The corrupting influence of AI as a boss or Counterparty

Hal Ashton [* 1]   **Matija Franklin** [* 2]

Abstract

In a recent article Köbis et al. (2021) propose a framework to identify four different primary roles in which Artificial Intelligence (AI) cause unethical or corrupt human behaviour; namely - role model, delegate, partner, and advisor. In this article we propose two further roles - AI as boss and AI as counterparty. We argue that the AI boss exerts coercive power over its employees whilst the different perceptual abilities of an AI counterparty provide an opportunity for humans to behave differently towards them than they would with human analogues. Unethical behaviour towards the AI counterparty is rationalised because it is not human. For both roles, the human will will typically not have any choice about their participation in the interaction.

1. Intro

In their recent review article "Bad machines corrupt good morals", Köbis et al. (2021) propose a framework to identify four different primary roles in which Artificial Intelligence (AI) agents cause unethical or corrupt human behaviour. Namely, they propose that an AIs can take the role of a role model - an agent humans imitate, delegate - an agent humans delegate to and oversee, partner - an agent humans collaborate with, and an advisor - an agent humans receive advice from. Advisors and role models corrupt humans by being influencers - the AI agents *nudge humans towards unethical behaviour by displaying or recommending actions with harmfulk consequences*. Partners and delegates corrupt humans by being enablers - the AI agents *are involved in (un)ethical behaviour and allow people to pursue selfish goals while reducing guilt*.

We would like to propose two further roles - AI as boss and

*Equal contribution [1]Causal Cognition Lab, University College London [2]Department of Engineering, University of Cambridge. Correspondence to: Matija Franklin <matija.franklin@ucl.ac.uk>.

AI as counterparty - which do not fit into the framework of Köbis et al. but are nevertheless pertinent when considering the corrupting influence of AI agents on humans. The characteristics of these roles are summarised in Figure 1. As with Köbis et al. we will use the canonical fraud triangle of pressure, rationalisation and opportunity (see (Tickner & Button, 2021) for its history) to analyse the corruption mechanisms involved. The AI boss is characterised by the coercive power it may wield over employees who have no choice but to engage with it. Similarly, people might have no option but to transact with the AI counterparty in roles which might have previously been filled by a human. Deficiencies in the perception or cognitive properties of the AI counterparty provide an opportunity for unethical behaviour. Unethical behaviour towards the AI counterparty is often rationalised because it is not human (and governed by social norms). Nevertheless transactions are still governed by standard laws even if the harms from unethical behaviour are less obvious to the actor. We also argue that behaviour which would be deemed unacceptable if the AI counterparty were human, is also problematic for reasons of character corruption.

2. The coercive power of the AI Boss

Workers in the 21st century are increasingly being told what to do by AI overseers. Adams-Prassl (2019) argues that wide ranging, cost effective data collection and surveillance abilities (Moore, 2020) in the workplace coupled with machine learning techniques to process the data and the acceptance of various control mechanisms pioneered in the gig economy have extended Digital Taylorism (Cole et al., 2020) to create the AI boss. (Parent-Rocheleau & Parker, 2022) identify six tasks of algorithmic management: Monitoring, Goal setting, Performance management, Scheduling, Compensation and Job termination. The power of the AI boss is acknowledged in the EU's recent proposed AI Act(CNECT, 2021) identifying AI management as a high risk application (Kelly-Lyth, 2021).

According to Allen Wood's definition of coercion (Wood, 2014), someone is coerced if they are forced to to something that they do not choose or if they choose to do that thing because they have no other acceptable choice. When someone has to act in a certain way or else they will lose their job,

they are being coerced. A coercive relationship between AI boss does not fit easily into the corruption framework of Köbis et al.. The roles that they describe (Advisor, Role Model, Partner and Delegate) are voluntary in their nature. Workers typically do not get given the chance to choose whether they come under the control of an AI boss[1]. As Maas (2022) puts it, workers with AI bosses are in danger of being subject to the neo-republican moral wrong of domination, that is to say 'subject to superior and unaccountable power'. Muldoon & Raekstad (2022) argue that workers in the gig economy are subject to algorithmic domination facilitated in part by the asymmetric access to information that the AI boss has.

It should be noted that as an intermediary between manager and worker, the same AI boss plays two roles. As an actor which human mangers have delegated powers to, it can be classified as an 'AI delegate enabler' and this fits well into the Köbis et al. framework. It is the second role, direct manager to employees, that does not fit. We argue that the causal mechanisms that might produce unethical behaviour in the employee when being managed by an AI boss are fundamentally different from the manager turning a blind eye to the AI boss' behaviour. We do not make the claim that AI bosses are instructing their charges to do unethical things directly. Instead we suggest that in order to fulfil their tasks, employees are necessarily required to behave unethically. The ways and degree to which they do this may be altered by the fact that their manager is an AI boss.

Evidence of bad behaviour engendered by the AI boss is beginning to emerge. Management Algorithms lack empathy and have difficulty contextualising (Todolí-Signes, 2021); Task scheduling algorithms operate in an idealised world, where there are no difficulties in finding items or locations and humans do not tire and require breaks. Drowsy driving by ride-sharing operatives (increasingly designated employees) is an acknowledged problem (Berneking et al., 2018) with drivers often working shifts after their day jobs. A survey of gig-economy (Christie & Ward, 2019) found that delivery drivers felt pressurised to violate traffic regulations (for example speeding) in order to complete more deliveries and were putting themselves in danger of collisions. They also found that workers' phones were distracting drivers with job offers with insistent notifications. Warehouse workers are put under pressure to work at a pace that risks their personal health (Delfanti, 2021). Whereas employees may have felt able to negotiate with their managers when presented with unreasonable work loads or requests, their ability to do so with an AI Boss is curtailed. Workers may blindly follow orders given to them by an AI boss due to obedience effects coupled with undue respect given

to the aptitude of AI (Christie & Ward, 2019). This is described as the 'marginalisation of human sense-making' by (Leicht-Deobald et al., 2019) who consider the problem of AI enabled HR systems reducing worker's integrity (defined as a built in moral compass which employees can use to compare their actions against their standards). They argue that such systems tip the balance in organisations too far in the direction of worker compliance from worker integrity. One reason for this is the incapability of current management algorithms to reason about courses of actions which necessarily involve unethical means to bring about targets.

3. The corrupting effect of the AI Counterparty

The corporate rush to harness AI is in part motivated by promise of its economic advantages over humans (Brynjolfsson & McAfee, 2012). AI agents can work much faster than humans at certain tasks, do not tire, don't need paying and are not regulated and taxed in the same way. As a result, humans are encountering AI agents in roles which were previously filled by other humans. Despite their strengths, the deployed AI agents also display perceptual or cognitive shortcomings which might be exploited by their human counterparties. Sometimes this exploitation is explicitly norm breaking and sometimes it might be more subtly corrupting.

We argue that this potentially corrupting effect is not covered in the Köbis et al. framework. The role of an AI agent as a counterparty does not fit comfortably either into the classes of influencer or enabler. We will argue for AI as a corrupting counterparty in two ways. First, humans will respond to the asymmetry between their abilities and that of the AI with adversarial behaviour. Second, because AI counterparties are not human, people will feel freer to abuse them. Further, virtue ethics would suggest that such behaviour is potentially damaging to one's moral character (Coeckelbergh, 2021).

Humans and AI have asymmetric abilities when transacting. There is evidence that people perceive this asymmetry and behave differently as a result. (Bogert et al., 2021) show that as task difficulty increases, people rely more on algorithmic advice than on their own judgements. This effect remained regardless of the quality of the AI's advice, the human's capacity, and in the presence of other sources of non-algorithmic advice. (Malle et al., 2015) show that people also have expectations about an AI's ability to optimise in the domain of moral judgements. Compared with humans, robots were expected more strongly to make a utilitarian moral choice. They were also blamed more than humans when they did not make such a choice.

In research exploring human-robot competition, (Fraune et al., 2019) show that groups of humans (rather than indi-

---

[1]Job seekers who have met with no success via automated video interviews have succeeded when insisting on human assessment (Ajunwa, 2021)
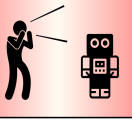
Figure 1. Two additional roles in which AI agents and humans influence ethical behavior. The main mechanisms and some fears (aka examples) are displayed in each role.

viduals) show more competition towards robots. When the number of members in a group of people is matched by the number of their robot competitors, participants displayed the most fear and competed more. (Karpus et al., 2021) present evidence in human-AI cooperation that people are able to detect an AI's weakness and are willing to exploit it. Participants were able to predict the cooperative behaviours of a benevolent AI and tended to exploit it more so than a benevolent human.

Aside from the opportunity that the asymmetries between AI and Human present, people may feel freer (or more justified) to treat AI agents in an abusive or unethical way than they would a human. A study using methods from Milgram's experiment on obedience, using a robot as the target, found that participants were less concerned for the abuse of robots than that of other humans (Bartneck et al., 2005). All participants used the highest voltage setting on the robot, which more than doubled Milgram's original findings. (Nomura et al., 2016) show that abuse of robots is observed in children, finding that the abuse continued even when half of them believed that the robot was capable of perceiving it. People's willingness to abuse can partially be explained with fMRI evidence showing that participants experience more emotional distress and empathetic concern for videos of abuse behaviour towards humans rather than robots (Rosenthal-Von Der Pütten et al., 2014). People's abuse gets worse for robots they perceive as less intelligent (Bartneck & Hu, 2008), bigger (rather than smaller; (Lucas et al., 2016)) and less human-like (Riek et al., 2009).

Given that an AI counterparty will be owned by someone, certain types unethical behaviour may be criminal (theft from a vending machine is theft irrespective of the sentience of the machine). There is evidence to suggest that people are more willing to commit crime in front of a machine observer rather than a human observer. A study of dishonest behaviour at supermarket self-service checkouts found that more customer thefts occurred when the staff's presence was reduced (Creighton et al., 2015). Both perceived and actual staff presence reduced thefts. (Forlizzi et al., 2016) show that although human observers elicit less attention from a participant than robots, they will evoke more of a socially normative presence. This presence, in turn causes participants to act honestly. Even when participants thought that they were being monitored they would still behave more dishonestly in the presence of a robot observer .

Virtue ethics would have it that actions that harm an AI are potentially damaging to one's moral character (Coeckelbergh, 2021). Many researchers, most notably Shannon Vallor (Vallor, 2016) and Robert Sparrow (Sparrow, 2016) apply virtue ethics to argue for the corrupting nature of abusing robots to one's moral character. Such a perspective argues that virtue is performed, creating habits. There is evidence for such a perspective. People's behaviour in one context influences their behaviour in another context (Dolan & Galizzi, 2015) and can also lead to the emergence of new preferences (Ariely & Norton, 2008). Indeed, people's everyday interactions with AI (Ashton & Franklin, 2022; Franklin et al., 2022; Franklin, 2022a), robots (Franklin & Ashton, 2022), and wearable devices (Franklin, 2022b) will also influence the emergence of future behaviour and preference. In the context of AI abuse, the act may have negative spillovers in the way an individual acts towards other people. Sánchez Ramos et al. (2018); Strait et al. (2018) find that people are more likely to make sexist comments about online videos depicting female-gendered robots (gynoids) than female (humans) and that they are more like to make dehumanizing comments if the robots are racialized as Asian or Black. In the context of sex robots, Danaher (2017) analyses a common related concern known as the 'symbolic consequences argument' which worries that the symbolism of sex with robots may reveal disturbing attitudes towards, amongst other things, consent. He concludes that an experimental approach is required to verify the consequences of this symbolism. Reviews of empirical research on virtue ethics suggests that there is reason to think that people's interactions with robots may causally affect their virtue (Coghlan et al., 2019). This has led some to argue for

the design of robots which promotes better virtue from people, such as incorporating an ability for consent (Sparrow, 2016).

4. Conclusion

This article extends the framework of Köbis et al. (2021) to include the roles of AI as boss and AI as counterparty which we argue are both current realities and deserve their own classification. A feature in common with both roles is that the human in the relationship will typically not have any choice about its existence. In both cases, corruption occurs in response to the presence of the AI agent. The AI boss can exert coercive power to corrupt humans by giving orders and, for employees that do not follow those orders, firing them from their jobs. As discussed, there is already evidence of coerced behaviour engendered by AI bosses (Christie & Ward, 2019). Employees managed by AI bosses risk corruption by following orders unquestioningly either through reverence to the all knowing algorithm or simply to keep their jobs. As AI agents proliferate, so will transactions between humans and AI agents. We argue that this counterparty relationship has the power to engender corruption of people's behaviour. People perceive the opportunity that the asymmetry in capabilities with their AI counterparty gives and behave differently. In particular they feel freer to treat autonomous machines in abusive ways. In many cases this behaviour norm breaking since the AI counterparty is owned and working on behalf of someone else. Even if this is not the case, we argue from a virtue ethics perspective, abusive behaviour towards AI counterparties is a corrupting and might encourage harms directed at humans in the future.

References

Adams-Prassl, J. What if your boss was an algorithm? economic incentives, legal challenges, and the rise of artificial intelligence at work. *Comp. Lab. L. & Pol'y J.*, 41:123, 2019.

Ajunwa, I. Automated Video Interviewing as the New Phrenology. 2021. doi: 10.15779/Z38RX93F1Q. URL https://lawcat.berkeley.edu/record/1246446. Publisher: Berkeley Technology Law Journal.

Ariely, D. and Norton, M. I. How actions create–not just reveal–preferences. *Trends in cognitive sciences*, 12(1): 13–16, 2008.

Ashton, H. and Franklin, M. The problem of behaviour and preference manipulation in ai systems. In *CEUR Workshop Proceedings*, volume 3087. CEUR Workshop Proceedings, 2022.

Bartneck, C. and Hu, J. Exploring the abuse of robots. *Interaction Studies*, 9(3):415–433, 2008.

Bartneck, C., Rosalia, C., Menges, R., and Deckers, I. Robot abuse-a limitation of the media equation. *Proceedings of the Interact 2005 Workshop on Abuse*, 2005.

Berneking, M., Rosen, I. M., Kirsch, D. B., Chervin, R. D., Carden, K. A., Ramar, K., Aurora, R. N., Kristo, D. A., Malhotra, R. K., Martin, J. L., Olson, E. J., Rosen, C. L., Rowley, J. A., and Gurubhagavatula, I. The Risk of Fatigue and Sleepiness in the Ridesharing Industry: An American Academy of Sleep Medicine Position Statement. *Journal of Clinical Sleep Medicine*, 14 (04):683–685, April 2018. ISSN 1550-9389, 1550-9397. doi: 10.5664/jcsm.7072. URL http://jcsm.aasm.org/doi/10.5664/jcsm.7072.

Bogert, E., Schecter, A., and Watson, R. T. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):1–9, 2021.

Brynjolfsson, E. and McAfee, A. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Brynjolfsson and McAfee, 2012.

Christie, N. and Ward, H. The health and safety risks for people who drive for work in the gig economy. *Journal of Transport & Health*, 13:115–127, June 2019. ISSN 22141405. doi: 10.1016/j.jth.2019.02.007. URL https://linkinghub.elsevier.com/retrieve/pii/S2214140518305772.

CNECT. Proposal for a regulation of the European parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Technical Report COM/2021/206, European Commission, Directorate-General for Communications Networks, Content and Technology, April 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

Coeckelbergh, M. Does kindness towards robots lead to virtue? A reply to Sparrow's asymmetry argument. *Ethics and Information Technology*, 23(4):649–656, December 2021. ISSN 1388-1957, 1572-8439. doi: 10.1007/s10676-021-09604-z. URL https://link.springer.com/10.1007/s10676-021-09604-z.

Coghlan, S., Vetere, F., Waycott, J., and Neves, B. B. Could social robots make us kinder or crueller to humans and animals? *International Journal of Social Robotics*, 11(5): 741–751, 2019.

Cole, M., Radice, H., and Umney, C. The political economy of datafication and work: A new digital Taylorism. *Socialist Register*, 57:78–99, 2020.

Creighton, S., Johnson, G., Robertson, P., Law, J., and Szymkowiak, A. Dishonest behavior at self-service checkouts. In *International Conference on HCI in Business*, pp. 267–278. Springer, 2015.

Danaher, J. The Symbolic-Consequences Argument in the Sex Robot Debate. In *Robot Sex: Social and Ethical Implications*. MIT Press, 2017.

Delfanti, A. Machinic dispossession and augmented despotism: Digital work in an Amazon warehouse. *New Media & Society*, 23(1):39–55, January 2021. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444819891613. URL http://journals.sagepub.com/doi/10.1177/1461444819891613.

Dolan, P. and Galizzi, M. M. Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16, 2015.

Forlizzi, J., Saensuksopa, T., Salaets, N., Shomin, M., Mericli, T., and Hoffman, G. Let's be honest: A controlled field study of ethical behavior in the presence of a robot. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 769–774. IEEE, 2016.

Franklin, M. The influence of explainable artificial intelligence: Nudging behaviour or boosting capability? *arXiv preprint arXiv:2210.02407*, 2022a.

Franklin, M. Virtual spillover of preferences and behavior from extended reality. *CHI Conference on Human Factors in Computing Systems (CHI '22)*, 2022b.

Franklin, M. and Ashton, H. Preference change in persuasive robotics. *arXiv preprint arXiv:2206.10300*, 2022.

Franklin, M., Ashton, H., Gorman, R., and Armstrong, S. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of ai. *arXiv preprint arXiv:2203.10525*, 2022.

Fraune, M. R., Sherrin, S., Šabanović, S., and Smith, E. R. Is human-robot interaction more competitive between groups than between individuals? In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 104–113. IEEE, 2019.

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., and Deroy, O. Algorithm exploitation: humans are keen to exploit benevolent ai. *iScience*, pp. 102679, 2021.

Kelly-Lyth, A. Dispatch No. 39 - European Union The AI Act and Algorithmic Management. *Comparitive Labor Law and Policy Journal*, (39):9, 2021.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6): 679–685, June 2021. ISSN 2397-3374. doi: 10.1038/s41562-021-01128-2. URL http://www.nature.com/articles/s41562-021-01128-2.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., and Kasper, G. The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. *Journal of Business Ethics*, 160(2):377–392, December 2019. ISSN 0167-4544, 1573-0697. doi: 10.1007/s10551-019-04204-w. URL http://link.springer.com/10.1007/s10551-019-04204-w.

Lucas, H., Poston, J., Yocum, N., Carlson, Z., and Feil-Seifer, D. Too big to be mistreated? examining the role of robot size on perceptions of mistreatment. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 1071–1076. IEEE, 2016.

Maas, J. Machine learning and power relations. *AI & SOCIETY*, February 2022. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-022-01400-7. URL https://link.springer.com/10.1007/s00146-022-01400-7.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 117–124. IEEE, 2015.

Moore, P. V. Data subjects, digital surveillance, AI and the future of work. Technical Report 656.305, European Parliamentary Research Service EPRS, 2020. URL https://data.europa.eu/doi/10.2861/879078.

Muldoon, J. and Raekstad, P. Algorithmic Domination in the Gig Economy. *European Journal of Political Theory*, pp. 147488512210820, February 2022. ISSN 1474-8851, 1741-2730. doi: 10.1177/14748851221082078. URL http://journals.sagepub.com/doi/10.1177/14748851221082078.

Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., and Yamada, S. Why do children abuse robots? *Interaction Studies*, 17(3):347–369, 2016.

Parent-Rocheleau, X. and Parker, S. K. Algorithms as work designers: How algorithmic management influences the design of jobs. *Human Resource Management Review*, 32(3):100838, September 2022. ISSN 10534822. doi: 10.1016/j.hrmr.2021.100838. URL https://linkinghub.elsevier.com/retrieve/pii/S1053482221000176.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 245–246, 2009.

Rosenthal-Von Der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., Brand, M., and Krämer, N. C. Investigations on empathy towards humans and robots using fmri. *Computers in Human Behavior*, 33:201–212, 2014.

Sparrow, R. Kicking a robot dog. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 229–229, 2016. doi: 10.1109/HRI.2016.7451756.

Strait, M., Ramos, A. S., Contreras, V., and Garcia, N. Robots Racialized in the Likeness of Marginalized Social Identities are Subject to Greater Dehumanization than those racialized as White. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 452–457, Nanjing, August 2018. IEEE. ISBN 978-1-5386-7980-7. doi: 10.1109/ROMAN.2018.8525610. URL https://ieeexplore.ieee.org/document/8525610/.

Sánchez Ramos, A. C., Contreras, V., Santos, A., Aguillon, C., Garcia, N., Rodriguez, J. D., Amaya Vazquez, I., and Strait, M. K. A Preliminary Study of the Effects of Racialization and Humanness on the Verbal Abuse of Female-Gendered Robots. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 227–228, Chicago IL USA, March 2018. ACM. ISBN 978-1-4503-5615-2. doi: 10.1145/3173386.3177075. URL https://dl.acm.org/doi/10.1145/3173386.3177075.

Tickner, P. and Button, M. Deconstructing the origins of Cressey's Fraud Triangle. *Journal of Financial Crime*, 28 (3):722–731, August 2021. ISSN 1359-0790, 1359-0790. doi: 10.1108/JFC-10-2020-0204. URL https://www.emerald.com/insight/content/doi/10.1108/JFC-10-2020-0204/full/html.

Todolí-Signes, A. Making algorithms safe for workers: occupational risks associated with work managed by artificial intelligence. *Transfer: European Review of Labour and Research*, pp. 102425892110350, August 2021. ISSN 1024-2589, 1996-7284. doi: 10.1177/10242589211035040. URL http://journals.sagepub.com/doi/10.1177/10242589211035040.

Vallor, S. *Technology and the virtues: A philosophical guide to a future worth wanting.* Oxford University Press, 2016.

Wood, A. W. Coercion, Manipulation, Exploitation. In *Manipulation: theory and practice.* Oxford University Press, Oxford ; New York, 2014. ISBN 978-0-19-933821-4 978-0-19-933820-7. URL DOI:10.1093/acprof:oso/9780199338207.003.0002.