

# The corrupting influence of AI as a boss or Counterparty

Hal Ashton, University of Cambridge  
Matija Franklin, UCL

The corrupting influence of AI as a boss or Counterparty



AI as coercer			AI as counterparty		
AI agents exert coercive power (redundancy) over humans by forcing them to behave a certain way to avoid a bad personal outcome			AI agents replace roles previously fulfilled by humans. People respond in unethical ways, feeling unconstrained from norms which normally regulate inter-human encounters.		
<ul style="list-style-type: none"><li>Unrealistic targets</li><li>Lack of compassion</li><li>No recourse for negotiation with AI</li></ul>	Delivery drivers drive recklessly and without rest in order to fulfil AI generated targets		<ul style="list-style-type: none"><li>Asymmetric abilities give opportunities</li><li>'Not a person' rationalisation</li></ul>	People more likely to steal from automated checkouts Robot abuse might spill over to human abuse	
Mechanisms	Fears	Boss	Mechanisms	Fears	Counterparty

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.

## AI Boss

- Workers in the 21st century are increasingly being told what to do by AI overseers
- Workers in the gig economy are subject to algorithmic domination facilitated in part by the asymmetric access to information that the AI boss has.
- 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.

## AI Counterpart

- Humans will respond to the asymmetry between their abilities and that of the AI with adversarial behaviour
- Because AI counterparties are not human, people will feel freer to abuse them.
- Virtue ethics would suggest that such behaviour is potentially damaging to one's moral character
- Such a perspective argues that virtue is performed, creating habits.
- Some argue for the design of robots which promotes better virtue from people, such as incorporating an ability for consent.