GUEST EDITORIAL

Computing morality: Synthetic ethical decision making and behaviour

Abstract
We find ourselves at a unique point of time in history. Following over two millennia of debate amongst some of the greatest minds that ever existed about the nature of morality, the philosophy of ethics and the attributes of moral agency, and after all that time still not having reached consensus, we are coming to a point where artificial intelligence (AI) technology is enabling the creation of machines that will possess a convincing degree of moral competence. The existence of these machines will undoubtedly have an impact on this age old debate, but we believe that they will have a greater impact on society at large, as AI technology deepens its integration into the social fabric of our world. The purpose of this special issue on Computing Morality is to bring together different perspectives on this technology and its impact on society. The special issue contains four very different and inspiring contributions.

1 | DEVELOPING MORAL MACHINES

Modelling a Socialised Chatbot Using Trust Development in Children: Lessons learnt from Tay.
Authors: Oliver Bridge, Rebecca Raper, Nicola Strong and Selin Nugent.

In 2016, Microsoft’s Twitter-chatbot Tay.ai was shut down after only 24 h for making many racist and sexist comments by virtue of learning too much from Twitter trolls. One of the difficulties in designing ethical artificial intelligence (AI) systems is, as Bridge et al. put it, not in getting AI to follow rules, but rather in figuring out a priori, which rules it should follow. To circumvent both the obvious and unforeseen consequences of any top-down imposition of morality on AI, Bridge et al. propose an alternative solution informed by interdisciplinary social scientific accounts of the evolution and ontogeny of human morality. Put simply, the authors propose that rather than teaching AI what it means to be moral in a given context—a more flexible and potentially more effective solution is to build AI that is better at learning morality akin to how humans do—socially but selectively from others in their socio-cultural contexts.

Functionalist approaches to human morality consider how these cultural norms about what is right and what is wrong contribute to sustaining cooperation between individuals. In this sense, morality is not an end in and of itself, but rather one culturally evolved solution to the problems associated with regulating and coordinating human social behaviours. While evidence suggests that the cognitive foundations of morality are early emerging in humans, these foundational capacities do little to account for either the vast diversity in the forms of human morality or how they constellate in varied socio-ecological contexts. Thus, there has been longstanding interest in the social scientific study of morality in studying how individuals come to acquire the moral norms of their local social environments. In reviewing the evidence, Bridge et al. highlight how domain-general learning biases related to epistemic vigilance (i.e., selectively learning from only trustworthy sources) give rise to efficient learning of all sorts of cultural information, including moral norms. If one does as trustworthy others do and trustworthy others are good moral role models, then one can learn how to be moral without specifically being explicitly taught any moral rules. Thus, if you can teach AI to do as trustworthy and moral sources do, then morality can become an emergent quality of an AI’s behaviour rather than the result of adherence to a priori programmed rules. In their article, Bridge et al. provide a foundation for building this capacity for selective social learning into AI—with a particular focus on how a new variant of Tay (A1B0T) can learn to be more moral by the consequence of selectively paying more attention to reliable as opposed to unreliable sources of information. By doing so, the authors lay the foundations for the design of ethical AI grounded in the social sciences’ best empirical accounts of how humans flexibly learn moral information from their complex social environments.
The anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics
Authors: Nigel Crook and Joseph Corneli.

In this article, the authors present a high-level description of VirtuosA, a model of the cognitive components and processes that would enable a moral machine to learn virtue. The authors argue that, amongst the large variety of ethical systems that are available, virtue ethics is in some key respects are foundational, and should therefore be one of the primary approaches in developing the moral competence of machines. Despite this, there is very little work reported in the moral machine literature that is devoted to the adoption of virtue ethics-based approaches.

The VirtuosA model presented in this paper is informed by both theological and neuroscience perspectives of the cognitive components and processes that underlie the embodiment of virtue. The theological inspiration for this work is based on the model of the human self, proposed by Dallas Willard, a philosopher and Christian writer. He identifies what he calls the five dimensions of the human self, which are, in his view, essential to the development of the moral character (i.e., virtue) of a person. Willard provides functional descriptions of how these dimensions operate in the context of the whole person, which the authors have used to develop a functional model of the acquisition of virtue by a robot. The authors have also mapped these functional components onto equivalent structures in the human brain which have similar functions, enabling them to formulate a modular design of VirtuosA.

The paper looks specifically modelling kindness as an example of virtue, illustrating what this might look like in a simple gridworld example. The example outlines three stages of development of the robot: an initial mentoring phase in which the robot learns a specific task, an exploration phase in which the robot discovers new behaviours, including being unkind to another robot, and finally a development of virtue phase where the robot is shown via a decision about global warming based upon moral norms.

Defining moral values as those wherein some actions are regarded as 'good', and others as 'bad', Kenward and Sinclair begin by defining a moral machine as one that operates on values derived from human morality. As such, they continue, because machines are better than humans at making certain predictions, moral machines are apt to give moral advice to humans in situations where predictions are required, that is, in determining which criminals are most likely to reoffend. However, because it is not obvious whose morality should be applied to make these decisions, what is often the case is that moral consensus is used to determine what should be done. In turn, the difficulty in deciding whose morals to place into a machine has led to trying to embed the values of the users in the machine.

This, however, becomes problematic when we begin to assess what these values might look like. Where moral progress is defined as the change of values to alter participant behaviour that would largely reduce the risk of a catastrophic event taking place, or large-scale suffering, without imposing on human rights; the argument from the authors is that quite often the users are not aware that these morals need to be imposed. It is no good using moral norms to decide how a machine should behave, because ultimately, it is often the case that these morals would go against what the really moral thing is to be done.

The article finishes by highlighting this risk in relation to the environment, whereby urgent action needs to be taken, but only 36% inhabitants of the USA see global warming as a real issue. The problem is that if a machine were to make a decision about global warming based upon moral norms, the weighting would suggest that it is not a problem. An environmental disaster looms.

Questioning ‘what makes us human’: how audiences react to an AI-driven show
Authors: Rob Eagle, Rik Lander, and Phil D Hall

In this paper, the authors present a case study in human-computer collaboration in moral evaluation of AI agents through performance art. They present an observational study of the ‘I am Echoborg’ theatrical performance, which allows an audience to direct an exchange with a conversational AI. The show is enabled by a dynamic conversational AI, which interacts with individual audience members in an interview format through the support of a human actor who reads the text responses to the audience.

In designing an audience study, the authors aim to answer two questions: (1) In what ways can an encounter with AI change participant levels of knowledge of, or opinions, attitudes about, or potential posture towards automation and the application of intelligent machines in society? and (2) What role might the participatory nature of an encounter with AI
play in its effectiveness in informing or provoking thoughts or responses to automation/AI?

The authors measured overall audience attitudes on AI through two qualitative methods. First, they ask audience members to interactively answer questions on a sliding scale before and after the show using sticky notes on posters dispelling the questions in the performance space. Conversations between audience members discussing the questions were recorded. Second, they conducted semi-structured interviews to capture a range of responses not captured in the sliding scales.

The authors report the results of evaluating three performances, each producing a unique audience dynamic. Each show delivers different outcomes based on the composition and behaviours of the audience. The variation and unpredictability of human audience and conversational AI interaction between performances demonstrates the complexity of human group dynamics that shapes their perception of AI.

The case study of I am Echoborg reminds us that encoding morality in autonomous systems requires complementary understanding of the audience, or user perceptions of the systems. Furthermore, data on user perceptions must be sourced through participatory observation and experimentation that advances our understanding of human-technology relationships beyond shallow deductive binaries of good versus bad robots.

### 3 SUMMARY/CONCLUSION

This special issue has brought together four very different perspectives on some of the most pressing questions we face today about moral machines; questions that relate to whether it is possible to build such machines in the first place, and if so, how should we build them. The special issue also addresses challenging questions relating to the impact of this technology on individual people and society as a whole. The work presented here seeks to move the debate forward and focus it on key areas of concern. It offers both practical advice and guidance on the technology, and highlights the potential impacts of developing it. It is very clear that we need to proceed with caution, as we continue to build machines with moral competence.

**ORCID**

Nigel Crook https://orcid.org/0000-0002-0793-6616

**AUTHOR BIOGRAPHIES**

**Prof Nigel Crook** (Institute for Ethical AI, Oxford Brookes University) Prof Crook is the Associate Dean for Research and Knowledge Exchange and Professor of AI and Robotics at Oxford Brookes University. Nigel is also the Founding Director of the Institute for Ethical AI (ethical-ai.ac.uk), which seeks to develop ethical and trustworthy intelligent software solutions for business, organisations and society. Nigel graduated from Lancaster University with a BSc (hons) in Computing and Philosophy in 1982 and was awarded his PhD at Oxford Brookes (CNAA) in 1991. He has over 30 years of experience as a lecturer and a researcher in AI, during which he has contributed to a wide range of topics in Artificial Intelligence, including medical diagnostic systems, explainable AI, machine learning and social robotics. His most recent work has been focussed on the development of autonomous moral machines.

**Dr Selin Nugent** (Institute for Ethical AI, Oxford Brookes University) Dr Nugent is the Assistant Director (Social Science Research) in the Institute for Ethical AI. She is an anthropologist interested in human behavioural and cognitive modelling as it relates to cooperation and societal impacts of major technological transitions. Her work integrates computational methods and an evolutionary perspective to tackle questions on the social roles of AI systems. Her current work explores the impacts of machine agency and human-computer interaction on social interactions in the workplace. Selin holds a BA in Anthropology from Emory University (2012), an MA (2013) and PhD (2017) in Biological Anthropology from The Ohio State University. She previously held a position as a postdoctoral researcher at the University of Oxford’s Centre for the Study of Social Cohesion.

**Dr Matthias Rolf** (School of Engineering, Computing and Mathematics, Oxford Brookes University) Dr Rolf is a Senior Lecturer in AI and Mathematics in the School of Engineering Computing and Mathematics at Oxford Brookes University. He has published over 40 peer-reviewed articles
on AI and machine learning. His most significant, and award-winning research investigates infant-learning-inspired algorithms for robot learning. Matthias is currently working on the mathematical approaches for control of AI systems which are placed under multiple simultaneous constraints or objectives for accuracy and ethical behaviours; autonomous goal-system development; explorative learning of inverse models; autonomous systems creating novel goals and autonomous ethics based on social value systems, and audiovisual signal-level synchrony. Matthias previously held a position as a Specially Appointed Assistant Professor at Osaka University, Japan. He received his PhD, with highest honours, from Bielefeld University, Germany, in 2012.

**Dr. Adam Baimel** (Centre for Psychological Research, Oxford Brookes University). Dr. Baimel is an Early Career Research Fellow in the Centre for Psychological Research at Oxford Brookes University. He received his B.A (Psychology & Religious Studies), M.A. (Social Psychology), and PhD (Social Psychology and Quantitative Methods) from the University of British Columbia (Vancouver, Canada). Dr. Baimel studies the cultural evolution of religion with a particular interest in the psychological causes and behavioural consequences of changes in the moral concerns of religious systems.

**Rebecca Raper** (Institute for Ethical AI, Oxford Brookes University) is a Senior Consultant at The Institute for Ethical Artificial Intelligence and a final year PhD candidate at Oxford Brookes University. Her thesis is an examination and Reinforcement Learning implementation of machine ethics, bringing insights from philosophy and psychology to tackle a problem within computer science. Prior to embarking on her PhD, Rebecca completed a BA and an MA in philosophy, and a further postgraduate certificate in psychology. She also spent several years working as an analyst in banking and government.