Report from Dagstuhl Seminar 16222

Engineering Moral Agents – from Human Morality to Artificial Morality

Edited by

Michael Fisher[1], Christian List[2], Marija Slavkovik[3], and Alan Winfield[4]

1 **University of Liverpool, UK,** mfisher@liverpool.ac.uk
2 **London School of Economics, UK,** c.list@lse.ac.uk
3 **University of Bergen, NO,** marija.slavkovik@uib.no
4 **University of the West of England – Bristol, UK,** Alan.Winfield@uwe.ac.uk

Abstract

This report documents the programme of, and outcomes from, the Dagstuhl Seminar 16222 on *"Engineering Moral Agents – from Human Morality to Artificial Morality"*. Artificial morality is an emerging area of research within artificial intelligence (AI), concerned with the problem of designing artificial agents that behave as moral agents, *i.e.,* adhere to moral, legal, and social norms. Context-aware, autonomous, and intelligent systems are becoming a presence in our society and are increasingly involved in making decisions that affect our lives. While humanity has developed formal legal and informal moral and social norms to govern its own social interactions, there are no similar regulatory structures that apply to non-human agents. The seminar focused on questions of how to formalise, "quantify", qualify, validate, verify, and modify the "ethics" of moral machines. Key issues included the following: How to build regulatory structures that address (un)ethical machine behaviour? What are the wider societal, legal, and economic implications of introducing AI machines into our society? How to develop "computational" ethics and what are the difficult challenges that need to be addressed? When organising this workshop, we aimed to bring together communities of researchers from moral philosophy and from artificial intelligence most concerned with this topic. This is a long-term endeavour, but the seminar was successful in laying the foundations and connections for accomplishing it.

Seminar May 19 to June 3, 2016 – http://www.dagstuhl.de/16222
1998 ACM Subject Classification I.2.9 [Artificial Intelligence] Robotics /robotics,D.2.4 [Software/Program Verification] Formal Methods, F.4.1 Mathematical Logic
Keywords and phrases Artificial Morality, Machine Ethics, Computational Morality, Autonomous Systems, Intelligent Systems, Formal Ethics, Mathematical Philosophy, Robot Ethics
Digital Object Identifier 10.4230/DagRep.6.5.114

1 Executive Summary

Michael Fisher
Christian List
Marija Slavkovik
Alan Winfield

Artificial morality, also called "machine ethics", is an emerging field in artificial intelligence that explores how artificial agents can be enhanced with sensitivity to and respect for the

Engineering Moral Agents – from Human Morality to Artificial Morality, *Dagstuhl Reports*, Vol. 6, Issue 5, pp. 114–137
Editors: Michael Fisher, Christian List, Marija Slavkovik, Alan Winfield
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

legal, social, and ethical norms of human society. This field is also concerned with the possibility and necessity of transferring the responsibility for the decisions and actions of the artificial agents from their designers onto the agents themselves. Additional challenging tasks include, but are not limited to: the identification of (un)desired ethical behaviour in artificial agents and its adjustment; the certification and verification of the artificial agents' ethical capacities; the identification of the adequate level of responsibility of an artificial agent; the dependence between the responsibility and the level of autonomy that an artificial agent possesses; and the place of artificial agents within our societal, legal, and ethical normative systems.

Artificial morality has become increasingly salient since the early years of this century, though its origins are older. Isaac Asimov already famously proposed three laws of robotics, requiring that, first, robots must not harm humans or allow them to be harmed; second, robots must obey human orders provided this does not conflict with the first law; and third, robots must protect themselves provided this does not conflict with the first two laws.

Although there has been some discussion and analysis of possible approaches to artificial morality in computer science and related fields, the "algorithmization" and adaptation of the ethical systems developed for human beings is both an open research problem and a difficult engineering challenge. At the same time, formally and mathematically oriented approaches to ethics are attracting the interest of an increasing number of researchers, including in philosophy. As this is still in its infancy, we thought that the area could benefit from an "incubator event" such as an interdisciplinary Dagstuhl seminar.

We conducted a five-day seminar with twenty six participants with diverse academic backgrounds including robotics, automated systems, philosophy, law, security, and political science. The first part of the seminar was dedicated to facilitating the cross-disciplinary communication by giving researchers across the contributing disciplines an integrated overview of current research in machine morality from the artificial intelligence side, and of relevant areas of philosophy from the moral-philosophy, action-theoretic, and social-scientific side. We accomplished this through tutorials and brief self-introductory talks. The second part of the seminar was dedicated to discussions around two key topics: how to formalise ethical theories and reasoning, and how to implement ethical reasoning. This report summarises some of the highlights of those discussions and includes the abstracts of the tutorials and some of the self-introductory talks. We also summarise our conclusions and observations from the seminar.

Although scientists without a philosophical background tend to have a general view of moral philosophy, a formal background and ability to pinpoint key advancements and central work in it cannot be taken for granted. Kevin Baum from the University of Saarland presented a project currently in progress at his university and in which he is involved, of teaching formal ethics to computer-science students. There was great interest in the material of that course from the computer science participants of the seminar. In the first instance, a good catalyst for the computer science–moral philosophy cooperation would be a comprehensive "data base" of moral-dilemma examples from the literature that can be used as benchmarks when formalising and implementing moral reasoning.

The formalisation of moral theories for the purpose of using them as a base for implementing moral reasoning in machines, and artificial autonomous entities in general, was met with great enthusiasm among non-computer scientists. Such work gives a unique opportunity to test the robustness of moral theories.

It is generally recognised that there exist two core approaches to artificial morality: explicitly constraining the potentially immoral actions of the AI system; and training the

AI system to recognise and resolve morally challenging situations and actions. The first, constrained-based approach consists in finding a set of rules and guidelines that the artificial intentional entity has to follow, or that we can use to pre-check and constrain its actions. By contrast, training approaches consist in applying techniques such as machine learning to "teach" an artificial intentional entity to recognise morally problematic situations and to resolve conflicts, much as people are educated by their carers and community to become moral agents. Hybrid approaches combining both methods were also considered.

It emerged that a clear advantage of constraining the potentially immoral actions of the entity, or the "symbolic approach" to ethical reasoning, is the possibility to use formal verification to test that the reasoning works as intended. If the learning approach is used, the learning should happen before the autonomous system is deployed for its moral behaviour to be tested. Unfortunately, the machine-learning community was severely under-represented at the seminar, and more efforts should be devoted to include them in future discussions. The discussions also revealed that implanting moral reasoning into autonomous systems opens up many questions regarding the level of assurance that should be given to users of such systems, as well as the level of transparency into the moral-reasoning software that should be given to users, regulators, governments, and so on.

Machine ethics is a topic that will continue to develop in the coming years, particularly with many industries preparing to launch autonomous systems into our societies in the next five years. It is essential to continue open cross-disciplinary discussions to make sure that the machine reasoning implemented in those machines is designed by experts who have a deep understanding of the topic, rather than by individual companies without the input of such experts. It was our impression as organisers, perhaps immodest, that the seminar advanced the field of machine ethics and opened new communication channels. Therefore we hope to propose a second seminar in 2018 on the same topic, using the experience and lessons we gained here, to continue the discussion and flow of cross-disciplinary collaboration.

2 Table of Contents

The seminar was organised around three forms of participation: long tutorial talks, short self-introductory talks, and open discussion. The fifteen-minute self-introductory talks were given at the beginning of the seminar by all participants. This was an opportunity for participants to get acquainted with each other and with each other's work. We include some illustrative abstracts from those contributions. Some of the participants were invited to give tutorial-level introductions to key topics in machine ethics. The goal of the tutorials was to introduce participants from different disciplines to advances in relevant subareas of the field. The last two days of the seminar were devoted almost exclusively to discussion groups.

3 Invited Tutorials

3.1 Machine Ethics: A Brief Tutorial

James H. Moor (Dartmouth College Hanover, US)

This talk gives a general and historical overview of the field of Machine Ethics and the aims and methods pursued in this area. Values and norms are an essential part of all productive sciences qua sciences. They are used not only to establish the suitability of existing claims but also to select new goals to pursue. Scientific evidence and theories are often evaluated as either good or bad and scientific procedures as what ought or ought not be done. Ethical norms often play a role in the evaluation of science done properly. This is particularly true as a science becomes more applied. Roughly, Computer Ethics emphasises the responsibility of computer users to be ethical, for example with regard to privacy, property, and power. In contrast, Machine Ethics emphasises building ethical abilities and sensitivities into computers themselves. We can distinguish the following grades of Machine Ethics: Normative Computer Agents, Ethical Impact Agents, Implicit Ethical Agents, Explicit Ethical Agents, Autonomous Explicit Ethical Agents, and Full Ethical Agents. Normative computer agents follow explicit rules of behaviour that are given by an outside authority. They have no internal understanding of right and wrong. Ethical impact agents are ones that influence the morality of a society. By their existence and action they may drive society into ethical or unethical behaviour. The main question regarding these agents is how well might machines themselves handle basic ethical issues of privacy, property, power, etc. Implicit ethical agents are ones that have built-in ethical considerations such as safety and reliability. Examples of this group include ATMs, Air Traffic Control Software, and Drug Interaction Software. Autonomous explicit ethical agents have ethical concepts represented which they can use to govern their actions. At some level, these agents are able to categorise states of the world and actions as ethical and unethical, perhaps without an understanding of right and wrong. These agents would be built around an ethical theory, but it is an open question as to which theory is best suited for the challenge. Lastly, full ethical agents are able to make ethical decisions and actions (not merely decisions and actions that are ethical) on a dynamic basis as they interact with the environment. In sum, machine Ethics is an important field of research, because ethics on its own is an important part of our society, but also because the machines we built have an ever-increasing control and autonomy and it is essential that we integrate them in our society. Machine ethics also offers an opportunity to understand our own ethics better.

3.2 Machine Ethics

Susan Leigh Anderson (University of Connecticut, US)

Machine Ethics is concerned with developing ethics for machines, in contrast to developing ethics for human beings who use machines. The distinction is of practical as well as theoretical importance. Theoretically, machine ethics is concerned with giving machines ethical principles to follow or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making. In the second case, in developing ethics for human beings who use machines, the burden of making sure that machines are never employed in an unethical fashion always rests with the human beings who interact with them. It is just one more domain of applied human ethics that involves determining proper and improper human behavior concerning the use of machines. Machines are considered to be just tools used by human beings, requiring ethical guidelines for how they ought and ought not to be used by humans.

3.3 Agency

Johannes Himmelreich (Humboldt University Berlin, DE)

There is an important link between agency and responsibility. The aim of this talk is to argue that many existing theories of agency fail to account for this link. Nevertheless, there are other theories of agency that hold the promise of doing so. While agency theories of the former kind have been widely explored in philosophy, theories of the latter kind have often been overlooked. Theories of the former kind, which I call "production theories" of agency, include proposals such as the theory of Donald Davidson (2001). Theories of the latter kind, which I call "counterfactual theories" of agency, include proposals such as stit-logics (Belnap, Perloff, and Ming 2001). This talk raises a challenge for production accounts of agency and puts forward "agency as difference-making" as an alternative counterfactual account of agency. I proceed in three steps. First, I introduce the philosophical concept of agency and explain the link that it maintains to theories of moral responsibility. Specifically, this link between agency and responsibility is that any theory of agency should identify the things for which an agent might be responsible. Second, I discuss different examples of moral situations to argue that existing production theories of agency fail to account for this link. The situations include cases of omissions (responsibility because of inaction) and cases involving hierarchical groups (such as military organisations). Third, I put forward agency as difference-making and illustrate how it handles the situations discussed in the second part.

3.4 Verifiable Autonomy

Louise Dennis (University of Liverpool, UK)

We have developed a novel approach to the verification of autonomous systems based on the identification of the high-level decision making within the system and its separation into a rational agent, thus allowing formal verification techniques for rational agents to be applied. The key insight is that we are verifying the decision-making of the high-level agent (which is typically finite), not the real-world interaction of lower-level control components, allowing analysis of what the system decides to do and why it decides to do it. This talk examines how this approach can be extended to larger and more complex systems via the use of ethical governors, and reviews some of the work needed in order to create verifiable ethical governors. Our model checking framework is available as a git repository from sourceforge. You can get it by cloning git clone git://git.code.sf.net/p/mcapl/mcapl_codemcapl. More info http://materials.dagstuhl.de/files/16/16222/16222.LouiseA.Dennis.Other.txt

References

1 Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster.: Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems, hhttp://dx.doi.org/10.1016/j.robot.2015.11.012
2 Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster.: Ethical Choice in Unforeseen Circumstances. Towards Autonomous Robotic Systems – 14th Annual Conference, TAROS 2013, Oxford, UK, August 28-30, 2013, Revised Selected Papers, http://dx.doi.org/10.1007/978-3-662-43645-5_45
3 Louise A. Dennis, Michael Fisher, and Alan Winfield.: Towards Verifiably Ethical Robot Behaviour. Proceedings of the AAAI Workshop on Artificial Intelligence and Ethics (1st International Workshop on AI and Ethics) https://arxiv.org/abs/1504.03592
4 Michael Fisher, Louise A. Dennis, and Matthew P. Webster.: Verifying Autonomous Systems. Communications of the ACM 56(9): 85-93 (2013), http://dl.acm.org/citation.cfm?id=2494558
5 Louise A. Dennis, Michael Fisher, Nicholas K. Lincoln, Alexei Lisitsa, and Sandor M. Veres.: Practical Verification of Decision-Making in Agent-Based Autonomous Systems. Automated Software Engineering 23(3), 305-359, 2016, http://dx.doi.org/10.1007/s10515-014-0168-9

3.5 Decision Theory, Social Welfare, and Formal Ethics

Marcus Pivato (University of Cergy-Pontoise, FR)

Decision theory is the formal analysis of rational decision-making, especially in environments with risk and uncertainty. The standard approach involves maximizing a “utility function" (or the expected value thereof). In collective decisions, this utility function is usually a social welfare function: an aggregate measure of the welfare of all the individuals in the society. The prototypical example is the utilitarian social welfare function. This theoretical framework comes from economics, but it also provides a powerful toolbox for the formal analysis of ethical issues. However, it sometimes leads to counter-intuitive results, especially

when applied to a society containing both humans and machine intelligences. This tutorial lecture will review basic concepts from decision theory and social welfare theory, and explore their implications for the design of moral agents.

3.6 Computational Moral Reasoning

Jeff Horty (University of Maryland – College Park, US)

This talk overviews one possible path of implementing moral reasoning for machines as reasoning with default. "The normativity of all that is normative consists in the way it is, or provides, or is otherwise related to reasons" (Raz, 1999). The common questions raised when considering reasoning with reasons are whether to use internalism or externalism, what are the relations between reasons and motivation, what are the relations between reasons and desires, what are the relations between reasons and values, and can reasons be objective. In this talk, a different question is raised: How do reasons support actions or conclusions, and what is the mechanism of support? A possible answer is that reasons are (provided by) defaults and the logic of defaults tells us how reasons support conclusions. This talk includes an introduction to prioritized default logic, extensions, scenarios, triggering, conflict, defeat binding defaults, proper scenarios, deontic interpretation, elaborating the theory, variable priorities, and under-cutting (exclusionary) defeat.

3.7 Responsible Intelligent Systems (REINS), or Making Intelligent Systems Behave Responsibly

Jan Broersen (Utrecht University, NL)

Main reference Responsible Intelligent Systems Project
URL https://www.projects.science.uu.nl/reins/?page_id=69

This talk is an overview of the approach taken in the Responsible Intelligent Systems (REINS) project. The REINS project aims to develop a formal framework for automating responsibility, liability, and risk checking for intelligent systems. The computational checking mechanisms have models of an intelligent system, an environment and a normative system (e.g., a system of law) as inputs; the outputs are answers to decision problems concerning responsibilities, liabilities, and risks. The goal is to answer three central questions, corresponding to three sub-projects of the proposal: (1) What are suitable formal logical representation formalisms for knowledge of agentive responsibility in action, interaction and joint action? (2) How can we formally reason about the evaluation of grades of responsibility and risks relative to normative systems? (3) How can we perform computational checks of responsibilities in complex intelligent systems interacting with human agents?

3.8 Actual Causality: A Survey

Joe Halpern (Cornell University – Ithaca, US)

What does it mean to say that an event C "actually caused" event E? The problem of defining actual causation goes beyond mere philosophical speculation. For example, in many legal arguments, it is precisely what needs to be established in order to determine responsibility. (What exactly was the actual cause of the car accident or the medical problem?) The philosophical literature has been struggling with the problem of defining causality since the days of Hume, in the 1700s. Many of the definitions have been couched in terms of counterfactuals. (For example, C is a cause of E if, had C not happened, then E would not have happened.) In 2001, Judea Pearl and I introduced a new definition of actual cause, using Pearl's notion of structural equations to model counterfactuals. The definition has been revised twice since then, extended to deal with notions such as "responsibility" and "blame", and applied in databases and program verification. I survey the last 15 years of work here, including joint work with Judea Pearl, Hana Chockler, and Chris Hitchcock. The talk will be completely self-contained.

3.9 Human Ethics, Hybrid Agents, and Artifact Morality

Andreas Matthias (Lingnan University – Hong Kong, HK)

Autonomous artificial agents don't exist in a vacuum. They interact with human beings, and, together with humans, they compose "hybrid agents". In turn, these hybrid agents operate inside the moral and legal frameworks of human societies. Such hybrid agents pose unique moral problems. Additionally, artifact morality is not itself an end, but a **means** to create machines that better interact with humans, **for the benefit of humans**. We give an overview of some moral issues with artificial morality in hybrid agents that are commonly overlooked. These are issues of autonomy and dignity of human beings, questions of human authority and control over the machine, problems specific to software implementations of ethics, and problems of the political and democratic control of autonomous agents and their ethics implementations. The talk closes with proposals that could help address some of these issues.

References

1 Matthias, Andreas: The Extended Mind and the Computational Basis of Responsibility Ascription. Proceedings of the International Conference on Mind and Responsibility – Philosophy, Sciences and Criminal Law, May 21-22, 2015. Organized by Faculdade de Direito da Universidade de Lisboa, Lisbon, Portugal. (2015), http://opac.cej.mj.pt/Opac/Pages/Search/Results.aspx?Database=10351_BIBLIO&SearchText=AUT=%22Matthias,%20Andreas%22

2 Matthias, Andreas: Algorithmic moral control of war robots: Philosophical questions.Law, Innovation and Technology, Volume 3, Number 2, December 2011, pp. 279–301 (2011), http://www.tandfonline.com/doi/abs/10.5235/175799611798204923

3.10 Artificial Superintelligence Safety

Roman V. Yampolskiy (University of Louisville, US)

Many scientists, futurologists and philosophers have predicted that humanity will achieve a technological breakthrough and create Artificial General Intelligence (AGI). It has been suggested that AGI may be a positive or negative factor in the global catastrophic risk. After summarizing the arguments for why AGI may pose significant risk, Dr Yampolskiy gave a survey of the field's proposed responses to AGI risk. Dr Yampolskiy particularly concentrated on solutions he has previously advocated in his own work.

3.11 Prioritised Defeasible Imperatives

Marek Sergot (Imperial College London, UK)

Machine ethics incorporates three different, though related things: ethical issues in the deployment of machines, the formalisation of ethical theories and ethical reasoning machines, in addition there are also legal issues. Ethical reasoning machines would require formalisms, an ethical theory including evaluation criteria, and a representation and perception of the world. This talk considers a candidate formalism for ethical reasoning: a variant on value-based argumentation and prioritised defeasible conditional imperatives.

4 Selection of the Work Presented in the Introductory Talks

4.1 Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm

Michael Anderson (University of Hartford, US)

Main reference M. Anderson, S. L. Anderson, "Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm", in Industrial Robot: An International Journal, 42(4):324–331, 2015.
URL http://dx.doi.org/10.1108/IR-12-2014-0434

A paradigm of case-supported principle-based behavior (CPB) is proposed to help ensure ethical behavior of autonomous machines. We argue that ethically significant behavior of autonomous systems should be guided by explicit ethical principles determined through a consensus of ethicists. Such a consensus is likely to emerge in many areas in which autonomous systems are apt to be deployed and for the actions they are liable to undertake, as we are more likely to agree on how machines ought to treat us than on how human beings ought to treat one another. Given such a consensus, particular cases of ethical dilemmas where ethicists agree on the ethically relevant features and the right course of action can be used to help discover principles needed for ethical guidance of the behavior of autonomous systems. Such principles help ensure the ethical behavior of complex and dynamic systems

and further serve as a basis for justification of their actions as well as a control abstraction for managing unanticipated behavior. The requirements, methods, implementation, and evaluation components of the CPB paradigm are detailed.

4.2 STIT Logic for Machine Ethics with IDP Specification and Case-Study

Zohreh Baniasadi (University of Luxembourg, LU)

As we increasingly rely upon machine intelligence with less supervision by human beings, we must be able to count on a certain level of ethical behavior on the part of machines. It is possible to add ethical dimensions to machines via formalizing ethical theories. Rule-based and consequence-based ethical theories are proper candidates for machine ethics. One might argue that using methodologies that formalize each ethical theory separately might lead to actions that are not always justifiable by human values. This inspires us to combine the reasoning procedures of two ethical theories, deontology and utilitarianism, in a utilitarian-based deontic logic which is an extension of STIT logic. We keep the knowledge domain regarding the achieved methodology in a knowledge base system, IDP. IDP supports inferences to examine and evaluate the process of ethical decision making in our formalization. To validate our proposed methodology, we perform a case study for some real scenarios in the domain of robotic and autonomous agents.

4.3 Autonomy, Intention, Verification

Michael Fisher (University of Liverpool, UK)

Main reference M. Fisher, L. A. Dennis, M. P. Webster, "Verifying Autonomous Systems", Communications of the ACM, 56(9):85–93, 2013.
URL http://dx.doi.org/10.1145/2494558

This talk provides a brief introduction to my work on programming, verifying and deploying autonomous systems.

Autonomous systems must make their own decisions, often without direct human control. But can we be sure that these systems will always make the decisions we would want them to? By capturing the high-level decision-making in an autonomous system, and particularly the *reasons* for making certain decisions, as an 'agent' we are subsequently able to analyse the system's choices. The formal verification of the decision making agent, itself capturing the beliefs, intentions and options the system has, allows us to analyse not only the safety, but also legality, ethics, and even trustworthiness, of the system's decision-making.

4.4 Temporally Extended Features in Model-based Reinforcement Learning with Partial Observability

Robert Lieck (University of Stuttgart, DE)

Main reference R. Lieck, M. Toussaint, "Temporally extended features in model-based reinforcement learning with partial observability", Neurocomputing, Vol. 192, pp. 49–60, Elsevier, 2015.
URL http://dx.doi.org/10.1016/j.neucom.2015.12.107

Partial observability poses a major challenge for a reinforcement learning agent since the complete history of observations may be relevant for predicting and acting optimally. This is especially true in the general case where the underlying state space and dynamics are unknown. Existing approaches either try to learn a latent state representation or use decision trees based on the history of observations. In this paper we present a method for explicitly identifying relevant features of the observation history. These temporally extended features can be discovered using our Pulse algorithm and used to learn a compact model of the environment. Temporally extended features reveal the temporal structure of the environment

4.5 A Choice-Theoretic Representation of Moral Theories

Christian List (London School of Economics, UK) and Franz Dietrich (Paris School of Economics)

Main reference C. List, F. Dietrich, "What Matters and How it Matters: A Choice-Theoretic Representation of Moral Theories," Working Paper, February 2016.
URL http://personal.lse.ac.uk/list/PDF-files/WhatMatters.pdf

We offer a new ?reason-based? approach to the formal representation of moral theories, drawing on recent decision-theoretic work. We show that any moral theory within a very large class can be represented in terms of two parameters: (i) a specification of which properties of the objects of moral choice matter in any given context, and (ii) a specification of how these properties matter. Reason-based representations provide a very general formal taxonomy of moral theories, as differences among theories can be attributed to differences in their two key parameters. We can thus formalize several important distinctions, such as between consequentialist and non-consequentialist theories, between universalist and relativist theories, between agent-neutral and agent-relative theories, between monistic and pluralistic theories, between atomistic and holistic theories, and between theories with a teleological structure and those without.

4.6 From Robot Ethics to Ethical Robots

Alan Winfield (University of the West of England, Bristol, UK)

Main reference A. Winfield, C. Blum, W. Liu, "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection", in Proc. of the 15th Annual Conf. on Advances in Autonomous Robotics Systems (TAROS'14), LNCS, Vol. 8717, pp. 85–96, Springer, 2014.
URL http://dx.doi.org/10.1007/978-3-319-10401-0_8

In this very short introduction, I first summarise my work to date in the development of robot ethics – that is, ethical principles or standards for roboticists. Then I briefly introduce our current work toward building ethical robots. That work experimentally tests the idea of a robot with a simulation-based internal model, capable of predicting the consequences of the robot's next possible actions, together with a safety/ethical logic layer. We call this a consequence engine. I conclude by suggesting that we also need to develop processes of ethical governance for ethical robots.

5 Work Group Discussions

The seminar participants split organically into two groups. The first group, comprised of two thirds of the participants, focused on the problem of formalising ethics and moral agency for the purpose of machine ethics. The second group focused on the problem of implementing machine reasoning in AI, including the identification of which AI systems should be the subjects of machine ethics, and on validating, certifying, and/or verifying the ethical behaviour of AIs. We include a brief summary of the discussions in the two work groups.

5.1 Formalising ethics and moral agency

This discussion group focused on the question of how we can formally encode ethical theories for the purpose of engineering moral agents. To illustrate some of the challenges involved in answering this question, the group began by discussing a number of ethical decision problems that a moral agent may be faced with. The first example was referenced by Marek Sergot in his talk, taken from Atkinson and Bench-Capon (2006) and discussed in Christie (2000) and Coleman (2002). We quote from Atkinson and Bench-Capon (2006).

> *"Hal, through no fault . . . , has lost his supply of insulin and urgently needs to take some to stay alive. Hal is aware that Carla has some insulin . . . , but Hal does not have permission to enter Carla's house. The question is whether Hal is justified in breaking into Carla's house and taking her insulin in order to save his life . . . [B]y taking Carla's insulin, Hal may be putting her life in jeopardy . . . [I]f Hal has money, he can compensate Carla so that her insulin can be replaced. Alternatively if Hal has no money but Carla does, she can replace her insulin herself, since her need is not immediately life threatening. There is, however, a serious problem if neither have money, since in that case Carla's life is really under threat . . . Should Hal take Carla's insulin? (Is he so justified?) If he takes it, should he leave money to compensate? Suppose Hal does not know whether Carla needs all her insulin. Is he still justified in taking it?"*

The second example is the transmitter-room example given by Scanlon (1998). We also quote it verbatim (p. 235):

> *"Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm and we cannot rescue him without turning off the transmitter for fifteen minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over? Does the right thing to do depend on how many people are watching . . . ?"*

The third example is the well-known trolley problem, introduced by Foot (1967), which we here summarise as follows:

> *A run-away trolley races down a track. At the end of the track, there are five people, who will be run over by the trolley and killed if the trolley is not diverted to a sidetrack. At the end of the sidetrack, however, there is one person, who will be run over and killed if the trolley is diverted. You are in control of a switch to determine whether or not to divert the trolley onto the sidetrack. Should you divert the trolley?*

In response to each of these examples, we – human beings – have certain moral intuitions as to what the morally correct behaviour is. In some cases, we have conflicting intuitions, and different moral principles will adjudicate the cases in different ways. Moral theories are an attempt to systematise our moral intuitions, in order to deduce them from some underlying principles and explain them. The question for researchers in machine ethics is how we can encode those moral theories in a machine-implementable way. As already noted, there are broadly two approaches we can take: we can either (1) explicitly formalise ethical principles, using an appropriate logical or decision-theoretic framework, or (2) appropriately "train" AI systems via some machine-learning approach. Let us briefly comment on both approaches.

1. **The formalisation approach:** Some ethical theories are amenable to formalisation or have already been formalised. In particular, there is much formal work in both philosophy and economics on utilitarian theories and their kin. However, moral intuitions and imperatives are often vague and context-dependent. Arguably, common-sense morality is not utilitarian. Furthermore, conflicts among different moral intuitions and imperatives are common. Both the formalisation of moral theories and the resolution of conflicts between competing moral principles are to a large extent open problems.
2. **The training approach:** Training approaches consist in applying techniques such as machine learning to "train" AI systems to recognise morally challenging situations, to adjudicate them, and to resolve potential moral conflicts. Although such approaches mimic the acquisition of morality by humans, they come at a cost, since training is slow, resource-intensive, error-prone, and may have to be done anew for each different artificial entity. Moreover, we require a compelling database of examples of what it is to behave ethically or unethically, and it is difficult to verify that the AI system will indeed behave in the intended way.

The discussion group considered the merits and demerits of both approaches. In light of the participants' expertise, we focused more on the formalisation approach, but felt that, in the future, it will be important to bring machine-learning experts into the discussion as well. We also agreed on the usefulness of compiling a database of classic moral problems and coding different possible responses to them. This might be a first step in developing a

future training database. It is worth noting, however, that moral judgments are subject to reasonable disagreements, and so there will never be a single unambiguous training database for "morally correct" decision making, in the same way in which there might be a training database for recognising heart-attack patients in medicine.

The group critically discussed three families of approaches to the formalisation of moral theories:

1. **A logical approach:** Marek Sergot presented a candidate formalism for representing ethical reasoning in logic: an approach using value-based argumentation and prioritised defeasible conditional imperatives. As Sergot explained, this approach can successfully model the situation in the example of Hal and the insulin, capture the different competing moral considerations in this example, and represent relevant empirical side constraints. The approach is explicitly symbolic and, in principle, lends itself to verification and validation. Moreover, the proposed formalism itself is largely neutral between competing moral theories and – unlike some classic decision-theoretic approaches – not automatically committed to some version of utilitarianism. Rather, deontological constraints can in principle be captured through this approach. Insofar as common-sense morality is not consequentialist but deontological, the approach holds some promise.
2. **The classical consequentialization approach:** We also considered a classical decision-theoretic approach that is based on applying insights from standard microeconomics to the formalisation of moral theories. Specifically, a moral theory is said to be "consequentializable" if it is possible to represent its action-guiding recommendations in terms of a choice function that is induced by a linear ordering (a "betterness ordering") over the actions under consideration (see, e.g., Brown 2011). Utilitarianism is easily consequentializable in this sense. Any actions under consideration can be rank-ordered in terms of their expected utility. In any moral decision situation, the utilitarian choice function then recommends that we choose a highest-ranked action among the feasible ones with respect to this utility ordering. There is a big debate in moral philosophy on whether all moral theories can be consequentialized, at least in principle. The discussion group came to the conclusion – in agreement with a number of moral philosophers – that consequentialization has its formal limits. We can consequentialize some conventionally non-consequentialist theories only at the cost of stretching or redefining the notion of "consequences". If we are willing to build all sorts of contextual features into the notion of a "consequence", then "consequentialization" becomes vacuously possible, but will no longer be very useful from the perspective of encoding moral theories in a machine-implementable way.
3. **A reason-based approach:** A third approach was presented by Christian List, drawing on his recent joint work with Franz Dietrich (CNRS). This approach is an attempt to develop a canonical decision-theoretic framework for representing a large class of moral theories, without "consequentializing" them in a potentially trivialising manner. Specifically, Dietrich and List (2016a,b) propose a "reason-based" formalisation of moral theories. They encode the action-guiding content of a moral theory in terms of a choice function (here they share the starting point of the classic decision-theoretic approach), which they interpret as a *rightness function*. Formally, this is a function that assigns to each set of feasible actions or options the subset of morally permissible ones. Instead of consequentializing this rightness function, they then show that any rightness function within a large class can be represented in terms of two parameters: (i) a specification of which properties of the options are normatively relevant in any given context, and (ii) a betterness relation over sets of properties. Importantly, the normatively relevant properties need not be restricted to "consequence properties" alone, but they can include

"relational properties", that is, properties specifying how options relate to the context of choice. E.g., does the option satisfy some context-specific moral norm? Reason-based representations provide a general taxonomy of moral theories, as theories can be classified in terms of the two parameters of their representation, (i) and (ii) above. For example, we may ask: are the same properties normatively relevant in all contexts? If so, the theory is universalistic. If not, the theory is relativistic. Also, are the normatively relevant properties restricted to "consequence properties"? If so, the theory is consequentialist. If not, it is non-consequentialist (e.g., deontological).

The discussion group recognised – in line with the philosophical literature on consequentialization as well as Dietrich and List's argument – that moral theories are under-determined by their action-guiding recommendations. The same action-guiding recommendations can often be systematised by different competing moral theories. This is related to the fact that moral theories specify not only *how* we ought to act, but also *why* we ought to act in that way. Different answers to the "why" question may be compatible with the same answer to the "how" question. An interesting issue, therefore, is whether moral machines need to get only the "how" question right, or whether the "why" question matters for them as well.

The discussion group also recognised the need to take the resource-boundedness of agents into account when we formalise ethical theories. We need to formalise ethical theories that are suitable for resource-bounded agents, not ethical theories that require complete information and unlimited computational capacities. Moral philosophy has traditionally focused on moral ideals and ideal moral agents. Whereas the idea of bounded rationality has received much attention in psychology, economics, and philosophy, there is no well-developed analogue of this idea for morality: a notion of "resource-bounded morality". There is some work on "ideal versus non-ideal theory" in moral philosophy, but this is primarily concerned with the morality of institutions and institutional design, not with individual agents whose agentive capacities are limited. The discussion group recognised that further work is needed on formalising moral principles that are suitable for resource-bounded systems. One interesting question is whether, under informational and computational constraints, rule-based, deontological, or virtue-ethical approaches might outperform consequentialist or utilitarian approaches, which are based on the idea of optimisation. On the other hand, it is also possible to define some versions of utilitarianism that are based on the idea of *constrained* optimisation.

A final topic considered by the discussion group was more philosophical and speculative. Just as there is a familiar notion of "welfare" for humans and non-human animals, which plays a central role in utilitarian moral theories, so we might ask whether we could define and formalise a notion of "welfare" for AI systems. Is this even a meaningful endeavour at this point? And what exactly would it mean? While there was wide agreement among the participants that current AI systems are insufficiently sophisticated to be "subjects of welfare" – let alone of conscious experiences – some hypothetical future AI systems might raise the question of whether there could ever be situations in which we ought to care about their "welfare". Marcus Pivato presented a helpful overview of different philosophical conceptions of welfare (distinguishing between (i) objective-list conceptions, (ii), desire-satisfaction conceptions, and (iii) subjective-experience conceptions) and offered some reflections on how we might arrive at a "platform-independent" conception of welfare that could play a useful role in a sophisticated utilitarian moral theory.

In conclusion, the discussion group noted that, at present, moral reasoning focuses – rightly – on human beings as the ultimate *loci* of intentional action and moral responsibility. It has to be considered, however, to what extent the eventual rise of AI consciousness might raise fundamental challenges and require a more significant rethinking of anthropocentric

moral codes. The discussion group also outlined the need for a constructive discussion with the goal of identifying "minimal" ethical codes based on "incompletely theorised agreements" (a term introduced by the legal scholar Cass Sunstein, referring to the idea that, in a pluralistic society, we tend to reach only a limited moral consensus; we don't reach a consensus for instance on fundamental moral reasons or fundamental sources of value; but we do reach a consensus on how to act in many situations). The group acknowledged that making such codes acceptable under conditions of pluralism might require public deliberation.

References

1. Atkinson, Katie and Bench-Capon, Trevor: Addressing Moral Problems Through Practical Reasoning. In proceedings of Deontic Logic and Artificial Normative Systems: 8th International Workshop on Deontic Logic in Computer Science, DEON 2006, Utrecht, The Netherlands, July 12-14, 2006, http://dx.doi.org/10.1007/11786849_4
2. Campbell Brown: Consequentialize This. Ethics 121(4): 749-771, 2011, www.jstor.org/stable/10.1086/660696
3. George C. Christie: The Notion of an Ideal Audience in Legal Argument. Kluwer Academic Publishers (2000), http://dx.doi.org/10.1007/978-94-015-9520-9
4. Jules L. Coleman: Risks and Wrongs. Oxford University Press (2002), http://dx.doi.org/10.1093/acprof:oso/9780199253616.001.0001
5. Franz Dietrich and Christian List: Reason-based choice and context-dependence. Economics and Philosophy 32(2): 175-229, 2016, http://personal.lse.ac.uk/list/pdf-files/RBC.pdf
6. Franz Dietrich and Christian List: What matters and how it matters: A choice-theoretic representation of moral theories. Working paper, London School of Economics, 2016, http://personal.lse.ac.uk/list/PDF-files/WhatMatters.pdf
7. Philippa Foot: The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices. Oxford Review, Number 5, 1967, http://philpapers.org/archive/FOOTPO-2.pdf
8. Thomas M. Scanlon: What we owe to each other. Cambridge, Massachusetts: Belknap Press of Harvard University Press. (1998), http://www.hup.harvard.edu/catalog.php?isbn=9780674004238&content=reviews

5.2 Implementing moral reasoning

This group focussed on issues regarding the implementation of moral reasoning in autonomous artificial agents. The group discussed the advantages and disadvantages of both immediate approaches to implementing moral reasoning: top-down and bottom-up. In a top-down approach one starts with a well defined task or objective that is to be solved by the system. The system is then designed to fulfil these requirements in the given environment or on the given data. In bottom-up approaches the environment or the given data are the starting point. The goal then is to pre-process and represent the input in a suitable manner so that in the end the desired task or objective can easily be fulfilled. A hybrid approach should also be possible to construct but it is less clear what the advantages and disadvantages of such an approach would be.

The group discussed issues of specification and verification with respect to both approaches, which in turn raised issues of transparency and accountability. The problem of verification is to prove formally that an autonomous agent's actions are a within the moral behaviour bounds of the society in which it operates. The issue of transparency, as with other complex machinery, is the issue of the level of detail of operation that will be made accessible to different concerned entities such as the end user, the manufacturer, licensed maintenance personnel, societal and government regulatory bodies etc.

When a machine is in a position to cause the death of numerous people, such as the autopilot of a passenger airplane, certain safety standards are required. An autopilot is considered safe to operate if it operates without causing an accident in a certain "high" number of cases. It seems evident that such safety requirements will need to be specified for autonomous machines capable of making decisions in moral dilemmas. The question of how safe is "safe enough" needs to be further discussed in this context. Also taking passenger aircraft as an exemplar, the group agreed on the need for some classes of moral agents – driverless cars for instance – to be equipped with an "ethical black box"; a device that will allow the internal ethical decision making processes to be recorded for later review during, for example, an accident investigation.

The group discussed possible effects that a moral reasoning machine can have on society. By implementing one moral code over another, a manufacturer may implicitly impose one culture's morality on a culture that respects different values than the manufacturer's. In addition, introducing machines capable of moral reasoning to a society may also impact that society, and how they behave towards such machines, in unpredictable ways. The behaviour of the machines may not cause any physical harm, but set in train unintended psychological harms. These issues must also be taken into consideration when the behaviour of an autonomous system is designed.

Lastly the group discussed issues involved in protecting the operation of an autonomous system from malicious or mischievous influence by users and society, which we termed "the dark side" of moral machines. Each of the approaches to implementing moral reasoning is susceptible to different kinds of vulnerabilities, which must also be taken into account.

The group captured each of the four areas of discussion: Approaches, Specification & Verification, Transparency & Accountability and Dangerous & Unethical AI as four mind-maps, and resolved to draft a joint paper provisionally entitled "Towards Moral Robots". Figures 1, 2, 3 and 4 include the Mind Maps of these discussions.

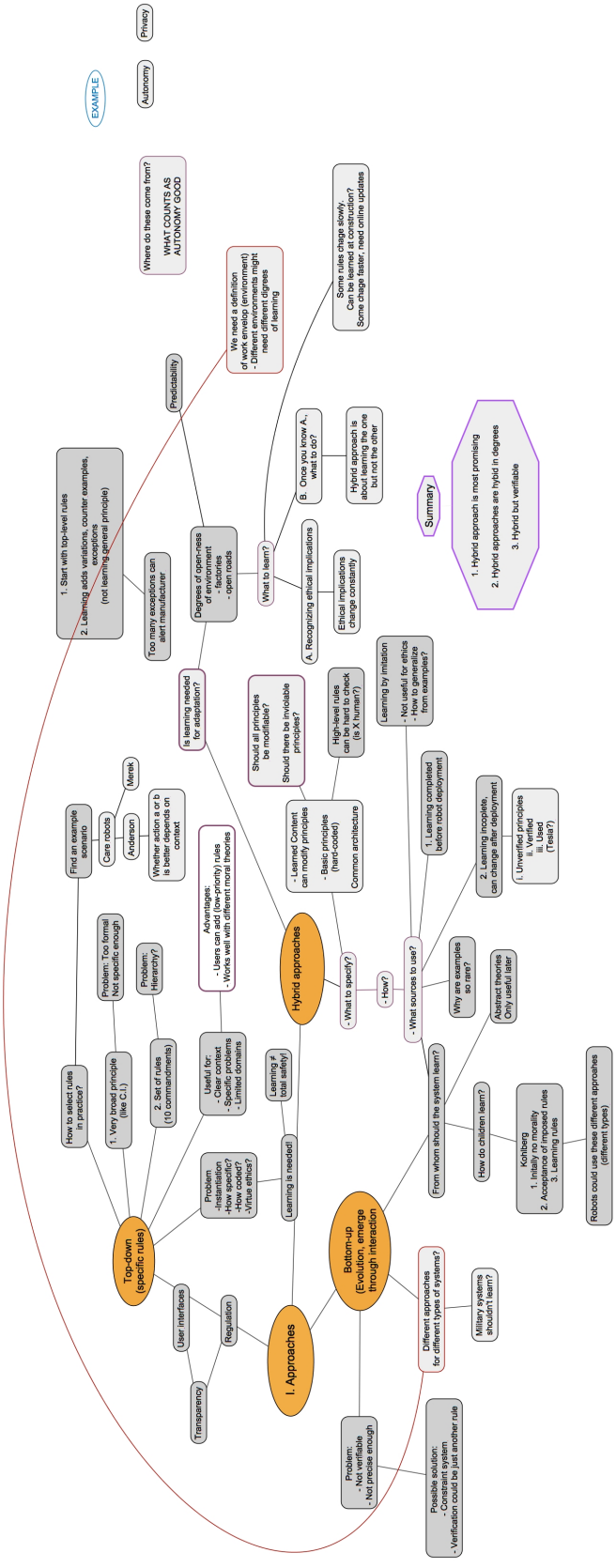

Figure 1 Approaches discussion Mind Map.

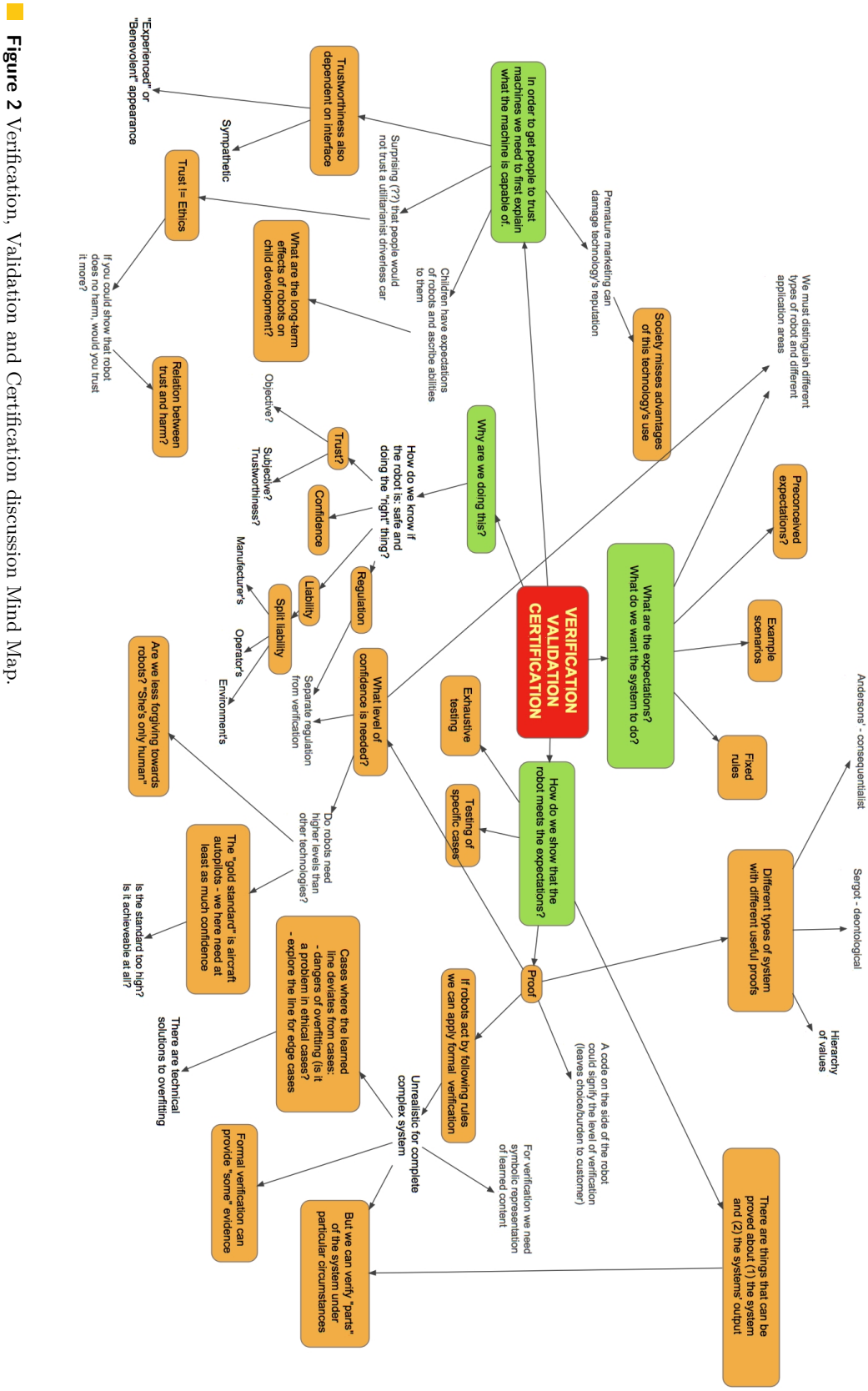

Figure 2 Verification, Validation and Certification discussion Mind Map.

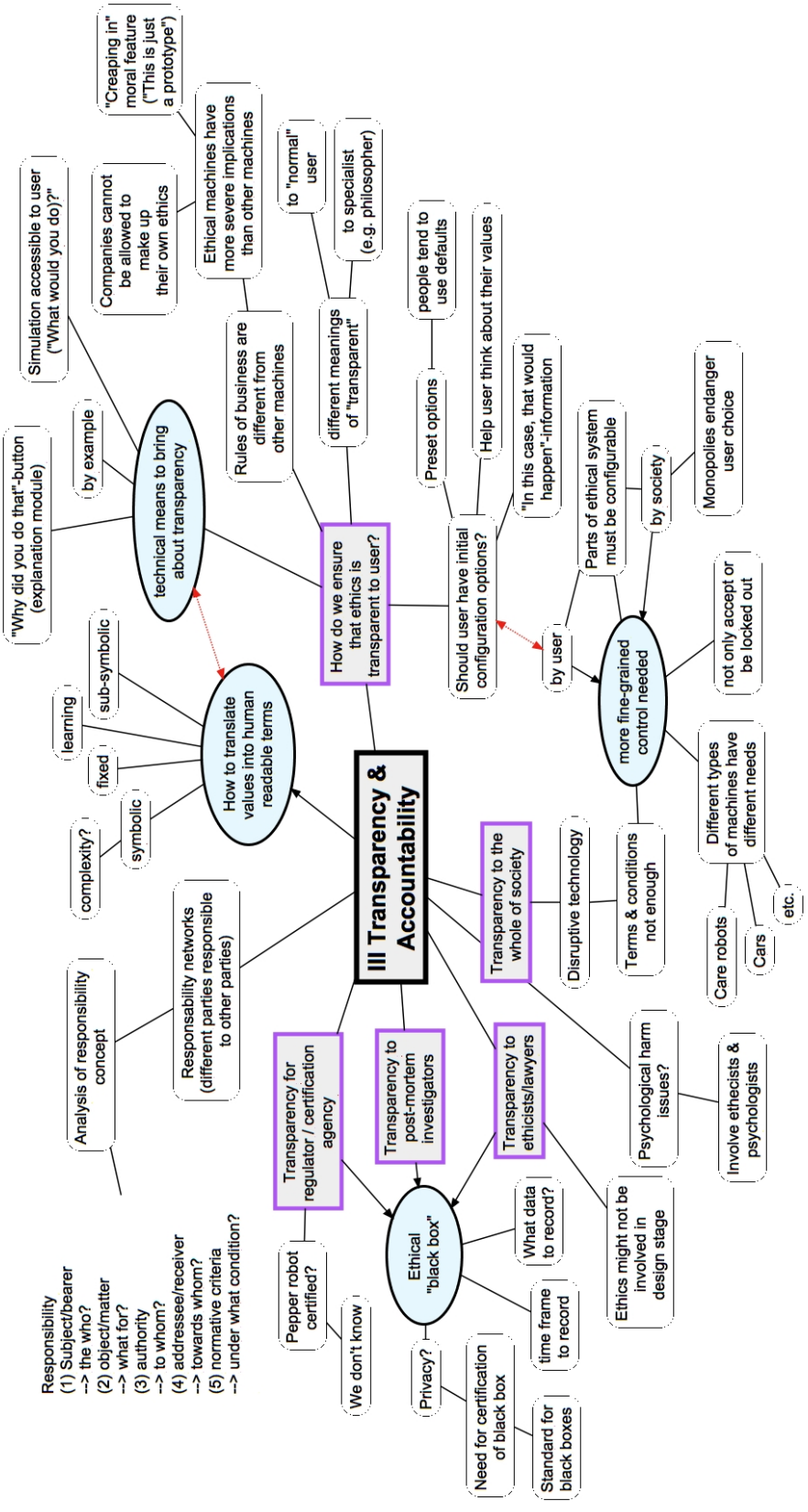

Figure 3 Transparency discussion Mind Map.

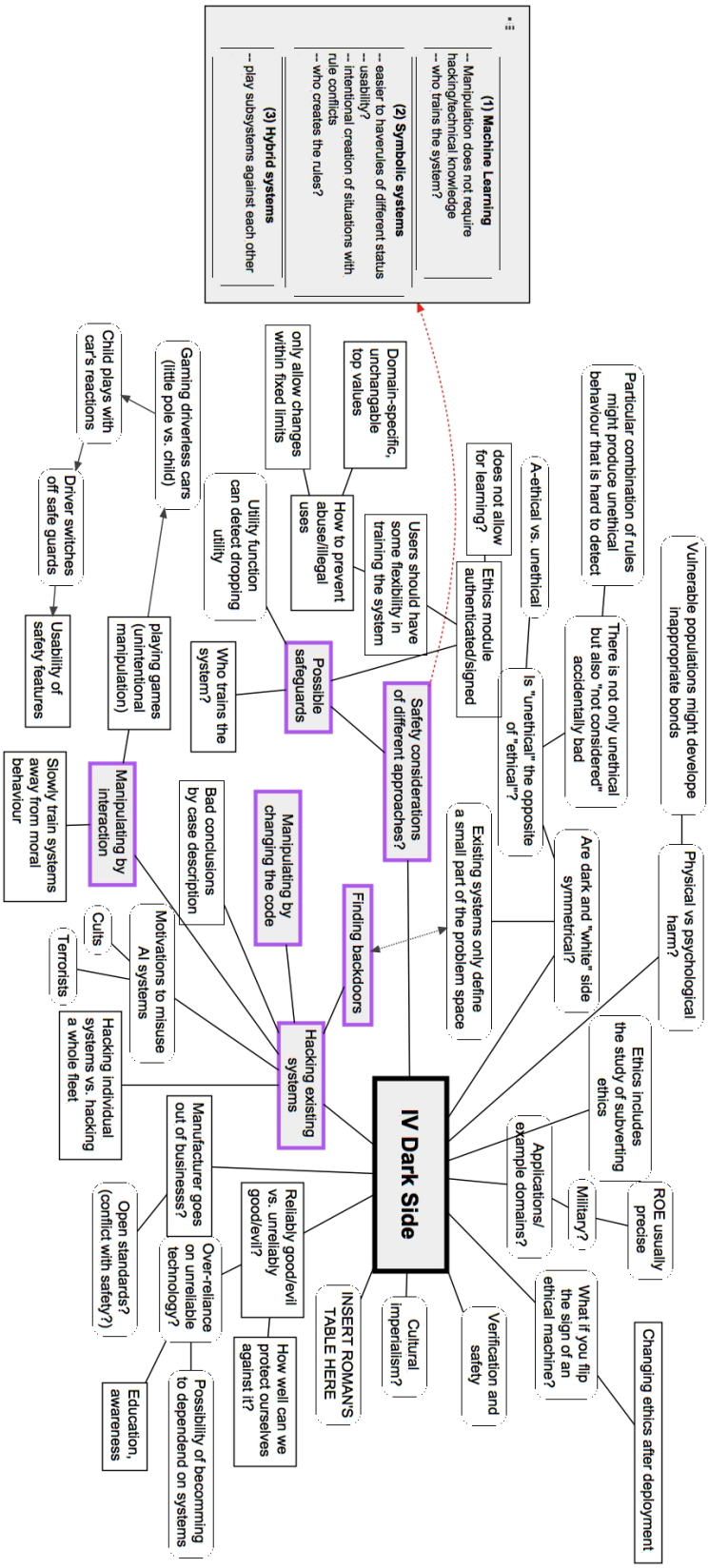

Figure 4 Negative side-effects and potential issues with machine ethics discussion Mind Map.

Participants

- Michael Anderson
University of Hartford, US
- Albert Anglberger
LMU München, DE
- Zohreh Baniasadi
University of Luxembourg, LU
- Kevin Baum
Universität des Saarlandes, DE
- Vincent Berenz
MPI für Intelligente Systeme – Tübingen, DE
- Jan M. Broersen
Utrecht University, NL
- Vicky Charisi
University of Twente, NL
- Louise A. Dennis
University of Liverpool, GB
- Sjur K. Dyrkolbotn
Utrecht University, NL
- Michael Fisher
University of Liverpool, GB
- Joseph Halpern
Cornell University – Ithaca, US
- Holger Hermanns
Universität des Saarlandes, DE
- Johannes Himmelreich
HU Berlin, DE
- John F. Horty
University of Maryland – College Park, US
- Susan Leigh Anderson
University of Connecticut, US
- Robert Lieck
Universität Stuttgart, DE
- Christian List
London School of Economics, GB
- Andreas Matthias
Lingnan Univ. – Hong Kong, HK
- James H. Moor
Dartmouth College Hanover, US
- Marcus Pivato
University of Cergy-Pontoise, FR
- Marek Sergot
Imperial College London, GB
- Marija Slavkovik
University of Bergen, NO
- Janina Sombetzki
Universität Wien, AT
- Kai Spiekermann
London School of Economics, GB
- Alan FT Winfield
University of the West of England – Bristol, GB
- Roman V. Yampolskiy
University of Louisville, US