nature human behaviour

REVIEW ARTICLE

https://doi.org/10.1038/s41562-021-01128-2

Bad machines corrupt good morals

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

As machines powered by artificial intelligence (AI) influence humans' behaviour in ways that are both like and unlike the ways humans influence each other, worry emerges about the corrupting power of AI agents. To estimate the empirical validity of these fears, we review the available evidence from behavioural science, human–computer interaction and AI research. We propose four main social roles through which both humans and machines can influence ethical behaviour. These are: role model, advisor, partner and delegate. When AI agents become influencers (role models or advisors), their corrupting power may not exceed the corrupting power of humans (yet). However, AI agents acting as enablers of unethical behaviour (partners or delegates) have many characteristics that may let people reap unethical benefits while feeling good about themselves, a potentially perilous interaction. On the basis of these insights, we outline a research agenda to gain behavioural insights for better AI oversight.

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 AI agents (see Box 1 for our use of this term) increasingly act in advisory, delegatory or cooperative roles[4–8], should we fear that AI may exert a corrupting force on human ethical behaviour?

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 more than US$220,000 (ref. [9]). AI can also tempt honest citizens into unethical behaviour by merely making cheating easier. For example, students have successfully used powerful natural language generation (NLG) algorithms to craft their essays[10]. Finally, even if AI does not directly offer the means to cheat, it may still give inappropriate advice or provide an example of inappropriate behaviour. Consider how traders might imitate manipulative market strategies from algorithmic traders[11], or that by now, many adolescents seek guidance on ethical dilemmas from their personal AI assistants or chatbot friends[12]. With more than 100 million people using AI-powered personal assistants such as 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[13]. For example, a recent report by the European Commission highlights that "citizens (…) worry that AI can have unintended effects or even be used for malicious purposes"[14]. Yet, such pessimistic views about new technologies are nothing new[15]. People have felt threatened by machines for centuries[16], and tend to meet innovations with exaggerated scepticism[17]. Developing a cool-headed policy agenda requires an evidence-based assessment[18] about which of the fears that AI will corrupt human ethical behaviour are warranted[18]. Put differently, developing effective AI oversight requires an overview of available empirical insights.

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

How can people and AI agents corrupt ethical behaviour?

Unethical behaviour is commonly defined as "acts that have harmful effects on others and are either illegal or morally unacceptable to the larger community"[19], on the basis of ref. [20]. Behavioural 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 behaviour[21–24], either in a material sense (for example, financial gains and legal punishment) or a psychological sense (for example, self-image)[25–29]. Meta-analyses of individual forms of unethical behaviour (situations in which people face temptations by themselves) indicate that people generally break ethical rules only to the extent that they can justify it[1,30]. The behavioural research we will focus on is concerned with the power of social forces shaping (un)ethical behaviour[3,31–33] (for a meta-analysis, see ref. [34])—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[35] (for example, by reproducing biases[36,37], fostering Internet addiction[38,39] or accelerating the spread of false information[40]), 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).

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[41]. 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 behaviour is deemed to be widespread[42–44]. Experimental research reveals that such normative perceptions in general and perceived descriptive norms in particular strongly influence unethical behaviour as people often imitate others. Put differently, when perceiving that others break versus adhere to ethical rules, people often follow suit[2,45,46] (for a review, see ref. [47]).

[1]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany. [2]Toulouse School of Economics (TSM-R, CNRS), University of Toulouse Capitole, Toulouse, France. ✉e-mail: n.c.kobis@gmail.com

Box 1 | What do we not mean by 'AI agents'?

AI encompasses various techniques in computer science (for example, machine learning) that allow for the autonomous execution of tasks that used to be reserved for humans[6,7]. As a result 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 this Review. 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 this Review is that humans (for example, programmers, designers and users) are always ultimately responsible for the behaviour of AI agents and its consequences[11].

In the digital world, people are exposed to both human and machine behaviour[4]. A machine that would display unethical or inappropriate behaviour may therefore shift people's perception of what is acceptable or appropriate. There is mixed evidence (and negative on balance) that adult humans might conform to machines the same way they conform to humans, although this evidence is restricted to non-moral behaviours[48–53].

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[54,55] (for example, when online traders imitate manipulative trading strategies that, unbeknownst to them, are executed by algorithmic traders[11]). There is concerning evidence that children, more than adults, may be influenced by machine role models[50], in a way that makes them change their perception of moral transgressions[56,57]. 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. Behavioural research has established that people do tend to follow advice and orders, particularly when they come from authority figures[58] (see ref. [59] for a replication). 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 behaviour themselves[60,61].

Many AI agents pursue persuasive goals[39,62], such as giving advice and recommendations[63]. This trend of AI agents swaying people's behaviour is only increasing. Anecdotally, Amazon's chief scientist, Rohit Prasad, remarked that people's relationship with their Alexas "keeps growing from more of an assistant to advisor"[64]. In parallel to home assistants, millions of users engage with advice-giving conversational agents such as Replika (https://www.replika.ai/), trained on large amounts of data reflecting personalized preferences[65]. Companies such as Gong (https://www.gong.io/) use natural language processing (NLP) and machine learning to analyse big data of recorded sales conversations to provide advice to salespeople about how to improve their performance. Given the difficulty of training AI advisors to be impartial moral guides[35,66]—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[67].

Are these fears warranted? Even if machines were to give unethical advice, a phenomenon that has yet to be documented, we know that people state that they are not necessarily keen on following algorithmic recommendations in non-technical domains[68,69]. While this aversion could, in theory, dampen the effect of unethical machine advice, recent evidence from a large-scale experiment tells a different story[70]. This experiment directly compared the effect of human and AI advice on people's actual (un)ethical behaviour—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. Other studies have further shown that people might overtrust robots in emergency situations[71]. These initial findings suggest that we should take seriously the possibility that humans may act on the basis of 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[3,32]. This happens when two or more individuals act together towards a mutually beneficial outcome, realize that this outcome can be achieved through unethical means, and collaborate in these unethical means[24,31]. Behavioural research shows that people are more likely to act unethically in these collaborative conditions than when they face temptations alone[3,32]. Besides people having a general tendency to conform to others[72] (see ref. [73] for a replication), 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[41,74]. This skewed balance facilitates justifications for unethical behaviour[27,31]. 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 (for example, it is much easier to passively accept a bribe than to actively request one[75–77]).

Humans have long cooperated with machines[78–80]. As the machine partners become 'smarter' and their behaviour less predictable, research is shifting from mostly looking at the physical relationships between humans and machines towards understanding their socio-cognitive relationships[79,81] (see ref. [82] for a review). 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[55,83]. Hence, we may be concerned that people cooperate collusively with machines and thereby break ethical rules, similar to algorithmic collusion among machines[84–86]. As there are few behavioural insights into unethical behaviour in hybrid human–machine teams[87], much of this proposal 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 use machines, the machines can be seen as sharing the responsibility for negative outcomes[88], both by their human partners[89] and by third parties[90]. Having said that, humans still see themselves as primarily responsible for the outcomes when they cooperate with relatively simple machines[91,92]. Third-party observers similarly attribute less blame to AI agents compared to humans if a hybrid team violates moral norms[93]. These results suggest that people may be cognitively disposed to deflect at least some blame onto machines when they engage in joint unethical behaviour with these machines.

Second, we do not know the extent to which people might frame joint unethical behaviour with machines as mutually beneficial, as it is not clear whether people think of machines as experiencing some form of utility[94]. What we do know is that people show less mentalizing brain activity when cooperating with machines (compared to humans)[95], which suggests that they are de-emphasizing the 'mental states' of the machines[96], including their experienced utility. People also experience less emotional and social responses when interacting with machines[82,97,98], which could be a double-edged sword: this muted response could make it harder to frame the unethical

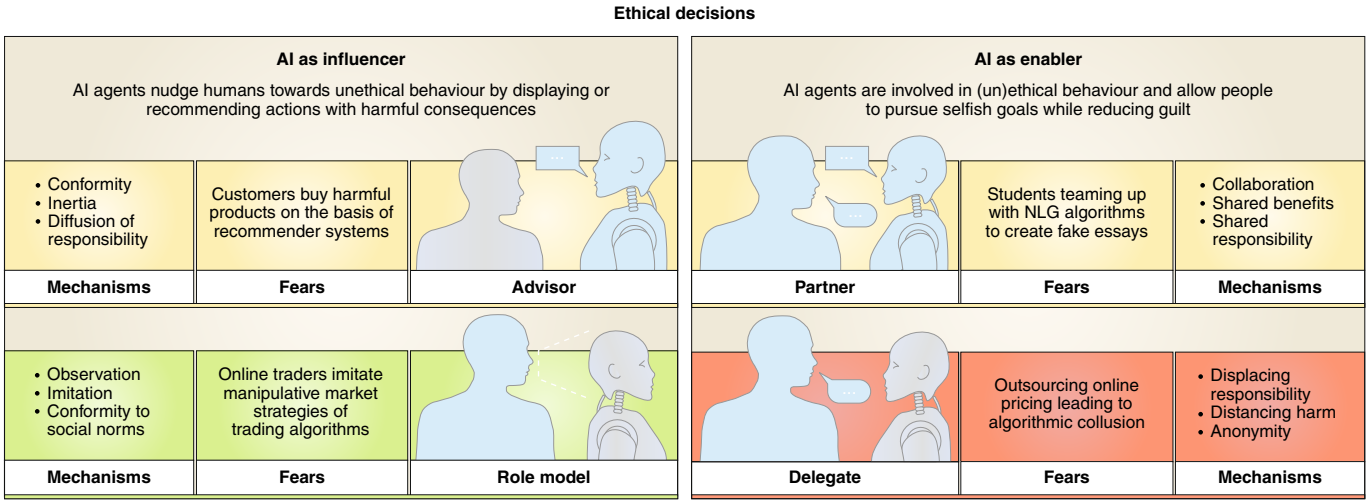

Fig. 1 | Four main roles in which AI agents and humans influence ethical behavior. An illustration of the main roles through which intelligent actors, whether human or AI, can corrupt ethical behaviour, grouped along the left panel for AI in the role of an influencer (role model and advisor) and along the right panel for AI being an enabler (partner and delegate). The main fears and mechanisms attached to each role are summarized. The colour coding indicates the strength of the corrupting force of AI: not reaching human levels yet (green), reaching but not surpassing human levels (yellow), and surpassing human levels (red).

act positively[99]—as a mutually beneficial win–win situation—but it could also facilitate unethical behaviour by weakening feelings of guilt[98].

Other factors may prove even more critical. For example, although AI systems have the potential to curb corruption[100], such as automated whistleblowing procedures, we do not know yet how much people will fear denunciation or whistleblowing when such systems are present. 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 behaviour. When people face the choice between breaking ethical rules themselves versus letting others do so on their behalf, they generally prefer delegation[101]. Acting through others can entail explicit instruction to break ethical rules, such as when using henchpeople. 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[101,102]. 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[5,103,104], as diverse as setting prices in online markets[85], interrogating suspects[105] or devising a sales strategy (https://www.gong.io/). New forms of ethical risks emerge because the delegation of ethically questionable behaviour to AI agents might be particularly attractive[106]: the often-incomprehensible workings of algorithms create ambiguity[107,108]. Letting such 'black box' algorithms execute tasks on one's behalf increases plausible deniability[105,109], and such 'moral wiggle room' obfuscates the attribution of responsibility for the harm caused[110]. On top of that, when entrusting machines to execute tasks that cause potential harm, victims generally remain psychologically distant and abstract[111].

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[35]—for example, one may use algorithmic prices to sell goods on online markets, without being aware that algorithms might coordinate and set collusive prices[84]. Those employing AI interrogators might merely specify the desired result of a confession without realizing the system has been programmed to also threaten torture[105]. Marketers drawing on AI-powered sales strategies might blind themselves to the fact that the AI agent employs deceptive tactics to reach the sales goals.

However, AI can also be of use for those who explicitly intend to do harm[109,112,113]. Recent developments in deep learning, particularly generative adversarial networks (GAN), have massively facilitated the production of fake content that appears realistic[113]. Employing such AI hench-agents bears key advantages for those with malicious intent: AI can act autonomously[114] and has the power to strike with unprecedented effectiveness[115]. Furthermore, such AI hench-agents are typically scalable[116] and leave little to no breadcrumb trail back to the original initiator of the wrongdoings[109,117]. For example, AI-powered deepfakes allow forging identities[118], and thereby put phishing attacks on a new—more personalized—level of spear phishing[112], which boosts the effectiveness of the attacks[115].

Reflecting on this emerging worry, a panel of experts has nominated deepfakes as the most dangerous tool for AI-enabled crime[113]. 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[119].

Delegating tasks to AI agents rather than to humans combines most factors conducive for unethical behaviour: anonymity[120], psychological distance from victims[121] and undetectability[111,122]. While people are hesitant to outsource tasks to static algorithms[104], recent studies show that delegating tasks to AI agents rather than a person reduces the remitters' (negative) emotional reactions[123]. These studies suggest that letting algorithms do the 'dirty job' of breaking ethical rules for profit on one's behalf probably reduces people's remorse and guilt. Thereby, there are reasons to worry 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[35,124]. 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[11]. Soon, not only scammers but everyone from high school students, to business owners, to disgruntled ex-partners could be tempted to use AI agents to engage in such delegated forms of unethical behaviour.

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 behaviour 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 surpass humans in their abilities to corrupt ethical behaviour on a single occasion, their aggregate influence can be worrisome[116]. Consider the vast effect that AI has by enabling 'personalized mass persuasion'[39]. AI recommender systems can slightly nudge consumers to purchase products that create harmful consequences for others[11]. 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 behaviour when employed widely. The subtle influence of AI agents might, in aggregate, have a substantial effect on human unethical behaviour.

When AI agents are actively involved in unethical behaviour—as partners and delegates—they become enablers that allow people to act on the basis of 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[125]. They frequently let moral concerns fade into the background and seek to obscure the moral implications of their behaviour, a process that can occur without conscious awareness[126]. AI enablers amplify this trend, probably more than human enablers do, and thus potentially increase people's ethical blind spots[127], a trend that warrants concern and, more importantly, empirical scrutiny.

Empirical insights to improve oversight

A pressing demand exists for behavioural insights into how interactions between humans and AI agents might corrupt human ethical behaviour[109]. Such research programmes need to be grounded in both computer science and social science[128–130]. 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[131–133]. However, little empirical knowledge exists on how dynamic human interactions with and through AI agents can cause unethical behaviour. Adopting such a behavioural ethics approach to AI will provide a better understanding of its potential to promote ethical behaviour and help to design evidence-based policies that reduce the corrupting risks of AI[134].

As part of the new research agenda, we need more experiments that directly compare the magnitude of AI-induced corruption versus human-induced corruption. This Review outlined several social roles that human and AI agents can play in corrupting human ethical behaviour. We note that these roles are archetypical, that they may overlap, that they might not capture every form of influence (for example, interactions with chatbots may disinhibit people to engage in harmful discourse[135,136]), 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. Previous research has compared the behaviour of humans who play economic games with humans to the behaviour of humans who play economic games with AI agents[55,83]. However, 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 what magnitude, and in which role.

Running experiments on unethical behaviour can raise practical and ethical challenges of its own. Many forms of unethical behaviour, such as corruption, are typically hidden from plain sight, rendering the search for valid proxies challenging[137]. Researchers who themselves introduce unethical behaviours in field experiments face warranted concerns from a research ethics perspective[138]. Overcoming these challenges requires adopting creative means to obtain behavioural data on unethical behaviour from the field[21,139] (see ref. [140] for a review) or running experiments using behavioural tasks of unethical behaviour in the laboratory or online[1,30]. The estimates obtained in such controlled environments correlate with unethical behaviour in the field, hinting at their external validity[141,142].

Even though unexpected behaviours by AI agents can emerge[143], their impact on humans' ethical behaviour 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. Suppose AI agents are programmed to follow objective functions that merely maximize financial payoffs. In that case, there is little reason to believe that they 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[83] 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[35]. 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. A rich body of 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 machine learning algorithms on desirable behavioural 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 behaviour.

Understanding is not enough, though. The next necessary step is to conduct policy-oriented behavioural research[149], particularly with a "focus on … AI-related social, legal and ethical implications and policy issues" as the Organisation for Economic Co-operation and Development recommends[150]. Anti-corruption research[18,151], AI safety research[107,152] 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[54,118], 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 probably

shapes AI agents' corrupting influence across all of the roles that we considered in this Review[54,55,70]. However, transparency can also create new tradeoffs (for example, by undermining efficiency)[55]. 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, as such strategic avoidance may nullify efforts towards 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 such as 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] and crowdsourcing citizens' preferences about ethical AI[131,159] are mostly concerned about the direct unethical consequences of AI behaviour 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 bad machines can corrupt good morals.

Received: 9 December 2020; Accepted: 26 April 2021;
Published online: 3 June 2021

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. USA* **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. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
7. Yang, G.-Z. et al. The grand challenges of science robotics. *Sci. Robot.* **3**, eaar7650 (2018).
8. Floridi, L. Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philos. Trans. A Math. Phys. Eng. Sci.* https://doi.org/10.1098/rsta.2016.0112 (2016).
9. Damiani, J. A voice deepfake was used to scam a CEO out of $243,000. *Forbes Magazine* https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/ (3 September 2019).
10. Robitzski, D. This grad student used a neural network to write his papers. *Futurism* https://futurism.com/grad-student-neural-network-write-papers (21 April 2020).
11. Lin, T. C. W. The new market manipulation. *Emory Law J.* **66**, 1253–1315 (2016).
12. Hakim, F. Z. M., Indrayani, L. M. & Amalia, R. M. A dialogic analysis of compliment strategies employed by Replika chatbot. In *Proc. 3rd International Conference of Arts, Language and Culture (ICALC 2018)* https://www.atlantis-press.com/proceedings/icalc-18/55913474 (Atlantis, 2019).
13. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nat. Mach. Intell.* **1**, 74–78 (2019).
14. *White Paper on Artificial Intelligence—A European Approach to Excellence and Trust* (EU Commission, 2020).
15. Plant, S. *Zeros and Ones: Digital Women and the New Technoculture* (Fourth Estate, 1997).
16. 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* (Wiley, 2017).
17. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence* (Knopf, 2017).
18. Mungiu-Pippidi, A. The time has come for evidence-based anticorruption. *Nat. Hum. Behav.* **1**, 0011 (2017).
19. Gino, F. Understanding ordinary unethical behavior: why people who value morality act immorally. *Curr. Opin. Behav. Sci.* **3**, 107–111 (2015).
20. Jones, T. M. Ethical decision making by individuals in organizations: an issue-contingent model. *Acad. Manag. Rev.* **16**, 366–395 (1991).
21. Cohn, A., Maréchal, M. A., Tannenbaum, D. & Zünd, C. L. Civic honesty around the globe. *Science* **365**, 70–73 (2019).
22. Treviño, L. K., Weaver, G. R. & Reynolds, S. J. Behavioral ethics in organizations: a review. *J. Manag.* **32**, 951–990 (2006).
23. Bazerman, M. H. & Gino, F. Behavioral ethics: toward a deeper understanding of moral judgment and dishonesty. *Annu. Rev. Law Soc. Sci.* **8**, 85–104 (2012).
24. Shalvi, S., Weisel, O., Kochavi-Gamlie, S. & Leib, M. in *Cheating, Corruption, and Concealment: the Roots of Dishonesty* (eds Van Prooijen, J. W. & Van Lange, P. A. M.) 134–148 (Cambridge Univ. Press, 2016).
25. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* **45**, 633–644 (2008).
26. Ariely, D. *The Honest Truth about Dishonesty: How We Lie to Everyone—Especially Ourselves* (HarperCollins, 2012).
27. Shalvi, S., Gino, F., Barkan, R. & Ayal, S. Self-serving justifications: doing wrong and feeling moral. *Curr. Dir. Psychol. Sci.* **24**, 125–130 (2015).
28. Cohn, A., Fehr, E. & Maréchal, M. A. Business culture and dishonesty in the banking industry. *Nature* **516**, 86–89 (2014).
29. Rahwan, Z., Yoeli, E. & Fasolo, B. Heterogeneity in banker culture and its influence on dishonesty. *Nature* **575**, 345–349 (2019).
30. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: a meta-analysis on dishonest behavior. *Psychol. Bull.* **145**, 1–44 (2019).
31. 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).
32. Gross, J., Leib, M., Offerman, T. & Shalvi, S. Ethical free riding: when honest people find dishonest partners. *Psychol. Sci.* **29**, 1956–1968 (2018).
33. Gross, J. & De Dreu, C. K. W. Rule following mitigates collaborative cheating and facilitates the spreading of honesty within groups. *Pers. Soc. Psychol. Bull.* **47**, 395–409 (2020).
34. Leib, M., Köbis, N. C., Soraperra, I., Weisel, O. & Shalvi, S. *Collaborative Dishonesty: a Meta-Study* CREED Working Paper Series (Univ. Amsterdam, 2021).
35. Thomas, P. S. et al. Preventing undesirable behavior of intelligent machines. *Science* **366**, 999–1004 (2019).
36. 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).
37. Koenecke, A. et al. Racial disparities in automated speech recognition. *Proc. Natl Acad. Sci. USA* **117**, 7684–7689 (2020).
38. He, Q., Turel, O. & Bechara, A. Brain anatomy alterations associated with social networking site (SNS) addiction. *Sci. Rep.* **7**, 45064 (2017).
39. Aral, S. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt* (Crown, 2020).
40. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
41. Soraperra, I. et al. The bad consequences of teamwork. *Econ. Lett.* **160**, 12–15 (2017).
42. 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.
43. Bicchieri, C. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (Oxford Univ. Press, 2016).
44. 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).
45. 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. *Behav. Public Policy* https://doi.org/10.1017/bpp.2019.37 (2019).
46. 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).
47. Köbis, N. C., Jackson, D. & Carter, D. I. in *A Research Agenda for Studies of Corruption* (eds Mungiu-Pippidi, A. & Heywood, P.) 41–53 (Edward Elgar, 2020).
48. Brandstetter, J. et al. A peer pressure experiment: recreation of the Asch conformity experiment with robots. In *Proc. 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* 1335–1340 (IEEE, 2014).
49. Shiomi, M. & Hagita, N. Do synchronized multiple robots exert peer pressure? In *Proc. 4th International Conference on Human Agent Interaction* 27–33 (Association for Computing Machinery, 2016).
50. Vollmer, A.-L., Read, R., Trippas, D. & Belpaeme, T. Children conform, adults resist: a robot group induced peer pressure on normative social conformity. *Sci. Robot.* **3**, eaat7111 (2018).
51. Salomons, N., van der Linden, M., Strohkorb Sebo, S. & Scassellati, B. Humans conform to robots: disambiguating trust, truth, and conformity. In *Proc. 2018 ACM/IEEE International Conference on Human–Robot Interaction* 187–195 (Association for Computing Machinery, 2018).

52. Hertz, N. & Wiese, E. Under pressure: examining social conformity with computer and robot groups. *Hum. Factors* **60**, 1207–1218 (2018).
53. Hertz, N., Shaw, T., de Visser, E. J. & Wiese, E. Mixing it up: how mixed groups of humans and machines modulate conformity. *J. Cogn. Eng. Decis. Mak.* **13**, 242–257 (2019).
54. Köbis, N. & Mossink, L. Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Human Behav.* **114**, 106553 (2021).
55. Ishowo-Oloko, F. et al. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Mach. Intell.* **1**, 517–521 (2019).
56. Song-Nichols, K. & Young, A. G. Gendered robots can change children's gender stereotyping. In *Proc. CogSci 2020* 2480–2485 (Cognitive Science Society, 2020).
57. 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 Proc. 17th *ACM Conference on Interaction Design and Children* 625–631 (Association for Computing Machinery, 2018).
58. Milgram, S. Behavioral study of obedience. *J. Abnorm. Psychol.* **67**, 371–378 (1963).
59. Burger, J. M. Replicating Milgram: would people still obey today? *Am. Psychol.* **64**, 1–11 (2009).
60. Gino, F., Moore, D. A. & Bazerman, M. H. *No Harm, No Foul: the Outcome Bias in Ethical Judgments* Harvard Business School NOM Working Paper (Harvard Univ., 2009).
61. Wiltermuth, S. S., Newman, D. T. & Raj, M. The consequences of dishonesty. *Curr. Opin. Psychol.* **6**, 20–24 (2015).
62. Fogg, B. J. Creating persuasive technologies: an eight-step design process. In *Proc. 4th International Conference on Persuasive Technology* 1–6 (Association for Computing Machinery, 2009).
63. Longoni, C. & Cian, L. Artificial intelligence in utilitarian vs. hedonic contexts: the 'word-of-machine' effect. *J. Mark.* https://journals.sagepub.com/doi/full/10.1177/0022242920957347 (2020).
64. AI reads human emotions. Should it? *MIT Technology Review* (14 October 2020).
65. How close is AI to decoding our emotions? *MIT Technology Review* (24 September 2020).
66. Giubilini, A. & Savulescu, J. The artificial moral advisor. The 'ideal observer' meets artificial intelligence. *Philos. Technol.* **31**, 169–188 (2018).
67. 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).
68. Castelo, N., Bos, M. W. & Lehmann, D. R. Task-dependent algorithm aversion. *J. Mark. Res.* **56**, 809–825 (2019).
69. Dietvorst, B., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
70. Leib, M., Köbis, N. C., Hagens, M., Rilke, R. & Irlenbusch, B. The corruptive force of AI-generated advice. Preprint at https://arxiv.org/abs/2102.07536
71. Robinette, P., Li, W., Allen, R., Howard, A. M. & Wagner, A. R. Overtrust of robots in emergency evacuation scenarios. In *Proc. 2016 ACM/IEEE International Conference on Human–Robot Interaction* 101–108 (2016).
72. Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monogr.* **70**, 1–70 (1956).
73. Larsen, K. S. The Asch conformity experiment: replication and transhistorical comparison. *J. Soc. Behav. Pers.* **5**, 163–168 (1990).
74. Wiltermuth, S. S. Cheating more when the spoils are split. *Organ. Behav. Hum. Decis. Process.* **115**, 157–168 (2011).
75. Ryvkin, D. & Serra, D. Corruption and competition among bureaucrats: an experimental study. *J. Econ. Behav. Organ.* **175**, 439–451 (2018).
76. 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).
77. Lambsdorff, J. G. & Frank, B. Corrupt reciprocity–experimental evidence on a men's game. *Int. Rev. Law Econ.* **31**, 116–125 (2011).
78. Schmidt, K. in *Distributed Decision Making: Cognitive Models for Cooperative Work* (eds Rasmussen, J. et al.) 75–110 (Wiley, 1991).
79. Hoc, J.-M. Towards a cognitive approach to human–machine cooperation in dynamic situations. *Int. J. Hum. Comput. Stud.* **54**, 509–540 (2001).
80. 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).
81. Suchman, L., Blomberg, J., Orr, J. E. & Trigg, R. Reconstructing technologies as social practice. *Am. Behav. Sci.* **43**, 392–408 (1999).
82. Chugunova, M. & Sele, D. *We and It: an Interdisciplinary Review of the Experimental Evidence on Human-Machine Interaction* https://doi.org/10.2139/ssrn.3692293 (SSRN, 2020).
83. Crandall, J. W. et al. Cooperating with machines. *Nat. Commun.* **9**, 233 (2018).
84. Calvano, E., Calzolari, G., Denicolò, V. & Pastorello, S. Artificial intelligence, algorithmic pricing and collusion. *Am. Econ. Rev.* **110**, 3267–3297 (2019).
85. 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).
86. Martinez-Miranda, E., McBurney, P. & Howard, M. J. W. Learning unfair trading: a market manipulation analysis from the reinforcement learning perspective. In *Proc. 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems* 103–109 (EAIS, 2016).
87. Mell, J., Lucas, G. & Gratch, J. in *Intelligent Virtual Agents* 273–282 (Springer, 2017).
88. 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).
89. Kirchkamp, O. & Strobel, C. Sharing responsibility with a machine. *J. Behav. Exp. Econ.* **80**, 25–33 (2019).
90. Pezzo, M. V. & Pezzo, S. P. Physician evaluation after medical errors: does having a computer decision aid help or hurt in hindsight? *Med. Decis. Mak.* **26**, 48–56 (2006).
91. Paravisini, D. & Schoar, A. *The Incentive Effect of Scores: Randomized Evidence from Credit Committees* Working Paper Series (National Bureau of Economic Research, 2013).
92. 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).
93. 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).
94. Houser, D. & Kurzban, R. Revisiting kindness and confusion in public goods experiments. *Am. Econ. Rev.* **92**, 1062–1069 (2002).
95. Coricelli, G. & Nagel, R. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proc. Natl Acad. Sci. USA* **106**, 9163–9168 (2009).
96. Frith, C. D. & Frith, U. The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
97. Schniter, E., Shields, T. W. & Sznycer, D. Trust in humans and robots: economically similar but emotionally different. *J. Econ. Psychol.* **78**, 102253 (2020).
98. De Melo, C., Marsella, S. & Gratch, J. People do not feel guilty about exploiting machines. *ACM Trans. Comput. Hum. Interact.* **23** (2016).
99. Mazar, N. & Ariely, D. Dishonesty in everyday life and its policy implications. *J. Public Policy Mark.* **25**, 117–126 (2006).
100. Köbis, N., Starke, C. & Rahwan, I. Artificial intelligence as an anti-corruption tool (AI-ACT)–potentials and pitfalls for top-down and bottom-up approaches. Preprint at https://arxiv.org/abs/2102.11567 (2021).
101. Drugov, M., Hamman, J. & Serra, D. Intermediaries in corruption: an experiment. *Exp. Econ.* **17**, 78–99 (2014).
102. 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).
103. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. The ethics of algorithms: mapping the debate. *Big Data Soc.* https://doi.org/10.1177/2053951716679679 (2016).
104. Gogoll, J. & Uhl, M. Rage against the machine: automation in the moral domain. *J. Behav. Exp. Econ.* **74**, 97–103 (2018).
105. 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**, 2527–2574 (2016).
106. 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).
107. Miller, T. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–36 (2019).
108. Gunning, D., Stefik, M., Choi, J. & Miller, T. XAI—explainable artificial intelligence. *Sci. Robot.* **4**, eaay7120 (2019).
109. 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).
110. Dana, J., Weber, R. A. & Kuang, J. X. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Econ. Theory* **33**, 67–80 (2007).
111. Hancock, J. T. & Guillory, J. in *The Handbook of the Psychology of Communication Technology* (ed. Sundar, S. S.) 270–289 (Wiley, 2015).
112. Seymour, J. & Tully, P. Weaponizing data science for social engineering: automated E2E spear phishing on Twitter. *Black Hat USA* https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf (2016).
113. Caldwell, M., Andrews, J. T. A., Tanay, T. & Griffin, L. D. AI-enabled future crime. *Crime Sci.* **9**, 14 (2020).

114. Sharkey, N., Goodman, M. & Ross, N. The coming robot crime wave. *Computer* **43**, 115–116 (2010).
115. Jagatic, T. N., Johnson, N. A., Jakobsson, M. & Menczer, F. Social phishing. *Commun. ACM* **50**, 94–100 (2007).
116. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
117. Brundage, M. et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Preprint at https://arxiv.org/abs/1802.07228 (2018).
118. Bendel, O. The synthetization of human voices. *AI Soc.* **34**, 83–89 (2019).
119. McKelvey, F. & Dubois, E. *Computational Propaganda in Canada: the Use of Political Bots* (Computational Propaganda Research Project, 2017).
120. Ostermaier, A. & Uhl, M. Spot on for liars! How public scrutiny influences ethical behavior. *PLoS ONE* **12**, e0181682 (2017).
121. 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).
122. 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).
123. Leyer, M. & Schneider, S. Me, you or AI? How do we feel about delegation. In *Proc. 27th European Conference on Information Systems (ECIS)* https://aisel.aisnet.org/ecis2019_rp/36 (2019).
124. Wellman, M. P. & Rajan, U. Ethical issues for autonomous trading agents. *Minds Mach.* **27**, 609–624 (2017).
125. Tenbrunsel, A. E. & Messick, D. M. Ethical fading: the role of self-deception in unethical behavior. *Soc. Justice Res.* **17**, 223–236 (2004).
126. Bazerman, M. H. & Banaji, M. R. The social psychology of ordinary ethical failures. *Soc. Justice Res.* **17**, 111–115 (2004).
127. Bazerman, M. H. & Tenbrunsel, A. E. *Blind Spots: Why We Fail to Do What's Right and What to Do about It*. (Princeton Univ. Press, 2012).
128. Sloane, M. & Moss, E. AI's social sciences deficit. *Nat. Mach. Intell.* **1**, 330–331 (2019).
129. Irving, G. & Askell, A. AI safety needs social scientists. *Distill* **4**, e14 (2019).
130. Crawford, K. & Calo, R. There is a blind spot in AI research. *Nature* **538**, 311–313 (2016).
131. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
132. Bigman, Y. E., Waytz, A., Alterovitz, R. & Gray, K. Holding robots responsible: the elements of machine morality. *Trends Cogn. Sci.* **23**, 365–368 (2019).
133. 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).
134. Fisman, R. & Golden, M. How to fight corruption. *Science* **356**, 803–804 (2017).
135. De Angeli, A. Ethical implications of verbal disinhibition with conversational agents. *PsychNology J.* **7**, 49–57 (2009).
136. McDonnell, M. & Baxter, D. Chatbots and gender stereotyping. *Interact. Comput.* **31**, 116–121 (2019).
137. Schwickerath, A. K., Varraich, A. & Smith, L.-L. How to research corruption. In *Conference Proceedings Interdisciplinary Corruption Research Forum* (eds Schwickerath, A. K. et al.) 7–8 (Interdisciplinary Corruption Research Network, 2016).
138. Salganik, M. J. *Bit by Bit* (Princeton Univ. Press, 2017).
139. Fisman, R. & Miguel, E. Corruption, norms, and legal enforcement: evidence from diplomatic parking tickets. *J. Polit. Econ.* **115**, 1020–1048 (2007).
140. Pierce, L. & Balasubramanian, P. Behavioral field evidence on psychological and social factors in dishonesty and misconduct. *Curr. Opin. Psychol.* **6**, 70–76 (2015).
141. Dai, Z., Galeotti, F. & Villeval, M. C. Cheating in the lab predicts fraud in the field: an experiment in public transportation. *Manag. Sci.* **64**, 1081–1100 (2018).
142. Cohn, A. & Maréchal, M. A. Laboratory measure of cheating predicts school misconduct. *Econ. J.* **128**, 2743–2754 (2018).
143. Floridi, L. & Sanders, J. W. On the morality of artificial agents. *Minds Mach.* **14**, 349–379 (2004).
144. Hagendorff, T. Ethical behavior in humans and machines–evaluating training data quality for beneficial machine learning. Preprint at https://arxiv.org/abs/2008.11463 (2020).
145. Mullainathan, S. Biased algorithms are easier to fix than biased people. *The New York Times* https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html (6 December 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 (IEEE, 2017).
148. Russell, S., Dewey, D. & Tegmark, M. Research priorities for robust and beneficial artificial intelligence. *AI Mag.* **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/LEGAL/0449 (OECD, 2020).
151. Fisman, R. & Golden, M. A. *Corruption: What Everyone Needs to Know* (Oxford Univ. 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.* https://doi.org/10.1007/s00146-020-01045-4 (2020).
158. Finkel, A. What will it take for us to trust AI? *World Economic Forum* https://www.weforum.org/agenda/2018/05/alan-finkel-turing-certificate-ai-trust-robot (12 May 2018).
159. Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A. & Rahwan, I. Crowdsourcing moral machines. *Commun. ACM* **63**, 48–55 (2020).

Acknowledgements

We thank A. Bouza da Costa for designing the illustrations, and M. Leib and L. Karim for valuable comments on the manuscript. J.-F.B. acknowledges support from the Institute for Advanced Study in Toulouse, grant ANR-19-PI3A-0004 from the Artificial and Natural Intelligence Toulouse Institute and grant ANR-17-EURE-0010 from Investissements d'Avenir.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to N.K.

Peer review information *Nature Human Behaviour* thanks Thilo Hagendorff and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.