobility and Preferences for Redistribution (NBER Working Paper 2037)*, National Bureau of Economic Research
175. Fast, N.J. and Schroeder, J. (2020) Power and decision making: new directions for research in the age of artificial intelligence. *Curr. Opin. Psychol.* 33, 172–176
176. Dellaert, B.G.C. *et al.* (2020) Consumer decisions with artificially intelligent voice assistants. *Mark. Lett.* 31, 335–347
177. Lazer, D. *et al.* (2009) Computational social science. *Science* 323, 721–723
178. Lazer, D.M.J. *et al.* (2020) Computational social science: obstacles and opportunities. *Science* 369, 1060–1062
179. Macy, M.W. and Willer, R. (2002) From factors to actors: computational sociology and agent-based modeling. *Annu. Rev. Sociol.* 28, 143–166
180. Bar-Yam, Y. *et al.* (1998) Dynamics of complex systems (studies in nonlinearity). *Comput. Phys.* 12, 335–336
181. Newman, M. (2018) *Networks*, Oxford University Press
182. Newman, M. *et al.* (2011) *The Structure and Dynamics of Networks*, Princeton University Press