20TH-22ND SEPTEMBER 2023
THE UNIVERSITY OF MANCHESTER

FORMALISING RESPONSIBILITY

BOOK OF ABSTRACTS

The Computational Agent Responsibility Project

TABLE OF CONTENTS

INTRODUCTION

The last decade has seen an intense increase in the usage of algorithmic decision-making in everyday life. As this technology is becoming more present in our lives, a seemingly never-ending storm of questions surrounding responsibility has naturally arisen. What are the different notions of responsibility and in what ways can these concepts be codified or made programmable? How does responsibility work in relation to task management in groups featuring multiple agents? Who is responsible for the outputs of autonomous decision-makers? And many more.

Formalising Responsibility is an inter-disciplinary workshop presented in collaboration between the University of Leeds and the University of Manchester. Over the course of three days, the workshop seeks to bring together philosophers and computer scientists in-person to discuss the notions of responsibility in relation to autonomous systems. The workshop comprises of 8 keynote speakers, a curated poster session, and a final round table discussion led by Prof. Helen Beebee and Prof. Michael Fisher.

The technological developments are blurring the lines between research in philosophy and computer science. As such, this workshop seeks to be a platform for innovative inter-disciplinary conversations and provide opportunity to take a bold leap forward for future inter-disciplinary work between philosophy and computer science.

ORGANIZATION

'Formalising Responsibility' is a workshop presented by the UKRI-funded ‘The Computational Agent Responsibility’ project. The project is a collaboration between the University of Leeds and the University of Manchester focusing on inter-disciplinary work in philosophy and computer science on the topic of responsibility of autonomous systems.

SCHEDULE

Wednesday - 20th of September 2023
Location: Uni Place, room 3.204, University of Manchester

11.00-12.00: Registration and Welcome

12.00-13.00: Al Mele (Florida State University):
"Responsibility: A Philosophical Toolkit, Some terminology, and a little philosophy"

13.00-14.00: Lunch

14.00-14.45: Daniela Vacek, née Glavaničová (Slovak Academy of Sciences)
"AI Control and Vicarious Responsibility"

14.45-15.15: Tea break

15.15-16.15: Pekka Mäkelä (University of Helsinki)
"Two Ways of Formalizing Responsibility"

Thursday - 21st of September 2023
Location: Kilburn, Atlas/Mercury Rooms University of Manchester

10.00-11.00: Hein Duijf (Utrecht University)
"Responsibility voids and collective obligations"

11.00-11.30: Tea break

11.30-12.15: Brian Logan (University of Aberdeen)
"Responsibility in Multi-Agent Systems"

12.30-13.30: Lunch

SCHEDULE

Thursday - 21st of September 2023 cont.
Location: Kilburn, Atlas/Mercury Rooms
University of Manchester

13.30-14.15: Emily Collins (University of Manchester)
"Relationships Mediating Trustworthy Human-Robot/AI Interaction: The Impact of Responsibility"

14.15-15.30: Poster Session and Tea break

15.30-16.30: Nick Schuster (Australian National University)
"Moral Expertise, Reasonably Pluralism, and Machine Ethics"

Friday - 22nd of September 2023
Location: Kilburn, Atlas/Mercury Rooms
University of Manchester

10.00-10.45: Virginia Dignum (Umeå University)
"Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour"

10.45-11.15: Tea break

11.15-12.15: Round-Table Workshop Discussion
led by Helen Beebee (University of Leeds) and
Michael Fischer (University of Manchester)

12.15: Goodbyes

ABSTRACTS

RESPONSIBILITY: A PHILOSOPHICAL TOOLKIT, SOME TERMINOLOGY, AND A LITTLE PHILOSOPHY

Al Mele
Florida State University

I asked myself how I might be able to help with the Computational Agent Responsibility project given that I know very little about engineered systems. I was in a similar position years ago regarding the bearing of various neuroscience studies on free will (though I knew more about neuroscience than I do about engineered systems). One thing neuroscientists in the group found useful were clearly stated definitions (or options for definitions) of various concepts: intention, decision, wanting, free will, etc. I will do something similar here with ability, intention, decision (for people), and intentional action. I will also discuss the "backward-looking" / "forward-looking" distinction in the philosophical literature on moral responsibility and three different conceptions of moral responsibility in that literature. I will close with some remarks on autonomy from a philosophical perspective.

AI CONTROL AND VICARIOUS RESPONSIBILITY

Daniela Vacek (née Glavaničová)
Slovak Academy of Sciences

Sven Nyholm recently formulated a new AI control problem and a corresponding dilemma. Nyholm undermines the central assumption of the traditional AI control problem, namely that controlling AI is always morally desirable. Nyholm argues that controlling advanced humanoid robots might be morally problematic; however, giving up control over such robots might well be unsafe, and thus not morally unproblematic either. This talk will accept the challenge. The talk will present a reformulation of the problem and the dilemma. A solution to these will be offered. It will draw some inspiration from the practices of vicarious responsibility and vicarious agency, and the idea of (in)appropriate control implicit in these practices. This suggestion in turn provides some insights on how responsibility can be ascribed in the context of AI.

TWO WAYS OF FORMALIZING RESPONSIBILITY

Pekka Mäkelä
University of Helsinki

The speed of progress in the development of automation, autonomously operating artificially intelligent systems, and social and industrial robotics is flabbergasting. Algorithms and robots functioning and making decisions in areas that used to be controlled by humans alone, for instance, in stock trading, medical diagnosing, and car driving are becoming ubiquitous. This development is mind-blowing and inspiring but also raising a lot of worries. One rather generic fear concerning increasingly autonomous systems has to do with responsibility. What happens to responsibility when technology is less and less in the control of human agents? Indeed, "responsibility" has become a catch word, which politicians, company representatives, and researchers frequently use to flag that they are sensitive to moral, social, and political risks that accompany the technological change and evolution. There is a good variety of notions of responsibility, and many debates and discussions might benefit from being a bit more exact about the sense of "responsibility" employed.

In this talk I will discuss whether "formalising responsibility" would be an answer to some of the worries concerning the societal risks brought about by autonomous machines. I will focus on moral and legal senses of responsibility. I will distinguish between two ways of understanding "formalization of responsibility": One that tracks the ideas discussed under the generic title responsibility of AI or responsibility of robots, and the other that tracks ideas discussed under the generic title of responsible AI or responsible robotics. At the core of the former is the idea that we could formalize responsibility in the sense of capturing moral responsibility into a computer program and by that way bring about a moral agent, say a robot, capable of bearing moral responsibility pretty much in the same sense as some human beings are considered to be morally responsible. This would provide us with a neat solution to the problem of responsibility gaps. I will critically evaluate the fruitfulness of this sense of formalizing responsibility, my critical argument builds in part on Alfred Mele's work on autonomous agency.

I end up arguing in favor of an institutional interpretation of "formalising responsibility" which tracks the ideas discussed under "responsible AI". Here I am thinking about social and institutional structures that can be identified at

least to an extent in terms of constitutive rules. Some such rules create social roles and positions which can be cashed out in terms of tasks, formal tasks if they are defined by codified rules. This provides us with a sense of both prospective and retrospective responsibility being formalized. I would claim that structural institutional responsibility allocation on the basis of formal rules is the most promising approach to the problems of moral and legal responsibility created by autonomous systems. This sense of formalising responsibility leads us to study and evaluate the responsibility of human beings either individually, jointly, or collectively. In this context I will briefly discuss regulation and hard and soft ethics. At the very end, if time allows, I will introduce a down to earth way of contributing to the implementation of this sense of formalizing responsibility by way raising the institutional sensitivity to moral reasons.

RESPONSIBILITY VOIDS AND COLLECTIVE OBLIGATIONS

Hein Duijf
Utrecht University

Societal challenges such as the widespread integration of black-box algorithms and climate change demonstrate that allocating moral responsibility is often difficult or impossible. In complex collective decision processes that involve several stages and multiple decision-makers, it will not always be clear who exactly contributes what and who can be held morally responsible for the final outcome. In the literatures on the ethics of technology and on corporate and group agency, the threat of responsibility voids or gaps is of central importance. A responsibility void obtains if a morally undesirable outcome or decision results from the interaction of several individuals even though none of these individuals can be held responsible for it.

In this talk, I will present recent and new research on responsibility voids and collective obligations. The basic underlying framework draws on deontic logic and game and decision theory. The main results indicate that facts about collective obligations cannot be explained by any set of conditions concerning individuals expressed in a well-established deontic logic of agency that models every combination of actions, omissions, abilities, and obligations of finitely many individual agents.

RESPONSIBILITY IN MULTI-AGENT SYSTEMS

Brian Logan
University of Aberdeen

When designing a multi-agent system, it is often necessary to specify which agents or groups of agents are (or can be) responsible for bringing about a particular state of affairs. Similarly, when analysing the operation of a multi-agent system, we may wish to determine which agents or groups of agents are responsible for the occurrence of a particular state of affairs, and the degree of responsibility of each agent for what occurred. In this talk, I will discuss how several notions of responsibility can be formalised in terms of the strategic abilities of agents, and how the degree of responsibility of individual agents can captured by their strategic power. Interestingly, it turns out that notions of forward-looking and backward-looking responsibility are equivalent in this framework. I discuss a number of examples from the literature, and argue that the strategic notion of responsibility allows a "natural" ascription of (degree of) responsibility to agents.

RELATIONSHIPS MEDIATING TRUSTWORTHY HUMAN-ROBOT/AI INTERACTION: THE IMPACT OF RESPONSIBILITY

Emily Collins
University of Manchester

Formalising responsibility in Robotic and AI (RAI) use, should consider, and perhaps even place as central, the humans using these technologies, and what they transparently understand are the consequences of such use in the short/long-term, asking: Who is responsible for the use of these technologies and what does their usage result in?

To formalise any of this in the broadest sense, an understanding of the dynamics between the human users is essential. Who are the users? Who are the employers of those users? Who deploys the technology? And what do these mediating relationships, and the trust between those branches, have to do with who is ultimately responsible for what happens when we use technology in real-world applied settings?

After a discussion of the increasing interest in considering, measuring, and implementing trust in Human-Robot Interaction (HRI), and relatedly

Human-AI Interaction, as it pertains to responsibility, I will propose that the dyadic model of HRI misses a key complexity which lies at the core of both trustworthy RAI systems, and how to approach formalising responsibility when it comes to RAI use: A robot's trustworthiness may be contingent on the user's relationship with, and opinion of, the individual or organisation deploying the robot.

I will discuss examples highlighting the need for trustworthy RAI in a variety of disparate environments. This will demonstrate that there is no one approach to the answer of trustworthy RAI, because a human's relationship with the person, employer, government, etc., who has given them RAI to work with is not consistent. Consequently, we should ask: Who is responsible when technology fails? The employer? The deployer? The user, instructed to work with the technology by said deployer? And ultimately, how does this affect how we formalise who is responsible for the outcomes of Human-Robot/AI use? What does a lack of answers here mean when it comes to building and deploying trustworthy RAI systems?

MORAL EXPERTISE, REASONABLY PLURALISM, AND MACHINE ETHICS

Nick Schuster
Australian National University

Value lock-in has emerged as a major concern for AI safety. It happens when AI systems are imbued with values that are oppressive to some of the people who stand to be affected by them. Value lock-in could be intentionally induced by authoritarian regimes, but it could also happen in a free society, where ideological minorities may find themselves oppressed by prevailing values which they reject. This makes reasonable pluralism—the fact that fellow members of society can and often do reasonably disagree with each other about substantive moral matters—a central problem for AI safety as well as machine ethics. Simply put, it's critical that AI systems resolve moral disagreements in ways that are themselves morally acceptable.

We begin this talk by exploring how certain current approaches to machine ethics—crowdsourcing moral judgements and/or training large language models to make moral judgements—might resolve moral disagreements by giving AI systems a kind of functional moral expertise. After discussing the merits of such systems, we then argue that they would have a

limited scope of application, since their outputs would be neither justified nor legitimate for populations where significant reasonable pluralism obtains. We therefore conclude that it's imperative to either find ways for AI systems to respect reasonable pluralism or else prohibit their use wherever they stand to impact diverse populations. Finally, we close by suggesting that meeting this immediate challenge for machine ethics would also stave off the authoritarian concerns of AI safety.

GOVERNANCE BY GLASS-BOX: IMPLEMENTING TRANSPARENT MORAL BOUNDS FOR AI BEHAVIOUR

Virginia Dignum
Umeå University

Artificial Intelligence (AI) applications are being used to predict and assess behaviour in multiple domains which directly affect human well-being. However, if AI is to improve people's lives, then people must be able to trust it, by being able to understand what the system is doing and why. Although transparency is often seen as the requirementin this case, realistically it might not always be possible, whereas the need to ensure that the system operates within set moral bounds remains.

In this paper, we present an approach to evaluate the moral bounds of an AI system based on the monitoring of its inputs and outputs. We place a 'Glass-Box' around the system by mapping moral values into explicit verifiable norms that constrain inputs and outputs, in such a way that if these remain within the box we can guarantee that the system adheres to the value. The focus on inputs and outputs allows for the verification and comparison of vastly different intelligent systems; from deep neural networks to agent-based systems.

The explicit transformation of abstract moral values into concrete norms brings great benefits interms of explainability; stakeholders know exactly how the system is interpreting and employing relevant abstract moral human values and calibrate their trust accordingly. Moreover, by operating at a higher level we can check the compliance of the system with different interpretations of the same value.

Full paper (with Andrea Aler, Andreas Theodorou and Frank Dignum):
https://arxiv.org/abs/1905.04994

POSTER PRESENTATIONS

The 'Formalising Responsibility' Workshop Poster Session takes place on Thursday the 21st of September in the Atlas/Mercury Lobby in the Kilburn building from 14.15-15.30.

1 ALGORITHMIC BIASES, DISCRIMINATION AND MORAL RESPONSIBILITY

Yuhan Fu
University of Sheffield

2 ARTIFICIAL INTELLIGENCE USE IN CLINICAL DECISION-MAKING: ALLOCATING LEGAL AND ETHICAL RESPONSIBILITY

Helen Smith
University of Bristol

3 MAKING ROBOTS RESPONSIBLE: NORM PSYCHOLOGY AND HUMAN-ROBOT RELATIONSHIPS

Stephen Setman
St. Bonaventure University

4 HOW TO GROUND RESPONSIBILITY ATTRIBUTIONS

Kristoffer Moody
University of Edinburgh

POSTER PRESENTATIONS

5 TRUST IN TECHNOLOGY IS A STANCE ABOUT TECHNOLOGISTS' RESPONSIBILITY WITH POWER

Chris McClean
University of Leeds

6 MORAL RESPONSIBILITY, AGENCY, AND AUTONOMOUS AI SYSTEMS

Mihaela Constantinescu
University of Bucharest

7 CRIMINAL RESPONSIBILITY OF AUTONOMOUS AI AGENTS: THE 'DETERRENCE TURN'

Elina Nerantzi
European University Institute

ANY QUESTIONS?

For more information about this book of abstracts, the workshop 'Formalising Responsibility' or in case of questions, please do not hesitate to reach out to the organisational team:

Sarah Moth-Lund Christensen:
s.m.l.christensen@leeds.ac.uk
Joe Collenette:
joe.collenette@manchester.ac.uk

20TH-22ND SEPTEMBER 2023
THE UNIVERSITY OF MANCHESTER

FORMALISING RESPONSIBILITY

This workshop and book of abstracts has been presented as part of the 'Computational Agent Responsibility' Project.

Organisational team:
Sarah Moth-Lund Christensen, University of Leeds
Joe Collenette, University of Manchester

A special thanks to:
The rest of the Computational Agent Responsibility' Team
The patient admin team at the University of Manchester
Our brilliant speakers and poster presenters

who have made this workshop possible

UNIVERSITY OF LEEDS

MANCHESTER 1824
The University of Manchester

UKRI Engineering and Physical Sciences Research Council