Bad machines corrupt good morals

Nils Köbis, Jean-François Bonnefon, Iyad Rahwan

WORKING PAPERS

N° 1212

May 2021

“Bad machines corrupt good morals”

Nils Köbis, Jean-François Bonnefon & Iyad Rahwan

Toulouse School of Economics

Bad machines corrupt good morals

Nils Köbis[1], Jean-François Bonnefon[2] & Iyad Rahwan[1]

[1] Center for Humans and Machines, Max Planck Institute for Human Development, Lentzealle 94, Berlin 14195, Germany

[2] Toulouse School of Economics (TSM-R, CNRS), University of Toulouse Capitole, Toulouse 31015, France

The authors thank Anna Bouza da Costa for designing the illustrations. JFB acknowledges support from the Institute for Advanced Study in Toulouse, the grant ANR-19-PI3A-0004 Artificial and Natural Intelligence Toulouse Institute, and the grant ANR-17-EURE-0010 Investissements d'Avenir.

Abstract

Machines powered by Artificial Intelligence (AI) are now influencing the behavior of humans in ways that are both like and unlike the ways humans influence each other. In light of recent research showing that other humans can exert a strong corrupting influence on people's ethical behavior, worry emerges about the corrupting power of AI agents. To estimate the empirical validity of these fears, we review the available evidence from behavioral science, human-computer interaction, and AI research. We propose that the main social roles through which both humans and machines can influence ethical behavior are (a) role model, (b) advisor, (c) partner, and (d) delegate. When AI agents become *influencers* (role models or advisors), their corrupting power may not exceed (yet) the corrupting power of humans. However, AI agents acting as *enablers* of unethical behavior (partners or delegates) have many characteristics that may let people reap unethical benefits while feeling good about themselves, indicating good reasons for worry. Based on these insights, we outline a research agenda that aims at providing more behavioral insights for better AI oversight.

Keywords: machine behavior; behavioral ethics; corruption; artificial intelligence

Although people generally prefer to behave ethically[1], they face many temptations to break rules for private benefits[2], especially when these ethical violations are facilitated by other individuals[3], who may be advisors, delegates, or active cooperation partners. Given that Artificial Intelligence (AI) agents (see Box 1 for our use of this term) increasingly act in advisory, delegation, or cooperation roles[4,5], should we fear that AI may exert a corrupting force on human ethical behavior?

BOX 1: What do we not mean by "AI agents"?

AI encompasses various techniques in computer science (e.g., machine learning) that allow for the autonomous execution of tasks that used to be reserved for humans[6,7]. Because of this autonomy of execution, some instantiations of AI-powered technology are commonly referred to as AI agents[8], and we will adopt this terminology in the current manuscript. It is important to note, however, that using the term "AI agent" should not carry any presupposition that the AI can be held morally or legally responsible for the outcomes of its tasks[9]. While liability issues can become complicated when AI technology increases in sophistication[10], our default stance in the article is that humans (e.g., programmers, designers, users) are always ultimately responsible for the behavior of AI agents and its consequences[11].

Of course, any new technology can be used for unethical purposes by savvy criminals, and such is the case for AI. For example, scammers made use of AI to create hyper-realistic deepfakes defrauding companies, with the damage in one single case amounting to over $220,000[12]. AI can also tempt honest citizens into unethical behavior by merely making cheating easier. For example, students have successfully used powerful Natural Language Generation

(NLG) algorithms to craft their essays[13]. Finally, even if AI does not directly offer the means to cheat, it may still give inappropriate advice or provide an example of inappropriate behavior. Consider how traders might imitate manipulative market strategies from algorithmic traders[14], or that by now, many adolescents seek guidance on ethical dilemmas from their personal AI assistants or chatbot friends[15]. With more than 100 million people using AI-powered personal assistants like Siri or Alexa, the potential for such an inappropriate influence cannot be ignored.

The trajectory of powerful AI tools quickly becoming widely accessible triggers fear and worry[16]. For example, a recent report by the EU commission highlights that "citizens (...) worry that AI can have unintended effects or even be used for malicious purposes"[17]. Yet, such pessimistic views about new technologies are nothing new[18]. People have felt threatened by machines for centuries[19], and tend to meet innovations with exaggerated skepticism[6] and fear-mongering. Developing a cool-headed policy agenda requires an evidence-based assessment about which of the fears that AI will corrupt human ethical behavior are warranted[20]. Put differently, developing effective AI *oversight* requires an overview of available empirical *insights*.

A growing literature in behavioral science examines how humans corrupt each other, yet research on how intelligent machines affect human ethical behavior remains scant. Based on a review of current findings on the *human* social forces shaping (un)ethical behavior, we identify four main roles through which AI agents might exert a corrupting force on human ethical behavior: (a) role model, (b) advisor, (c) partner, and (d) delegate. We critically evaluate the potential severity of the AI agents' corrupting force for each of these roles. Based on the identified gaps in knowledge, we sketch a research agenda on how interacting with and through AI agents affects human ethical behavior.

How can people and AI agents corrupt ethical behavior?

Unethical behavior is commonly defined as "acts that have harmful effects on others and are either illegal or morally unacceptable to the larger community"[21, based on 22]. Behavioral ethics investigates how people behave when faced with the temptation to act unethically and, in particular, how they weigh the personal benefits and risks of such behavior[23–26], either in a material sense (e.g., financial gains, legal punishment) or a psychological sense (e.g., self-image)[27–31]. Meta-analyses of *individual* forms of unethical behavior [1,32] (situations in which people face temptations by themselves) indicate that people generally break ethical rules only to the extent that they can justify it[1,32]. The behavioral research we will focus on is concerned with the power of *social* forces shaping (un)ethical behavior[3,33–35,for a meta-analyis, see 36], that is, the corrupting influence people can have on other people. Likewise, there is ample research on the harm that AI agents can themselves inflict[37], for example, by reproducing biases[38,39], fostering internet addiction[40,41], or accelerating the spread of false information[42]; but the research we will focus on is concerned with the way AI agents can perform social roles that make people harm each other. We now review in turn four such social roles (See Fig. 1 for a summary).

Fig. 1. Illustration of the main roles through which humans and AI can corrupt human ethical behavior, grouped along the left panel for AI in the role of an influencer (role model & advisor) and along the right panel for AI being an enabler (partner & delegate). Boxes summarize the main fears and mechanisms attached to each role. Color coding indicates the strength of the corrupting force of AI, either not reaching human levels yet (green), reaching but not surpassing human levels (yellow), and surpassing human levels (red).

Role Model

When deciding whether to break or adhere to ethical rules, people often consider what others would do to gauge the normative standards of the particular situation[43]. Social norms theory outlines that such perceptions fall into two main categories: Injunctive norms convey information about whether a particular course of action is considered acceptable and descriptive norms outline whether a behavior is deemed to be widespread[44–46]. Experimental research reveals that such normative perceptions in general, and perceived descriptive norms in particular, strongly influence unethical behavior as people often imitate others. Put differently, when perceiving that others break versus adhere to ethical rules, people often follow suit [2,47,48,for a review see 49].

In the digital world, people are exposed to both human and machine behavior[4]. A machine that would display unethical or inappropriate behavior may therefore shift people's perception of what is acceptable or appropriate. There is only mixed evidence (and negative on balance) that adult humans might conform to machines the same way they conform to humans, though, and this evidence is restricted to non-moral behaviors[50–55].

Note that even if people were shown not to conform to machine role models, the possibility would remain for them to be influenced by machines passing as humans online[56,57]; for example, when online traders imitate manipulative trading strategies that, unbeknownst to them, are executed by algorithmic traders[14]. There is concerning evidence that children, more than adults, may be influenced by machine role models[52], in a way that makes them change their perception of moral transgressions[58,59]. Overall, though, the current state of experimental evidence would

suggest that machines acting as unethical role models are less of a concern than humans acting in the same capacity.

Advisor

People can have a more direct corrupting influence than role-models when giving advice to act (un)ethically. Behavioral research has established that people do tend to follow advice and orders, particularly when they come from authority figures[60, see replication 61]. Advisors who have a vested interest in an unethical course of action may encourage advisees to act unethically, and research shows that such advice may lead advisees to break ethical rules, especially if they can benefit from this behavior themselves[62,63].

AI agents may follow persuasive goals[41,64], such as giving advice and recommendations[65]. This trend of AI agents swaying people's behavior is only increasing. In fact, Amazon's chief scientist Rohit Prasad remarked that people's relationship with their Alexas "keeps growing from more of an assistant to advisor"[66]. In parallel to home assistants, millions of users engage with advice-giving conversational agents like Replika (replika.ai), trained on large amounts of data reflecting personalized preferences[67]. Companies like Gong (Gong.io) use NLP and machine learning to analyze big data of recorded sales conversations in order to provide advice to salespeople about how to improve their performance. Given the difficulty of training AI advisors to be impartial moral guides[37,68], however we define this standard, their personalized advice could lead people to break ethical rules. This concern is compounded by the fact that people may feel "algorithmically dumbfounded" by AI advice, in the sense that they may be complacent to follow it, even if they anticipate its (ethical) shortcomings[69].

Are these fears warranted? Even if machines were to give unethical advice, a phenomenon

which has yet to be documented, we know that people state that they are not necessarily keen on following algorithmic recommendations in non-technical domains[70,71]. While this aversion could, in theory, dampen the effect of unethical machine advice, recent evidence from a large-scale experiment tells a different story[72]. This experiment directly compared the effect of human and AI advice on people's actual (un)ethical behavior - not their stated preferences. The results revealed that AI and human advice exerted an equally strong corrupting effect on people's willingness to break ethical rules for profit. These initial findings suggest that we should take seriously the possibility that humans may act based on corrupting advice from AI agents, as seriously as we take the possibility that humans may receive and follow corrupting advice from other humans.

Partner

People can be corrupted by unethical advisors, but they can also corrupt each other, becoming partners in crime[3,34]. This happens when two or more individuals act together toward a mutually beneficial outcome, realize that this outcome can be achieved through unethical means, and collaborate in these unethical means[26,33]. Behavioral research shows that people are more likely to act unethically in these collaborative conditions than when they face temptations alone[3,34]. Besides people having a general tendency to conform to others[73,see for a replication 74], another reason for the appeal of collaborative corruption is that the salient, positive effect of helping one another can overshadow the negative impact of harming some third-party[43,75]. This skewed balance facilitates justifications for unethical behavior[29,33]. Furthermore, partners in crime can deflect blame on one another, which is even easier if one was not the one to initiate the unethical act (e.g., it is much easier to passively accept a bribe than to actively request one[76–78]).

Humans have long cooperated with machines[79–81]. As the machine partners become "smarter" and their behavior less predictable, research is shifting from mostly looking at the physical relationships between humans and machines towards understanding their socio-cognitive relationships[80,82,see for a review 83]. As a testimony of this trend, thanks to recent breakthroughs in machine learning, algorithms now can establish and sustain cooperation with humans across multiple strategic situations[57,84]. Hence, we may be concerned that they collude with them and break ethical rules for mutual benefits, just as machines may engage in algorithmic collusion among themselves[85–87]. Since there are few behavioral insights into unethical behavior in hybrid human-machine teams[88], much of this section is speculation.

First, we do not know the extent to which people might strategically deflect blame on their *machine partners in crime*. What we do know is that when people team up with machines, the machines can be seen as sharing the responsibility for negative outcomes[89], both by their human partners[90] and by third parties[91]. Having said that, humans still see themselves as primarily responsible for the outcomes when they cooperate with relatively simple machines[92,93]. Third-party observers similarly attribute less blame to AI agents compared to humans if a hybrid team violates moral norms[94]. These results suggest that people may be cognitively disposed to deflect at least some blame onto machines when they engage in joint unethical behavior with these machines.

Second, we do not know the extent to which people might frame joint unethical behavior with machines as mutually beneficial[75], since it is not clear whether people think of machines as experiencing some form of utility[95]. What we do know is that people show less mentalizing brain activity when cooperating with machines (compared to humans)[96], which suggests that they are de-emphasizing the 'mental states' of the machines[97], including its experienced utility. People

also experience less emotional and social responses when interacting with machines[83,98,99], which could be a double-edged sword: this muted response could make it harder to frame the unethical act positively[100] — as a mutually beneficial win-win situation — but it could also facilitate unethical behavior by weakening feelings of guilt[99].

Other factors may prove even more critical. For example, although some have recommended drawing on automated, and ideally incorruptible, whistleblowing machines[101], we do not know yet how much people will fear that the machine may denounce them or blow the whistle if they initiate or accept unethical cooperation. Given the prevalence of human-human corrupt collaboration and our sizable uncertainty about its human-machine version, future research needs to give it serious consideration.

Delegate

Besides active partners, others can also serve as delegates to whom people can outsource the execution of unethical behavior. When people face the choice between breaking ethical rules themselves versus letting others do so on their behalf, they generally prefer delegation[102]. Acting through others can entail explicit instruction to break ethical rules, such as when using henchpersons. Yet, more often than not, people do not explicitly instruct the delegates to break ethical rules but instead merely define their desired outcome and turn a blind eye to the modalities of achieving this goal. Thereby, the remitter avoids direct contact with the victims and can willfully ignore any possible ethical rule violations[102,103]. Moreover, if inflicted harm becomes apparent, blame and responsibility can be deflected to the delegate, which alleviates the guilt experienced.

People also delegate a growing number of tasks to AI agents[5,104,105], as diverse as setting

prices in online markets[86], interrogating suspects[106], or devise a sales strategy[107]. New forms of ethical risks emerge because the delegation of ethically questionable behavior to AI agents might be particularly attractive[108]: The often-incomprehensible workings of algorithms create ambiguity[109,110]. Letting such "black box" algorithms execute tasks on one's behalf increases plausible deniability[10,106], and obfuscates the attribution of responsibility for the harm caused[111]. On top of that, when entrusting machines to execute tasks that cause potential harm, victims generally remain psychologically distant and abstract[112].

One key consequence of these dynamics is that in many cases, people may cause harm without explicitly knowing so because they only specified a goal they wanted to achieve and left the execution to an algorithm[37] — for example, one may use algorithmic prices to sell goods on online markets, without being aware that algorithms might coordinate and set collusive prices[85]. Those employing AI interrogators might merely specify the desired result of a confession without explicitly preventing the AI agent from threatening torture[106]. Marketers drawing on AI-power sales strategies might blind themselves to the fact that the AI agent employs deceptive tactics to reach the sales goals.

But AI can also be of use for those who explicitly intend to do harm[10,113,114]. Recent developments in deep learning, particularly Generative Adversarial Networks (GANs), have massively facilitated the production of fake content that appears realistic[114]. Employing such AI hench-agents bears key advantages for those with malicious intent: AI can act autonomously[115] and has the power to strike with unprecedented effectiveness[116]. Furthermore, such AI hench-agents are typically scalable[117] and leave little to no breadcrumb trail back to the original initiator of the wrongdoings[10,118]. For example, AI-powered deepfakes allow forging identities[119], and thereby put phishing attacks on new, i.e., a more personalized level of spear phishing[113],

which boosts the effectiveness of the attacks[116].

Reflecting on this emerging worry, a panel of experts has nominated deepfakes as the most dangerous tool for AI-enabled crime[114]. Soon their use could exceed the scam and cyberwarfare contexts and become an attractive tool for ordinary citizens. Consider, for example (online) shop owners, who outsource the task of writing fake reviews to NLG algorithms, or political competitors, who use deepfakes to sully the reputation of their rivals[120].

Delegating tasks to AI agents rather than to humans combines most factors conducive for unethical behavior: anonymity[121], psychological distance from victims[122], and undetectability[112,123]. While people are hesitant to outsource tasks to static algorithms[105], recent studies show that delegating tasks to AI agents rather than a person reduces the remitters (negative) emotional reactions[124]. These studies suggest that letting algorithms do the "dirty job" of breaking ethical rules for profit on one's behalf likely reduces people's remorse and guilt. Thereby, reasons to worry exist that algorithmic delegation could contribute to well-intended people doing bad things, often without realizing it. Although not explicitly instructed to, AI delegates might neglect ethical rules when executing such tasks[37,125]. On top of that, AI agents become an increasingly attractive tool for those who have the intention to advance selfish goals, acting as a hench-agent on one's behalf[14]. Soon not only scammers but everyone from high school students, over business owners, to disgruntled ex-partners could be tempted to use AI to engage in such delegated forms of unethical behavior.

AI as an influencer versus enabler

Examining the fears about the corrupting force of AI reveals a key difference between cases when AI agents themselves are actively involved in the ethical behavior or not. When they

are not, such as when acting as a role model or advisor, AI agents become *influencers* that target people's moral preferences. In these roles, available evidence suggests that AI agents do not yet exceed humans in their ability to change what people consider right and wrong. However, when it comes to the scale of influence, such AI agents' abilities vastly exceed those of humans. That is, even though AI agents do not significantly surpass humans in their abilities to corrupt ethical behavior on a single occasion, their aggregate influence can be worrisome[117]. Consider the vast effect that AI has by enabling "personalized mass persuasion"[41]. AI recommender systems can slightly nudge consumers to purchase products that create harmful consequences for others[14]. Even if AI agents succeed at a low rate on a given occasion, overall, they might lead to a non-negligible shift towards more unethical behavior when employed widely. The subtle influence of AI agents might, in aggregate, have a substantial effect on human unethical behavior.

When AI agents are actively involved in unethical behavior — as partners and delegates — they become *enablers* that allow people to act based on their (selfish) preferences. AI agents offer the dangerous trifecta of opacity, anonymity, and social distance that enables people to psychologically dissociate themselves from the unethical act. That is, people often deceive themselves to achieve the dual goals of behaving self-interestedly, but at the same time retain the belief that their moral standards are upheld[126]. They frequently let moral concerns fade into the background and seek to obscure the moral implications of their behavior, a process that can occur without conscious awareness[127]. AI enablers amplify this trend, likely more than human enablers do, and thus potentially increase people's ethical blind spots[128], a trend that warrants concern and, more importantly, empirical scrutiny.

How to gain behavioral insight for better oversight?

A pressing demand exists for behavioral insights into how interactions between humans and AI agents might corrupt human ethical behavior[10]. Such research programs need to be grounded in both computer science and social science[129–131]. Studies using hypothetical scenarios ("what would you want the algorithm to do?") and self-reported data ("how do you rate the algorithm's decision?") have produced valuable insights into people's stated preferences[132–134]. However, little empirical knowledge exists on how dynamic human interactions with and through AI agents can cause unethical behavior. Experimental research that treats AI as autonomous agents — similar to how social science treats humans — can help draw causal inferences into the potentially corrupting effects of AI on human ethical behavior[4]. Adopting such a behavioral ethics approach to AI will provide a better understanding of its potential to promote ethical behavior, and help to design evidence-based policies that reduce the corrupting risks of AI[20,135].

First, we need more experiments that directly compare the magnitude of AI-induced corruption versus human-induced corruption. This article outlined several social roles that human and AI agents can play in corrupting human ethical behavior. We note that these roles are somewhat archetypical, that they may overlap, that they might not capture every form of influence (e.g., interactions with chatbots may disinhibit people to engage in harmful discourse [136,137]), and that the shift from one to the other may be a matter of degree. However, differentiating between these roles helps to identify their unique corrupting powers. Due to the lack of experimental work that directly pits human against AI agents in these different roles, assessing the fears about AI's corrupting force largely relies on extrapolating from research on humans corrupting humans. Previous research has compared the behavior of humans, who play economic games with humans, to the behavior of humans, who play economic games with AI

agents[57,84], but these tasks mostly lack a clear ethical component. The next step would be to conduct experiments in which humans face the temptation to behave unethically and can be pushed in that direction by AI agents acting as role-models, advisors, partners, or delegates — and to assess whether such AI agents can surpass the corruptive influence of other humans, by which magnitude, and in which role.

Running experiments on unethical behavior can raise practical and ethical challenges of its own. Many forms of unethical behavior, like corruption, are typically hidden from plain sight, rendering the search for valid proxies challenging[138]. Researchers who themselves introduce unethical behaviors in field experiments face warranted concerns from a research ethics perspective[139]. Overcoming these challenges requires adopting creative means to obtain behavioral data on unethical behavior from the field[23,140,for a review see 141] or running experiments using behavioral tasks of unethical behavior in the lab or online[1,32]. The estimates obtained in such controlled environments correlate with unethical behavior in the field, hinting at their external validity[142,143].

Even though unexpected behaviors by AI agents can emerge[8], their impact on humans' ethical behavior largely depends on the way they are programmed and trained[144]. To assess the corrupting effects of AI, future research needs to make difficult choices when it comes to programming the AI agents used in experiments. AI agents can be programmed to follow a specific objective function while humans often follow multiple goals, which are difficult to change or predict[145]. Hence, the results of AI agents in these experiments will largely depend on how the algorithms are programmed. If AI agents follow objective functions that merely maximize financial payoffs, there is little reason to believe that it would abstain from breaking unethical rules to achieve this goal. In fact, first simulations reveal that the same algorithm that

achieves human-like cooperation levels in strategic games[84] lies to the maximum extent when placed in a collaborative cheating task. To enable transparent and reproducible research, we will need an open and standardized protocol to use diversely calibrated algorithms as agents in experiments[146].

This methodological challenge echoes the broader technical challenge of how to avoid algorithmic harm. Many fears about AI corrupting humans could be assuaged if algorithms simply never caused harm[37]. For example, if we can make sure that algorithms never give unethical advice, then we need not fear that humans be corrupted by this advice. The large literature dealing with ethical AI and its alignment to human ethical value has made it clear, though, that identifying, specifying, and programming human values into machines is a thorny challenge[147,148]. One strategy proposes to train ML algorithms on desirable behavioral patterns, rather than blindly opting for the largest datasets available for training[144]. Such efforts provide an interesting point of departure to understand how people imitate or leverage machines into unethical behavior.

Understanding is not enough, though. The next necessary step is to conduct policy-oriented behavioral research[149], particularly with a "focus on (...) AI-related social, legal and ethical implications and policy issues" as the OECD recommends[150]. Anti-corruption research[20,151], AI safety research[109,152], and policy guidelines[150], alike point towards transparency as a key policy to reduce potential harm. In particular, we need to investigate whether mere knowledge about the existence of an algorithm, known as transparency about algorithmic presence[153], could alleviate its corrupting power. As algorithms become increasingly difficult to detect with the naked eye[56,119], researchers and policymakers have called for legal regulations that demand AI agents to disclose themselves as such at the beginning of interactions[154]. Such

knowledge about algorithmic presence likely shapes AI agents' corrupting influence across all the roles we considered in this article[56,57,72]. However, transparency can also create new tradeoffs, for example, by undermining efficiency[57]. In any case, we need to know more about the situations in which people actively seek out information about whether a fellow human or an AI executes a given role and the situations in which they intentionally avoid such information, since such strategic avoidance may nullify efforts toward transparency.

Another policy-relevant research question is how to integrate awareness for the corrupting force of AI tools into the innovation process. New AI tools hit the market on a daily basis. The current approach of "innovate first, ask for forgiveness later" has caused considerable backlash[155] and even demands for banning AI technology like facial recognition[156]. As a consequence, ethical considerations must enter the innovation and publication process of AI developments[157]. Current efforts to develop ethical labels for responsible AI[158] or crowdsourcing citizens' preferences about ethical AI[132,159] are mostly concerned about the direct unethical consequences of AI behavior and not its influence on the ethical conduct of the humans who interact with and through it. A thorough experimental approach to responsible AI will need to expand concerns about direct AI-induced harm to concerns about how machine behavior affects the unethical behavior of humans.

References

1. Abeler, J., Nosenzo, D. & Raymond, C. Preferences for truth‐telling. *Econometrica* **87**, 1115–1153 (2019).
2. Gächter, S. & Schulz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499 (2016).
3. Weisel, O. & Shalvi, S. The collaborative roots of corruption. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10651–10656 (2015).
4. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
5. de Melo, C. M., Marsella, S. & Gratch, J. Social decisions and fairness change when people's interests are represented by autonomous agents. *Auton. Agent. Multi. Agent. Syst.* **32**, 163–187 (2018).
6. Tegmark, M. *Life 3.0: Being human in the age of artificial intelligence*. (Knopf, 2017).
7. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* 78–87 (2012).
8. Floridi, L. & Sanders, J. W. On the morality of artificial agents. *Minds Mach.* **14**, 349–379 (2004).
9. Yang, G.-Z. *et al.* The grand challenges of Science Robotics. *Sci Robot* **3**, 1–14 (2018).
10. King, T. C., Aggarwal, N., Taddeo, M. & Floridi, L. Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Sci. Eng. Ethics* **26**, 89–120 (2020).
11. Floridi, L. Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 1–12 (2016).
12. Damiani, J. A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. *Forbes Magazine* (2019).
13. Robitzski, D. This Grad Student Used a Neural Network to Write His Papers. *Futurism* https://futurism.com/grad-student-neural-network-write-papers (2020).
14. Lin, T. C. W. The new market manipulation. *Emory LJ* **66**, 1253–1315 (2016).
15. Hakim, F. Z. M., Indrayani, L. M. & Amalia, R. M. A Dialogic Analysis of Compliment Strategies

Employed by Replika Chatbot. in *Third International Conference of Arts, Language and Culture (ICALC 2018)* (Atlantis Press, 2019).

16. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* **1**, 74–78 (2019).
17. EU Commission. *White Paper on Artificial Intelligence—A European Approach to Excellence and Trust*. (2020).
18. Plant, S. *Zeros and ones: Digital women and the new technoculture*. vol. 4 (London, 1997).
19. Frank, M., Roehrig, P. & Pring, B. *What To Do When Machines Do Everything: How to Get Ahead in a World of AI, Algorithms, Bots, and Big Data*. (John Wiley & Sons, 2017).
20. Mungiu-Pippidi, A. The time has come for evidence-based anticorruption. *Nature Human Behaviour* **1**, 1–3 (2017).
21. Gino, F. Understanding ordinary unethical behavior: why people who value morality act immorally. *Current Opinion in Behavioral Sciences* **3**, 107–111 (2015).
22. Jones, T. M. Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *The Academy of Management Review* **16**, 366–395 (1991).
23. Cohn, A., Maréchal, M. A., Tannenbaum, D. & Zünd, C. L. Civic honesty around the globe. *Science* **365**, 70–73 (2019).
24. Treviño, L. K., Weaver, G. R. & Reynolds, S. J. Behavioral Ethics in Organizations: A Review. *J. Manage.* **32**, 951–990 (2006).
25. Bazerman, M. H. & Gino, F. Behavioral Ethics: Toward a Deeper Understanding of Moral Judgment and Dishonesty. *Annual Review of Law and Social Science* **8**, 85–104 (2012).
26. Shalvi, S., Weisel, O., Kochavi-Gamlie, S. & Leib, M. Corrupt collaboration: a behavioral ethics approach. in *Cheating, corruption, and concealment: The roots of dishonesty* (eds. Van Prooijen, J. W. & Van Lange, P. A. M.) 134–148 (Cambridge University Press, 2016).
27. Mazar, N., Amir, O. & Ariely, D. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *J. Mark. Res.* **45**, 633–644 (2008).

28. Ariely, D. *The Honest Truth About Dishonesty: How We Lie to Everyone---Especially Ourselves*. (HarperCollins, 2012).

29. Shalvi, S., Gino, F., Barkan, R. & Ayal, S. Self-Serving Justifications: Doing Wrong and Feeling Moral. *Curr. Dir. Psychol. Sci.* **24**, 125–130 (2015).

30. Cohn, A., Fehr, E. & Maréchal, M. A. Business culture and dishonesty in the banking industry. *Nature* **516**, 86–89 (2014).

31. Rahwan, Z., Yoeli, E. & Fasolo, B. Heterogeneity in banker culture and its influence on dishonesty. *Nature* **575**, 345–349 (2019).

32. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: A meta-analysis on dishonest behavior. *Psychol. Bull.* **145**, 1–44 (2019).

33. Köbis, N. C., van Prooijen, J.-W., Righetti, F. & Van Lange, P. A. M. Prospection in Individual and Interpersonal Corruption Dilemmas. *Rev. Gen. Psychol.* **20**, 71–85 (2016).

34. Gross, J., Leib, M., Offerman, T. & Shalvi, S. Ethical Free Riding: When Honest People Find Dishonest Partners. *Psychol. Sci.* **29**, 1956–1968 (2018).

35. Gross, J. & De Dreu, C. K. W. Rule Following Mitigates Collaborative Cheating and Facilitates the Spreading of Honesty Within Groups. *Pers. Soc. Psychol. Bull.* **online first**, 0146167220927195 (2020).

36. Leib, M., Köbis, N. C., Soraperra, I., Weisel, O. & Shalvi, S. Collaborative Dishonesty: A Meta-Study. *CREED Working Paper Series* (2021).

37. Thomas, P. S. *et al.* Preventing undesirable behavior of intelligent machines. *Science* **366**, 999–1004 (2019).

38. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).

39. Koenecke, A. *et al.* Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7684–7689 (2020).

40. He, Q., Turel, O. & Bechara, A. Brain anatomy alterations associated with Social Networking Site

(SNS) addiction. *Sci. Rep.* **7**, 45064 (2017).

41. Aral, S. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health--and How We Must Adapt*. (Crown, 2020).
42. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
43. Soraperra, I. *et al.* The bad consequences of teamwork. *Economics Letters* **160**, 12–15 (2017).
44. Cialdini, R. B., Reno, R. R. & Kallgren, C. A. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *J. Pers. Soc. Psychol.* **58**, 1015–1026.
45. Bicchieri, C. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. (Oxford University Press, 2016).
46. Efferson, C., Vogt, S. & Fehr, E. The promise and the peril of using social influence to reverse harmful traditions. *Nat Hum Behav* **4**, 55–68 (2020).
47. Köbis, N. C., Troost, M., Brandt, C. O. & Soraperra, I. Social norms of corruption in the field: social nudges on posters can help to reduce bribery. *Behavioural Public Policy* **online first**, 1–28.
48. Köbis, N. C., van Prooijen, J.-W., Righetti, F. & Van Lange, P. A. M. 'Who Doesn't?'—The Impact of Descriptive Norms on Corruption. *PLOS One* **10**, e0131830 (2015).
49. Köbis, N. C., Jackson, D. & Carter, D. I. Recent approaches to the study of social norms and corruption. in *A Research Agenda for Studies of Corruption* (eds. Mungiu-Pippidi, A. & Heywood, P.) 41–53 (Edward Elgar Publishing, 2020).
50. Brandstetter, J. *et al.* A peer pressure experiment: Recreation of the Asch conformity experiment with robots. *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2014) doi:10.1109/iros.2014.6942730.
51. Shiomi, M. & Hagita, N. Do Synchronized Multiple Robots Exert Peer Pressure? in *Proceedings of the Fourth International Conference on Human Agent Interaction* 27–33 (Association for Computing Machinery, 2016).
52. Vollmer, A.-L., Read, R., Trippas, D. & Belpaeme, T. Children conform, adults resist: A robot group

induced peer pressure on normative social conformity. *Science Robotics* **3**, eaat7111 (2018).

53. Salomons, N., van der Linden, M., Strohkorb Sebo, S. & Scassellati, B. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* 187–195 (Association for Computing Machinery, 2018).
54. Hertz, N. & Wiese, E. Under Pressure: Examining Social Conformity With Computer and Robot Groups. *Hum. Factors* **60**, 1207–1218 (2018).
55. Hertz, N., Shaw, T., de Visser, E. J. & Wiese, E. Mixing It Up: How Mixed Groups of Humans and Machines Modulate Conformity. *Journal of Cognitive Engineering and Decision Making* **13**, 242–257 (2019).
56. Köbis, N. & Mossink, L. Artificial Intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior* **114**, 106553 (2021).
57. Ishowo-Oloko, F. *et al.* Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* **1**, 517–521 (2019).
58. Song-Nichols, K. & Young, A. G. Gendered Robots Can Change Children's Gender Stereotyping. in *CogSci 2020* 2480–2485 (Cognitive Science Society, 2020).
59. Williams, R., Machado, C. V., Druga, S., Breazeal, C. & Maes, P. 'My doll says it's ok': a study of children's conformity to a talking doll. in *Proceedings of the 17th ACM Conference on Interaction Design and Children* 625–631 (Association for Computing Machinery, 2018).
60. Milgram, S. Behavioral Study of Obedience. *J. Abnorm. Psychol.* **67**, 371–378 (1963).
61. Burger, J. M. Replicating Milgram: Would people still obey today? *Am. Psychol.* **64**, 1–11 (2009).
62. Gino, F., Moore, D. A. & Bazerman, M. H. No harm, no foul: The outcome bias in ethical judgments. *Harvard Business School NOM Working Paper* (2009).
63. Wiltermuth, S. S., Newman, D. T. & Raj, M. The consequences of dishonesty. *Current Opinion in Psychology* **6**, 20–24 (2015).
64. Fogg, B. J. Creating persuasive technologies: an eight-step design process. in *Proceedings of the 4th*

International Conference on Persuasive Technology 1–6 (Association for Computing Machinery, 2009).

65. Longoni, C. & Cian, L. Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The 'Word-of-Machine' Effect. *Journal of Marketing* **online first**, (2020).
66. AI Reads Human Emotions. Should it? *MIT Technology Review* (2020).
67. How close is AI to decoding our emotions? *MIT Technology Review* (2020).
68. Giubilini, A. & Savulescu, J. The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence. *Philos. Technol.* **31**, 169–188 (2018).
69. Hoc, J.-M. & Lemoine, M.-P. Cognitive Evaluation of Human-Human and Human-Machine Cooperation Modes in Air Traffic Control. *Int. J. Aviat. Psychol.* **8**, 1–32 (1998).
70. Castelo, N., Bos, M. W. & Lehmann, D. R. Task-Dependent Algorithm Aversion. *J. Mark. Res.* **56**, 809–825 (2019).
71. Dietvorst, B. J., Simmons, J. & Massey, C. Understanding Algorithm Aversion: Forecasters Erroneously Avoid Algorithms After Seeing them Err. *Proc. AMIA Annu. Fall Symp.* **2014**, 12227 (2014).
72. Leib, M., Köbis, N.C., Hagens, M., Rilke, R., & Irlenbusch, B. The corruptive force of AI-generated advice. *CREED Working Paper Series* (2021).
73. Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied* **70**, 1–70 (1956).
74. Larsen, K. S. The Asch conformity experiment: Replication and transhistorical comparison. *J. Soc. Behav. Pers.* **5**, 163–168 (1990).
75. Wiltermuth, S. S. Cheating more when the spoils are split. *Organ. Behav. Hum. Decis. Process.* **115**, 157–168 (2011).
76. Ryvkin, D. & Serra, D. Corruption and competition among bureaucrats: An experimental study. *J. Econ. Behav. Organ.* **175**, 439–451 (2018).
77. Köbis, N. C., van Prooijen, J.-W., Righetti, F. & Van Lange, P. A. M. The Road to Bribery and

Corruption: Slippery Slope or Steep Cliff? *Psychol. Sci.* **28**, 297–306 (2017).

78. Lambsdorff, J. G. & Frank, B. Corrupt reciprocity--Experimental evidence on a men's game. *Int. Rev. Law Econ.* **31**, 116–125 (2011).
79. Schmidt, K. Cooperative work: A conceptual framework. in *Distributed decision making: Cognitive models for cooperative work* (eds. Rasmussen, J., Brehmer, B. & Leplat, J.) 75–110 (John Willey and Sons, 1991).
80. Hoc, J.-M. Towards a cognitive approach to human–machine cooperation in dynamic situations. *Int. J. Hum. Comput. Stud.* **54**, 509–540 (2001).
81. Flemisch, F. *et al.* Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cogn. Technol. Work* **14**, 3–18 (2012).
82. Suchman, L., Blomberg, J., Orr, J. E. & Trigg, R. Reconstructing Technologies as Social Practice. *Am. Behav. Sci.* **43**, 392–408 (1999).
83. Chugunova, M. & Sele, D. We and It: An Interdisciplinary Review of the Experimental Evidence on Human-Machine Interaction. (2020) doi:10.2139/ssrn.3692293.
84. Crandall, J. W. *et al.* Cooperating with machines. *Nature Communications* **9**, (2018).
85. Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. Artificial Intelligence, Algorithmic Pricing and Collusion. *American Economic Review* **110**, 3267–3297 (2019).
86. Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., Jr & Pastorello, S. Protecting consumers from collusive prices due to AI. *Science* **370**, 1040–1042 (2020).
87. Martinez-Miranda, E., McBurney, P. & Howard, M. J. W. Learning unfair trading: A market manipulation analysis from the reinforcement learning perspective. in *2016 IEEE Conference on Evolving and Adaptive Intelligent Systems* 103–109 (EAIS, 2016).
88. Mell, J., Lucas, G. & Gratch, J. Prestige Questions, Online Agents, and Gender-Driven Differences in Disclosure. in *Intelligent Virtual Agents* 273–282 (Springer International Publishing, 2017).
89. Hohenstein, J. & Jung, M. AI as a moral crumple zone: The effects of AI-mediated communication

on attribution and trust. *Comput. Human Behav.* **106**, 106190 (2020).

90. Kirchkamp, O. & Strobel, C. Sharing responsibility with a machine. *Journal of Behavioral and Experimental Economics* **80**, 25–33 (2019).
91. Pezzo, M. V. & Pezzo, S. P. Physician Evaluation after Medical Errors: Does Having a Computer Decision Aid Help or Hurt in Hindsight? *Medical Decision Making* **26**, 48–56 (2006).
92. Paravisini, D. & Schoar, A. The Incentive Effect of Scores: Randomized Evidence from Credit Committees. *National Bureau of Economic Research (NBER) Working Paper Series* (2013).
93. Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F. & Shah, J. A. Decision-making authority, team efficiency and human worker satisfaction in mixed human--robot teams. *Auton. Robots* **39**, 293–312 (2015).
94. Shank, D. B., DeSanti, A. & Maninger, T. When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Inf. Commun. Soc.* **22**, 648–663 (2019).
95. Houser, D. & Kurzban, R. Revisiting Kindness and Confusion in Public Goods Experiments. *American Economic Review* **92**, 1062–1069 (2002).
96. Coricelli, G. & Nagel, R. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9163–9168 (2009).
97. Frith, C. D. & Frith, U. The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
98. Schniter, E., Shields, T. W. & Sznycer, D. Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology* **78**, 102253 (2020).
99. De Melo, C., Marsella, S. & Gratch, J. People Do Not Feel Guilty About Exploiting Machines. *ACM Trans. Comput.-Hum. Interact.* **23**, 1–17 (2016).
100. Mazar, N. & Ariely, D. Dishonesty in Everyday Life and Its Policy Implications. *Journal of Public Policy & Marketing* **25**, 117–126 (2006).
101. Waytz, A. Why Robots Could Be Awesome Whistleblowers. *The Atlantic* (2014).
102. Drugov, M., Hamman, J. & Serra, D. Intermediaries in corruption: an experiment. *Exp. Econ.* **17**,

78–99 (2014).

103. Van Zant, A. B. & Kray, L. J. 'I can't lie to your face': Minimal face-to-face interaction promotes honesty. *J. Exp. Soc. Psychol.* **55**, 234–238 (2014).

104. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data & Society* **3**, 2053951716679679 (2016).

105. Gogoll, J. & Uhl, M. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics* **74**, 97–103 (2018).

106. McAllister, A. Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture. *Minn. Law Rev.* **101**, 2527 (2016).

107. Start having incredible sales conversations through science. *Gong* https://www.gong.io/.

108. Mell, J., Lucas, G., Mozgai, S. & Gratch, J. The Effects of Experience on Deception in Human-Agent Negotiation. *J. Artif. Intell. Res.* **68**, 633–660 (2020).

109. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–36 (2019).

110. Gunning, D., Stefik, M., Choi, J. & Miller, T. XAI—Explainable artificial intelligence. *Science Robotics* **4**, eaay7120 (2019).

111. Dana, J., Weber, R. A. & Kuang, J. X. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econom. Theory* **33**, 67–80 (2007).

112. Hancock, J. T. & Guillory, J. Deception with technology. in *The Handbook of the Psychology of Communication Technology* (ed. Sundar, S. S.) 270–289 (Wiley Online Library, 2015).

113. Seymour, J. & Tully, P. Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter. *Black Hat USA* **37**, 1–39 (2016).

114. Caldwell, M., Andrews, J. T. A., Tanay, T. & Griffin, L. D. AI-enabled future crime. *Crime Science* **9**, 14 (2020).

115. Sharkey, N., Goodman, M. & Ross, N. The coming robot crime wave. *Computer* **43**, 115–116

(2010).

116. Jagatic, T. N., Johnson, N. A., Jakobsson, M. & Menczer, F. Social phishing. *Commun. ACM* **50**, 94–100 (2007).

117. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).

118. Brundage, M. *et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. (2018).

119. Bendel, O. The synthetization of human voices. *AI Soc.* **34**, 83–89 (2019).

120. McKelvey, F. & Dubois, E. Computational propaganda in Canada: The use of political bots. *Computational Propaganda Research Project* (2017).

121. Ostermaier, A. & Uhl, M. Spot on for liars! How public scrutiny influences ethical behavior. *PLoS One* **12**, e0181682 (2017).

122. Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D. & Shalvi, S. Intuitive Honesty Versus Dishonesty: Meta-Analytic Evidence. *Perspect. Psychol. Sci.* **14**, 778–796 (2019).

123. Rauhut, H. Beliefs about lying and spreading of dishonesty: undetected lies and their constructive and destructive social dynamics in dice experiments. *PLoS One* **8**, e77878 (2013).

124. Leyer, M. & Schneider, S. Me, You or Ai? How Do We Feel About Delegation. in *Proceedings of the 27th European Conference on Information Systems (ECIS)* (2019).

125. Wellman, M. P. & Rajan, U. Ethical Issues for Autonomous Trading Agents. *Minds Mach.* **27**, 609–624 (2017).

126. Tenbrunsel, A. E. & Messick, D. M. Ethical fading: The role of self-deception in unethical behavior. *Soc. Justice Res.* **17**, 223–236 (2004).

127. Bazerman, M. H. & Banaji, M. R. The social psychology of ordinary ethical failures. *Soc. Justice Res.* **17**, 111–115 (2004).

128. Bazerman, M. H. & Tenbrunsel, A. E. *Blind Spots: Why We Fail to Do What's Right and What to Do about It*. (Princeton University Press, 2012).

129. Sloane, M. & Moss, E. AI's social sciences deficit. *Nature Machine Intelligence* **1**, 330–331 (2019).

130. Irving, G. & Askell, A. AI safety needs social scientists. *Distill* **4**, e14 (2019).

131. Crawford, K. & Calo, R. There is a blind spot in AI research. *Nature* **538**, 311–313 (2016).

132. Awad, E. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018).

133. Bigman, Y. E., Waytz, A., Alterovitz, R. & Gray, K. Holding Robots Responsible: The Elements of Machine Morality. *Trends Cogn. Sci.* **23**, 365–368 (2019).

134. Burton, J. W., Stein, M. & Jensen, T. B. A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **33**, 220–239 (2020).

135. Fisman, R. & Golden, M. How to fight corruption. *Science* **356**, 803–804 (2017).

136. De Angeli, A. Ethical implications of verbal disinhibition with conversational agents. *PsychNology Journal* **7**, (2009).

137. McDonnell, M. & Baxter, D. Chatbots and Gender Stereotyping. *Interact. Comput.* **31**, 116–121 (2019).

138. How to research corruption. in *Conference Proceedings Interdisciplinary Corruption Research Forum June* (eds. Schwickerath, A. K., Varraich, A. & Lee Smith, L.) 7–8 (Interdisciplinary Corruption Research Network, 2016).

139. Salganik, M. J. *Bit by bit*. (Princeton University Press, 2017).

140. Fisman, R. & Miguel, E. Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets. *J. Polit. Econ.* **115**, 1020–1048 (2007).

141. Pierce, L. & Balasubramanian, P. Behavioral field evidence on psychological and social factors in dishonesty and misconduct. *Current Opinion in Psychology* **6**, 70–76 (2015).

142. Dai, Z., Galeotti, F. & Villeval, M. C. Cheating in the Lab Predicts Fraud in the Field: An Experiment in Public Transportation. *Manage. Sci.* **64**, 1081–1100 (2018).

143. Cohn, A. & Maréchal, M. A. Laboratory Measure of Cheating Predicts School Misconduct. *Econ J* **128**, 2743–2754 (2018).

144. Hagendorff, T. Ethical behavior in humans and machines -- Evaluating training data quality for

beneficial machine learning. *arXiv [cs.CY]* (2020).

145. Mullainathan, S. Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times* (2019).
146. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
147. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2*. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (2017).
148. Russell, S., Dewey, D. & Tegmark, M. Research priorities for robust and beneficial artificial intelligence. *AI Magazine* **36**, 105–114 (2015).
149. Amir, O. *et al.* Psychology, behavioral economics, and public policy. *Mark. Lett.* **16**, 443–454 (2005).
150. OECD. *Recommendation of the Council on Artificial Intelligence (OECD)*. http://dx.doi.org/10.1017/ilm.2020.5 (2021) doi:10.1017/ilm.2020.5.
151. Fisman, R. & Golden, M. A. *Corruption: What Everyone Needs to Know*. (Oxford University Press, 2017).
152. Shin, D. & Park, Y. J. Role of fairness, accountability, and transparency in algorithmic affordance. *Comput. Human Behav.* **98**, 277–284 (2019).
153. Diakopoulos, N. Accountability in algorithmic decision making. *Commun. ACM* **59**, 56–62 (2016).
154. Walsh, T. Turing's red flag. *Commun. ACM* **59**, 34–37 (2016).
155. Webb, A. *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*. (Hachette UK, 2019).
156. Crawford, K. Halt the use of facial-recognition technology until it is regulated. *Nature* **572**, 565 (2019).
157. Hagendorff, T. Forbidden knowledge in machine learning reflections on the limits of research and publication. *AI Soc.* online (2020).

158. Finkel, A. *What will it take for us to trust AI?* https://www.weforum.org/agenda/2018/05/alan-finkel-turing-certificate-ai-trust-robot (2018).

159. Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A. & Rahwan, I. Crowdsourcing moral machines. *Commun. ACM* **63**, 48–55 (2020).