Moral Skill and Artificial Intelligence

Executive Summary

As humans, our skills define us. No skill is more human than the exercise of moral judgment. We are already using Artificial Intelligence (AI) to automate morally-loaded decisions. In other domains of human activity, automating a task diminishes our skill at that task. Will 'moral automation' diminish our moral skill? If so, how can we mitigate that risk, and adapt AI to enable moral 'upskilling'? Our project—a partnership with the Humanising Machine Intelligence (HMI) grand challenge at the Australian National University—will use philosophy, social psychology, and computer science to answer these questions.

The first stage of treatment is diagnosis. We begin by identifying both existing and prospective varieties of moral automation, before exploring the philosophical and social-psychological foundations of the argument from moral automation to moral deskilling. In doing so, we will determine just why, and how much, we should be worried about moral deskilling.

Treatment in this case comprises both mitigation and adaptation. We will propose technological and institutional solutions to mitigate the risk of moral deskilling. But we will also argue that AI systems will enable us to adapt to the challenge of automation, by morally upskilling in other areas. In particular, some measure of moral automation will free us up to pursue the morally most demanding aspects of our personal relationships; AI research will enable new kinds of moral knowledge and moral inquiry; and by affording us new understandings and capacities, AI can make new kinds of moral behaviour possible.

Our project will produce scholarship of the highest order. But our goals are not narrowly academic. Through the HMI project, we will translate our research to maximise its impact at all levels of Australian, and global society.

Personnel

Project Director: Professor Seth Lazar

Project Co-Director: Dr Claire Benn

Chief Investigator: Colin Klein

Chief Investigator: Jenny Davis

Chief Investigator: Toni Erskine

Project Description

INTRODUCTION

As humans, our skills define us. No skill is more human than the exercise of moral judgment. We are already using Artificial Intelligence (AI) to automate morally-loaded decisions. In other domains of human activity, automating a task diminishes our skill at that task. Will 'moral automation' diminish our moral skill (Vallor (2015))? If so, how can we mitigate that risk, and adapt AI to enable moral 'upskilling'? Drawing on philosophy, social psychology, and computer science, we will answer these questions.

MORAL AUTOMATION

Moral automation is already here. We use AI in criminal justice, in policing, in the allocation of welfare, and in recruitment and finance. Autonomous vehicles and weapons systems, rescue and care robots, are ready to go. However, moral automation has the potential to become even more socially pervasive than it is today. AI is a general-purpose technology. It can help us make any kind of decision better. As private individuals, we already rely on AI for many morally lightweight purposes. The stakes will rise. Personal AI already helps us manage our calendars; it may soon decide between conflicting commitments for us. Smart devices help us parent; AI's role in childcare will only grow. Robots sweep our floors; soon we will vest many other duties of mutual care in AI-powered service robots. AI already mediates our electronic communications with others; today's word-choice recommendations may turn into much more substantive moral advice in the future.

FROM AUTOMATION TO DESKILLING

Since these technologies are so powerful and so new, we cannot await a longitudinal study of how moral automation affects human moral skill. Instead, we must use philosophical analysis to test the theory behind the deskilling argument, and experimental social psychology to test its empirical claims.

There are two philosophical arguments from moral automation to moral deskilling. First: moral judgment is a kind of practical wisdom, a skill developed only through practical experience. If moral automation deprives people of the experience necessary to develop this practical wisdom, then it will undermine their acquisition of moral skill.

This argument is controversial. While moral behaviour is unarguably a variety of practical wisdom, perhaps we can acquire skill in moral judgment purely intellectually. We will explore this dialectic, drawing on moral philosophy and the philosophy of action. Although the argument from practical wisdom may be strongest from some specific philosophical standpoints (e.g. that of virtue ethics), we expect to argue that almost everyone should endorse its conclusions, since moral judgment plausibly requires sensitivity to morally relevant properties, which one is unlikely to have in the absence of significant relevant experience.

The second argument draws on two well-established phenomena. First, people tend to defer to automated systems (Parasuraman and Riley (1997)). Second, AI systems tend to be inscrutable—though they deliver reliable verdicts, they do not 'show their working'. If moral automation is inscrutable in this way, and if we defer to these systems, then we will increasingly dissociate verdicts in morally-loaded decisions from the reasons for which they are reached. Imagine, for example, that instead of a high court supporting their verdicts with an analysis of the case, we simply had an AI that ruled one way or the other. The task of jurisprudence would be to predict which way the system would decide, rather than think through the reasons behind its judgments. This would lead to a significant reduction in jurisprudential skill. The same would be true for moral automation and moral skill.

We expect to show that the argument for moral deskilling has solid philosophical foundations. So, it is worth also exploring its empirical support. This will involve drawing on existing sociological and psychological research on deskilling in other areas. But we will also construct experiments to test for the underlying causal mechanisms through which moral deskilling may occur. For example, we can give lab participants the opportunity to cheat on a task, and test if

they are more likely to take this opportunity following activities in which moral decision-making is done by the participant, partially automated, or entirely automated.

ASSESSING DESKILLING

If moral automation poses a risk of moral deskilling, the next task is to figure out whether and how much this matters. This will help us both show the urgency of intervening to prevent deskilling, and identify which interventions are most promising.

Even if moral automation helps implement our values in the short run, if it leads to moral deskilling then we will likely be worse off in the end, for at least two reasons.

First, one of the 'ironies of automation' is that the more efficient the automated system becomes, the more prone it is to lead to human deskilling, but also the more important human skill becomes when the system fails. If the automated system can easily handle less complex tasks, then when it fails, it will call upon high-level human skills that its widespread employment has made it hard for us to develop (Bainbridge (1983)).

For example, existing AI systems are notoriously easy to spoof—they are overly sensitive to irrelevant properties of the things they seek to classify. This makes them ill-equipped to deal with novel cases that do not have any straightforward counterparts in their training data. Humans, by contrast, are expert at reasoning by incomplete analogy, and extending their understanding to new edge cases. But our expertise derives from our engagement with the simple cases. Moral automation may deprive us of the training necessary to address the very cases in which automated systems fail.

Second, even if automated systems do not fail, the norms we want them to implement will evolve. But AI systems are inherently conservative—they train on past data, and reproduce past values. To ensure that moral automation keeps pace with evolving norms—as well as to accommodate reasonable moral pluralism—we must continue to cultivate the moral skill necessary to continually perform 'moral upgrades' on those automated systems.

These are instrumental worries about moral deskilling. But we will also explore non-instrumental reasons, derived from both Aristotle and Kant. On the one hand, the existence of morally skilled human agents is itself non-instrumentally valuable (Vallor (2015)). And on the other, part of what it means to treat one another with appropriate concern and respect may just be to exercise moral skill in deliberating over actions that seriously affect others' lives.

REMEDIATING DESKILLING, ENABLING UPSKILLING

The goal of diagnosing the nature and risk of deskilling is to treat it. Different varieties of moral automation will necessitate different kinds of response. In some areas, moral deskilling may be a price we have to pay for the associated moral benefits. In other areas, the risks of deskilling may be so great, and so hard to mitigate, that we should avoid moral automation entirely.

Where remediation is necessary, we will focus on two treatment paths: institutional and technological. Our participation in the broader Humanising Machine Intelligence (HMI) project at ANU will be invaluable on both counts (the central task of the HMI project is the technical challenge of designing moral AI, therefore the TWCF-funded project would supplement, rather than overlap with, core activities of HMI—see hmi.anu.edu.au). Through the HMI team, we have access not only to social scientists with insight into institutional design, but also to a project manager who will help translate our research to maximise impact, and to developers and end-

users of AI technologies, who can bring that impact about. And the HMI team includes some of Australia's leading AI researchers, who will be invaluable as we explore technological solutions.

The institutional structures that shape moral automation can also shape the risk of moral deskilling. Whether we acquire moral skill through practical experience, or through theoretical enquiry, we can clearly maintain our collective moral competence without humans taking every morally-loaded decision. To avoid deskilling, we must maintain pluralism in how we take morally- loaded decisions. This means ensuring that we do not excessively rely on automated systems, but maintain enough human- operated systems to preserve moral capacity.

It will also be crucial to ensure plurality within the AI systems themselves—thus enabling a greater role for skilled human judgment both in designing new systems, and in selecting among them, as well as promoting reasonable moral pluralism generally. To achieve this, we may need to find ways to lower barriers to entry for companies seeking to develop new approaches to AI.

Automation encourages complacency in human overseers, right to the point of catastrophic collapse. Conversely, requiring human input at every stage defeats the object of automation. The technological solution, then, may be to ensure that we design AI systems that seek human input for a random sample of the different kinds of decisions being made, ensuring productive (and engaging) ways for human operators to exercise their judgment, without losing the benefits of automation.

It will also be vital to address the second path to moral deskilling, by developing AI systems that can present reasons for their verdicts, potentially enhancing theoretical moral knowledge. In partnership with research fellows on the HMI project, we will explore how work on the philosophy of explanation can help us make AI less inscrutable.

The capstone of our project will be a sustained exploration of how humans can adapt ourselves and our AI systems to enable genuine moral growth, in at least three ways:

First, if personal AI takes some kinds of mutual care out of our hands, it can enable other means of attending to our loved ones' needs. For example, if service robots managed the ordering of groceries and preparation of meals, that would be moral automation—but it could enable us to be more involved, committed parents.

Second, technological progress often enables theoretical moral progress—think of the invention of writing, or of the printing press. AI too will profoundly enhance the scope of moral inquiry. Basic research into moral AI, of the kind that the HMI project will independently pursue, will make new kinds of moral knowledge possible. Teaching morality to autonomous agents through reinforcement learning will shed light on the nature of morality. AI systems will also help us do moral philosophy—for example, by helping knit together our considered judgments on a wider range of cases, and separate out reliable from self-serving moral intuitions.

Third, AI offers us unprecedented abilities to understand and intervene in the world. Philosophers often argue that 'ought implies can'. But sometimes, the converse is true. Consider an analogy: technological advances now enable us to fulfil our duties to aid other humans in ways that were unthinkable just decades ago. In the past, to help the most vulnerable people in the world one would have to relocate. A new kind of moral behaviour is now possible, and with it a new kind of moral inquiry. AI will afford us an understanding of the world, and of the potential consequences of our actions, that vastly exceeds anything available to people in the past. This will enable an equally significant opportunity to expand our understanding of what it means to be moral, in the age of AI.

Bainbridge, L. (1983) 'Ironies of Automation', Automatica, 19/6: 775-79.

Parasuraman, R. and V. Riley (1997) 'Humans and Automation', Human Factors, 39/2: 230-53.

Vallor, S. (2015) 'Moral Deskilling and Upskilling in a New Machine Age', Philosophy & Technology, 28/1: 107-24.