Report from Dagstuhl Seminar 19171

Ethics and Trust: Principles, Verification and Validation

Edited by

Michael Fisher[1], Christian List[2], Marija Slavkovik[3], and Astrid Weiss[4]

1 **University of Liverpool, GB,** `mfisher@liverpool.ac.uk`
2 **London School of Economics, GB,** `c.list@lse.ac.uk`
3 **University of Bergen, NO,** `marija.slavkovik@uib.no`
4 **TU Wien, AT,** `astrid.weiss@tuwien.ac.at`

Abstract

This report documents the programme of, and outcomes from, the Dagstuhl Seminar 19171 on *"Ethics and Trust: Principles, Verification and Validation"*. We consider the issues of ethics and trust as crucial to the future acceptance and use of autonomous systems. The development of new classes of autonomous systems, such as medical robots, "driver-less" cars, and assistive care robots has opened up questions on how we can integrate truly autonomous systems into our society. Once a system is truly autonomous, i.e. learning from interactions, moving and manipulating the world we are living in, and making decisions by itself, we must be certain that it will act in a safe and ethical way, i.e. that it will be able to distinguish 'right' from 'wrong' and make the decisions we would expect of it. In order for society to accept these new machines, we must also trust them, i.e. we must believe that they are reliable and that they are trying to assist us, especially when engaged in close human-robot interaction. The seminar focused on questions of how does trust with autonomous machines evolve, how to build a 'practical' *ethical* and *trustworthy* system, and what are the societal implications. Key issues included: Change of trust and trust repair, AI systems as decision makers, complex system of norms and algorithmic bias, and potential discrepancies between expectations and capabilities of autonomous machines. This workshop was a follow-up to the 2016 Dagstuhl Seminar 16222 on *Engineering Moral Agents: From Human Morality to Artificial Morality.* When organizing this workshop we aimed to bring together communities of researchers from moral philosophy and from artificial intelligence and extend it with researchers from (social) robotics and human-robot interaction research.

Seminar April 22–26, 2019 – http://www.dagstuhl.de/19171

2012 ACM Subject Classification Computer systems organization → Robotic autonomy, Hardware → Functional verification, Human-centered computing → HCI theory, concepts and models

Keywords and phrases Verification, Artificial Morality, Social Robotics, Machine Ethics, Autonomous Systems, Explain-able AI, Safety, Trust, Mathematical Philosophy, Robot Ethics, Human-Robot Interaction

Digital Object Identifier 10.4230/DagRep.9.4.59

Ethics and Trust: Principles, Verification and Validation, *Dagstuhl Reports*, Vol. 9, Issue 4, pp. 59–86
Editors: Michael Fisher, Christian List, Marija Slavkovik, and Astrid Weiss
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Astrid Weiss (TU Wien, AT)
Michael Fisher (University of Liverpool, GB)
Christian List (London School of Economics, GB)
Marija Slavkovik (University of Bergen, NO)

Academics, engineers, and the public at large, are all wary of *autonomous systems*, particularly robots, drones, "driver-less" cars, etc. Robots will share our physical space, and so how will this change us? With the predictions of roboticists in hand, we can paint portraits of how these technical advances will lead to new experiences and how these experiences may change the ways we function in society. Two key issues are dominant once robot technologies have advanced further and yielded new ways in which we and robots share the world: (1) will robots behave *ethically*, i.e. as we would want them to, and (2) can we *trust* them to act to our benefit. It is more these barriers concerning ethics and trust than any engineering issues that are holding back the widespread development and use of autonomous systems. One of the hardest challenges in robotics is to reliably determine desirable and undesirable behaviours for robots. We are currently undergoing another technology-led transformation in our society driven by the outsourcing of decisions to intelligent, and increasingly autonomous, systems. These systems may be software or embodied units that share our environment. The decisions they make have a direct impact on our lives. With this power to make decisions comes the responsibility for the impact of these decisions – legal, ethical and personal. But how can we ensure that these artificial decision-makers can be *trusted* to make safe and *ethical* decisions, especially as the responsibility placed on them increases?

The related previous Dagstuhl Seminar 16222 on *Engineering Moral agents: From human morality to artificial morality* in 2016, highlighted further important areas to be explored, specifically:

- the extension of 'ethics' to also address issues of 'trust';
- the practical problems of implementing ethical and trustworthy autonomous machines;
- the new verification and validation techniques that will be required to assess these dimensions.

Thus, we thought that the area would benefit from a follow-up seminar which broadens up the scope to Human-Robot Interaction (HRI) and (social) robotics research.

We conducted a four-day seminar (1 day shorter than usual due to Easter) with 35 participants with diverse academic backgrounds including AI, philosophy, social epistemology, Human-Robot Interaction, (social) robotics, logic, linguistics, political science, and computer science. The first day of the seminar was dedicated to seven invited 20-minute talks which served as tutorials. Given the highly interdisciplinary nature of the seminar, the participants from one discipline needed to be quickly brought up to speed with the state of the art in the discipline not their own. Moreover, the goal of these tutorials was to help develop a common language among researchers in the seminar. After these tutorials we gave all participants the chance to introduce their seminar-related research in 5-minute contributed talks. These talks served as a concise way to present oneself and introduce topics for discussion.

Based on these inputs four topics were derived and further explored in working groups through the rest of the seminar: (1) Change of trust, including challenges and methods to foster and repair trust; (2) Towards artificial moral agency; (3) How do we build practical systems involving ethics and trust? (2 sub-groups) (4) The broader context of trust in HRI:

Discrepancy between expectations and capabilities of autonomous machines. This report summarizes some of the highlights of those discussions and includes abstracts of the tutorials and some of the contributed talks. Ethical and trustworthy autonomous systems are a topic that will continue to be important in the coming years. We consider it essential to continue these cross-disciplinary efforts, above all as the seminar revealed that the “interactional perspective” of the “human-in-the-loop” is so far underrepresented in the discussions and that also broadening the scope to STS (Science and Technology Studies) and sociology of technology scholars would be relevant.

2 Table of Contents

Executive Summary

Overview of Tutorials

Overview of Contributed Talks

Working groups

3 Overview of Tutorials

3.1 Tutorial: Robot Ethics – Towards Trustworthy AI agents

Raja Chatila (Sorbonne University – Paris, FR)

Joint work of Contributions of authors of Ethically Aligned Design (IEEE) and members of the EU High-Level Expert Group on AI.
URL https://ethicsinaction.ieee.org,
https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

A computational intelligent system, a robot, is a set of algorithms designed by humans, using data (big/small/sensed) to solve [more or less] complex problems in [more or less] complex situations. The system might include the capability of improving its performance based on data classification (e.g., deep learning) or on evaluating previous decisions (e.g., reinforcement learning).

Such systems could be regarded as "autonomous" in a given domain and for given tasks as long as they are capable of accomplishing their tasks despite environment changes within this domain (this is close to the notion of robustness). Autonomy is related to the complexity of the domain and of the task.

Computerized technical systems, especially those used in critical applications, must be trustworthy to reliably deliver the expected correct service. The academic and industrial communities developing software-based systems have produced several techniques to achieve their dependability or resilience. Software validation and verification techniques, such as error detection and recovery mechanisms, model checking, detection of incorrect or incomplete system knowledge, and resilience to unexpected changes due to environment or system dynamics, have been developed and used.

However, as decisions usually devoted to humans are being more and more delegated to machines, sometimes running computational algorithms based on learning techniques using data, operating in complex and evolving environments, new issues have to be considered.

First, can such systems make ethical decisions? The answer is negative. Ethical discernement is not a mere computational process. Second, should the AI "black-box" justify moving away from procedures that guarantee a trusted operation of the system? This is both an ethical and a technical question to the designers. Key features such as transparency, explainability and accountability become of prime importance. What technical and non-technical new measures should be taken then in the design process and in the governance of these systems?

A summary of the IEEE global Initiative on Ethics of Autonomous and Intelligent Systems, as well as the Ethics Guidelines for Trustworthy AI of the European High-Level Expert Group on AI shed light on these issues.

References

1 *Ethically Aligned Design 1st Edition.* https://ethicsinaction.ieee.org, March 2019.

2 High-Level Expert Group on AI. *Ethics Guidelines for Trustworthy AI.* https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence, April, 2019.

3.2 Tutorial: Formalizing Ethical Choice

Franz Dietrich (Paris School of Economics & CNRS, FR)

Our reason-based formalism for rational choice (which Christian List and I are developing) can be used to represent moral theories. Almost any plausible moral theory can indeed 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. This yields a very general taxonomy of moral theories, in which we can formally distinguish 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 and without a teleological structure. (based on joint work with Christian List)

References

1 Franz Dietrich and Christian List. What matters and how it matters: a choice-theoretic representation of moral theories. *Philosophical Review*, 126(4):421–479, 2017.

3.3 Tutorial: Social Robots – To Be Trusted?

Marc Hanheide (University of Lincoln, GB)

This talk aims to provide a (quite shallow) overview into the domain of social robots. It collates a number of (sometimes controversial) definitions and their criticism, as well as offering links to challenges and open debates grounded in practical experience. I consider that (i) robots are not treated as “social equals”, (ii) social robots are often about social evocation (or deception?) and that (iii) “sociabilty” can serve as a means to build “better robots”.

References

1 Dautenhahn, K. Encyclopedia of Human-Computer Interaction. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction

2 Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. Robotics and Autonomous Systems, 42(3–4), 143–166.

3 Breazeal, C. (2004). Social interactions in HRI: the robot view. Systems, Man and Cybernetics, Part C, IEEE Transactions On, 34(2), 181–186. https://doi.org/10.1109/TSMCC.2004.826268

4 Mathur, M. B., and Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. Cognition, 146, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

5 Seibt, Johanna. (2016). “Integrative Social Robotics” – A New Method Paradigm to Solve the Description Problem And the Regulation Problem? de Graaf, M. M. A. (2016). An Ethical Evaluation of Human–Robot Relationships. International Journal of Social Robotics. https://doi.org/10.1007/s12369-016-0368-5

6 Hegel, F. (2012). Effects of a Robot's Aesthetic Design on the attribution of social capabilities. In Proceedings – IEEE International Workshop on Robot and Human Interactive Communication. https://doi.org/10.1109/ROMAN.2012.6343796

3.4 Tutorial: Trust in Human-Robot Interaction

James E. Young (University of Manitoba – Winnipeg, CA)

The field of Human-Robot Interaction proposes that, in many ways, people treat and respond to robots as life-like things. Drawing from this, social robotics is the broad study of how robots themselves can be seen as social actors, which leads to the investigation of robots using human-like social interaction techniques to work with people. Following, if we consider robots as social actors, then issues of trust arise. The human-robot interaction community has broadly explored trust and related concepts. This includes human trust in robot informers (e.g., kiosks), including for vulnerable populations such as children. As part of this, the community has mapped out robot and interaction design strategies for managing trust (e.g., increasing or decreasing) in a range of situations, leading to work in persuasion, and even obedience to robots. With all of this in perspective, I raise the question of whether we, as a society, should accept the idea that machines without emotional or moral regulating systems, with perfect memories and algorithmic accuracy, can use human language to impact people.

3.5 Tutorial: Trust in Robots is Multi-Dimensional Too

Bertram F. Malle (Brown University – Providence, US)

Different definitions, theories, and measurements of trust are distributed over multiple literatures, and it is unclear how they can all be integrated. I suggest that there is not one correct definition of trust but people have a multidimensional conception of trust. On the one hand, they can experience capacity trust, which breaks into perceptions of the agent in question as capable to a certain degree and as reliable to a certain degree; on the other hand, they can experience moral trust, which breaks into perceptions of the agent as sincere to a certain degree and as ethical to a certain degree. I offer empirical evidence for these conceptual distinctions in people's lay understanding of trust and introduce a new measurement instrument for assessing these multiple dimensions. Finally, I draw implications for the role of trust in human-robot interaction, including how one would conduct verification tests for moral trust and how one could better assess calibrated trust within a multidimensional framework.

3.6 Tutorial: Two Kinds of Trust in Robots

Andreas Matthias (Lingnan University – Hong Kong, HK)

We can distinguish two different kinds of trust: trust in a process and trust in value alignment. Only process-trust can be achieved through certification of robots. Values-trust requires an individual, personal alignment of values between the user and the robot that is the basis for the robot to treat the user as a Kantian "end". Is it doubtful whether values-trust can be achieved within the framework of existing economical structures in the technology sector.

3.7 Tutorial: Machine Ethics – Philosophical Approaches

Thomas Michael Powers (University of Delaware – Newark, US)

After reviewing some formalizable ethical theories, I argue for a particular minimal conception of machine ethics as a starting point—a "coded ethics". Coded ethics begins with conventional, accepted moral rules that apply to specific contexts, and implements them in engineered systems in response to a growth in morally-relevant capabilities of the system. The central feature of a coded ethics is the implementation of ethical rules when any new capability of the system threatens moral values (privacy, safety, etc.). There is no artificial consciousness or intentionality required for coded ethics; the machine follows accepted normative reasoning–it makes moral decisions–that protect or promote the interests of humans and other moral patients. A coded ethics might implement any number of basic deontological prescriptions and proscriptions, depending on the given context: rules against privacy violations, non-combatant harm, and non-compensated costs, etc. The proposed coded ethics is an elaboration of Adaptive Incremental Machine Ethics found in Powers (2011), Incremental Machine Ethics, IEEE Robotics and Automation 18:1.

4 Overview of Contributed Talks

4.1 How Is This Fair? Formalising Contextual Adherence to Moral Values

Andrea Aler Tubella (University of Umeå, SE)

Joint work of Andrea Aler Tubella, Virginia Dignum, Andreas Theodorou, Frank Dignum

In this short presentation, I introduce ongoing research on the formalisation of contextual adherence to moral values. If AI is to be deployed safely, then people need to understand how the system is interpreting and whether it is adhering to the relevant moral values. Even though transparency is often seen as the requirement in this case, realistically it might not always be possible or desirable, whereas the need to ensure that the system operates within set moral bounds remains. 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 contextual 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(s) in a specific context.

4.2 Being Responsible for Someone Else’s Actions

Jan M. Broersen (Utrecht University, NL)

Human agents are responsible for their own actions. And in so far AIs are mere tools, humans are also responsible for the actions of the AIs they employ. However, in a future where AIs are ubiquitous, things will not be so clean cut. First of all there is the possibility that AIs will become so advanced that some would want to attribute agency to them of the kind that comes with the responsibility that humans have. Second, there is the way in which actions of AIs express the agency of many different humans involved in the deployment or design of an AI. One central theme that pervades these issues is how one agent can be responsible for another agent’s actions (the second agent maybe being an AI or a cooperating human). The formal logic study of the transitivity of the responsibility relation has not been taken up yet in any serious way.

4.3 Four Papers in the Philosophy of Technology

Einar Duenger Bøhn (University of Agder, NO)

I work on four papers. One where I defend what I call informationalism, which is the view that reality is most fundamentally pure information. Second, a paper called AlphaMoral, where I develop the idea that artificial morality can be developed through board games. Third, a paper called The Moral Turing test, where I defend the moral Turing test as a good test for artificial morality. Fourth, popular pieces where I argue that smartphones should have an age limit.

4.4 What Can We Prove About Ethical Reasoning Systems?

Louise A. Dennis (University of Liverpool, GB)

Joint work of Louise A. Dennis, Michael Fisher, Alan Winfield, Martin Mose Bentzen, Felix Lindner, Paul Bremner, Matt Webster, Marija Slavkovik

I discussed work on the verification of ethical reasoning systems, specifically the properties that could be verified. I characterised these systems as one where some explicit encoding of ethics was given to a decision system. These systems were then verified using model-checking. I tentatively categorised the properties into those that verified the implementation of the decision process (properties of the form "the most ethical choice is always made according to the ethical theory used to make the decision"); "sanity checking" properties of the encoding of the ethics (for instance that a house is always evacuated in the case of a fire); and checking of specific scenarios which might also allow inclusion of probabilistic evaluation of outcomes.

4.5 Verification for Robotics and Autonmous Systems

Clare Dixon (University of Liverpool, GB)

Main reference Matt Webster, David Western, Dejanira Araiza-Illan, Clare Dixon, Kerstin Eder, Michael Fisher, Anthony G. Pipe: "An Assurance-based Approach to Verification and Validation of Human-Robot Teams", CoRR, Vol. abs/1608.07403, 2016.
URL http://arxiv.org/abs/1608.07403

In this introductory talk I explained my background and interests related to the workshop themes. In particular I discussed the EPSRC funded project Trustworthy Robot Assistants a joint project between the Universities of Liverpool, Hertfordshire and Bristol Robotics Lab. We considered two use cases, a domestic robot assistant and collaborative manufacture and three verification and validation (V&V) methods. The V&V methods were formal verification, simulation based testing and real robot experiments. We believe that using these methods to inform and update the inputs to the other methods leads to improved V&V for systems. More details can be found at www.robosafe.org.

I also mentioned interests relating to formal verification for swarm robotics and the development of calculi and provers for temporal and agent logics and their application to problems. Publications can be found at http://cgi.csc.liv.ac.uk/∼clare/.

4.6 Learning Rules for Ethical Machines

Abeer Dyoub (University of L'Aquila, IT)

Codes of ethics are abstract rules. These rules are often quite difficult to apply. Abstract principles such as these contain open textured terms that cover a wide range of specific situations. These codes are subject to interpretations and might have different meanings in different contexts. There is an implementation problem from the computational point of view with most of these codes, they lack clear procedures for implementation. In this work we present a new approach based on Answer Set Programming and Inductive logic Programming for monitoring the employees behavior w.r.t. ethical violations of their company's codes of ethics.

4.7 Perspicuous Computing

Holger Hermanns (Universität des Saarlandes, DE)

Joint work of Christel Baier, Raimund Dachselt, Vera Demberg, Bernd Finkbeiner, Holger Hermanns, Jörg Hoffmann, Markus Krötzsch, Rupak Majumdar
URL https://www.perspicuous-computing.science/

From autonomous vehicles to smart homes and cities – increasingly computer programs participate in actions and decisions that affect humans. However, our understanding of how these applications interact and what are the causes of a specific automated decision cascade is lagging far behind. It is nowadays virtually impossible to provide scientifically well-founded answers to questions about the exact reasons that lead to a particular decision, let alone about accountability in case of the malfunctioning of, say, an exhaust aftertreatment system in a modern car. The root of the problem is that contemporary systems do not have any built-in concepts to explicate their behaviour. They calculate and propagate outcomes of computations, but are not designed to provide explanations. They are not perspicuous.

This talk highlights the need for establishing a science of perspicuous computing as the key to enable comprehension in a cyber-physical world. And it surveys focused activities that are currently being ramped up as part of the DFG-funded Transregional Collaborative Research Centre 248 – CPEC.

4.8 Crowd-Sourcing Tests – Can This Increase Public Trust?

Kerstin Eder (University of Bristol, GB)

Joint work of Kerstin Eder, Dejanira Araiza-Illan, Tony Pipe
Main reference Matt Webster, David Western, Dejanira Araiza-Illan, Clare Dixon, Kerstin Eder, Michael Fisher, Anthony G. Pipe: "An Assurance-based Approach to Verification and Validation of Human-Robot Teams", CoRR, Vol. abs/1608.07403, 2016.
URL https://arxiv.org/abs/1608.07403
Main reference Dejanira Araiza-Illan, Anthony G. Pipe, Kerstin Eder: "Intelligent Agent-Based Stimulation for Testing Robotic Software in Human-Robot Interactions", in Proc. of the 3rd Workshop on Model-Driven Robot Software Engineering, MORSE@RoboCup 2016, Leipzig, Germany, July 1, 2016, pp. 9–16, ACM, 2016.
URL https://doi.org/10.1145/3022099.3022101

While *trust* is subjective – it can be gained and lost, re-gained and lost again over time – the *trustworthiness* of a system should be demonstrable. Because no single technique is adequate to cover a whole system in practice [1], at the Trustworthy Systems Laboratory in Bristol (http://www.bristol.ac.uk/tsl) we are working on a variety of complementing techniques to enable system designers and robotics engineers to gain confidence in the correctness of the robotic and autonomous systems they develop. These techniques include, but are not limited to:

- design techniques – systems that are simple by design are also understandable;
- analysis techniques that enable transparency – systems that provide an insight into how they make decisions, why they act in a certain way or how they use resources become understandable;
- verification and validation techniques – rigorous proof complemented by simulation-based testing and real-world testing can provide convincing evidence of a system's trustworthiness.

Systems that use Artificial Intelligence (AI) are a particular challenge when it comes to demonstrating their trustworthiness. Nevertheless, to make robots and autonomous systems truly useful, they have to be both powerful and smart. To achieve the latter, AI techniques, Machine Learning in particular, are what we rely on, with research actively exploring the use of these techniques in safety-critical applications such as autonomous driving.

An important research question we are exploring in this context is how we can exploit the power of AI in verification [2]. In addition, we are currently developing a game-based application that aims to crowd-source test cases for autonomous driving. This opens up interesting opportunities for research and also offers a platform for public engagement where players can gain confidence in the behaviour of autonomous vehicles in a simulated environment.

References

1 Webster M, Western D, Araiza-Illan D, Dixon C, Eder K, Fisher M, Pipe AG. A corroborative approach to verification and validation of human–robot teams. *CoRR* 2016; abs/1608.07403.

2 Araiza-Illan D, Pipe AG, Eder K. Intelligent agent-based stimulation for testing robotic software in human-robot interactions. *Proceedings of the 3rd Workshop on Model-Driven Robot Software Engineering*, MORSE '16, ACM, 2016; 9–16.

4.9 Transparency: For Interaction or for Societal Discourse?

Kerstin Fischer (University of Southern Denmark – Sonderborg, DK)

URL https://portal.findresearcher.sdu.dk/en/publications/when-transparent-does-not-mean-explainable

In this talk, I discuss some problems with transparency about robot capabilities: First, signaling what a robot can and cannot do is far from trivial due to the way social signals work. Second, in order to achieve social interaction with robots, i.e. in order to share social spaces with robots, robots need to use (and understand) social signals – which are often shortcuts to rich meanings and invite inferences to many further capabilities. If robots don't use these signals, they will be very tiresome to use; if they use them, they contribute to the illusion of life-like beings with more capabilities than they actually have, which is desired in the case of social robots, but which may hinder societal discourse about robots in society.

4.10 Hybrid Ethical Reasoning in HERA

Felix Lindner (Universität Freiburg, DE)

Joint work of Martin Mose Bentzen, Robert Mattmueller, Bernhard Nebel
Main reference Felix Lindner, Martin Mose Bentzen, Bernhard Nebel: "The HERA approach to morally competent robots", in Proc. of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017, pp. 6991–6997, IEEE, 2017.
URL https://doi.org/10.1109/IROS.2017.8206625

Hybrid Ethical Reasoning Agents (HERA) are capable of computing permissibility judgments under various ethical principles. The talk briefly gives an introduction to the technical aspects of HERA, presents a generalization to the case of judging action sequences rather than individual actions, and shows how explanations of permissibility judgments can be computed. Finally, Immanuel is presented–a robot that implements HERA and which can have moral discussions with humans.

4.11 Enabling People Who Design Machines That Influence People

AJung Moon (Open Roboethics Institute – Vancouver, CA)

From recommender systems to interactive robots, many autonomous intelligent systems we design and deploy today hold the promise to address some of the world's toughest problems. They have also been the source of social, ethical, and legal issues on a global scale. Open Roboethics Institute conducted a series of studies that demonstrate multiple approaches to incorporating human values into machines that influence people's decisions and behaviours. This includes the discovery of what factors affect our design decisions that have moral implications, and analysis of organizational values to create value-alignment in the design and operational decisions pertaining to the autonomous intelligent machines.

4.12 Robot Wrongs and Robot Rights: What Can Economic Theory Tell Us?

Marcus Pivato (University of Cergy-Pontoise, FR)

There are three kinds of morally relevant interactions with artificial intelligences ("robot"):

(1) Robots can act on humans
(2) Robots can interact with other robots
(3) Humans can act on robots

Class (1) raises the question: How to design robots that behave ethically? This leads to the question: what is ethical? Social choice theory and social welfare theory provide a mathematical framework for specifying and analysing consequentialist ethical theories.

Class (2) suggests treating robot-robot interactions using game theory. Indeed, game theory might be more suitable for robots than for humans, because robots can be programmed to be perfectly "rational" (in the economic sense of the word) and commit to strategies which lead to socially efficient equilibria.

Class (3) raises the question: Can a robot be a moral patient? On the plausible premise that moral patiency depends on "consciousness" or "sentience", this raises the question: Could a robot ever be conscious or sentient? If we someday design robots that are conscious or sentient, then we will need to develop a theory of "sentient agent well-being" which applies to both humans and robots.

4.13 Ethics and Trust in Sociotechnical Systems

Munindar P. Singh (North Carolina State University – Raleigh, US)

I make the case for a sociotechnical systems perspective on ethics and trust. Specifically, I advocate moving away from the current emphasis on individual decision making about ethical dilemmas to how we might design (micro)societies in which humans and agents coexist. Concerns of justice and norms are essential in developing computational formalizations of sociotechnical systems, both to evaluate such systems as designers and for agents to function in them as members.

4.14 What Does It Mean to Trust a Robot?

Kai Spiekermann (London School of Economics, GB)

We experience all kinds of different "trust talk":

- "I trust my car to get us to Trier."
- "I trust the IT guy to set up my network credentials correctly."
- "I trust my neighbour to water the plants while I'm away."
- "I trust my friends [to do what friends do]."

Some trust talk is about reliability (car, IT guy). In the reliability sense, we trust because we think the system is well built, the person well trained, incentivized, etc. But in the deeper sense we trust because we expect the trusted agent to be committed. The difference can be seen when considering adequate responses to failure: If we have reliability-trust and experience failure, we tend to feel disappointed. By contrast, if we have commitment-trust and experience failure, we tend to feel betrayed. We can draw a further distinction by looking at how precise and explicit the expected actions are codified. On the face of it, in the context of moral machines the idea of reliability trust for highly specified behaviour is most immediately applicable, but the most challenging issues arise in relation to commitment trust and in relation to expected actions that are not precisely codified.

4.15 From Values to Support

Myrthe Tielman (TU Delft, NL)

I am interested in personal assistive technology, systems which support us in daily life activities. In order to enable such systems to understand our motivations better, we propose to use value-based reasoning. Through linking values to actions we gain insight into what to support people with. Ideally, we will also be able to use values to reason about support itself, as well as being able to use them when explaining the systems behavior back to the user.

4.16 Machine Ethics: Test, Proof or Trust?

Suzanne Tolmeijer (Universität Zürich, CH)

In this talk, I introduced work in progress on a survey paper for the field of machine ethics. A classification framework is introduced with three dimensions: purely ethical, purely technical, and the overlapping implementation category. Some interesting findings are presented, including that half of the selected papers do not present a proper evaluation for their ethical machine.

4.17 Designing Normative Theories of Ethical Reasoning: Formal Framework, Methodology, and Tool Support

Leon van der Torre (University of Luxembourg, LU)

Joint work of Christoph Benzmüller, Xavier Parent, Leendert W. N. van der Torre
Main reference Christoph Benzmüller, Xavier Parent, Leendert W. N. van der Torre: "Designing Normative Theories of Ethical Reasoning: Formal Framework, Methodology, and Tool Support", CoRR, Vol. abs/1903.10187, 2019.
URL http://arxiv.org/abs/1903.10187

The area of formal ethics is experiencing a shift from a unique or standard approach to normative reasoning, as exemplified by so-called standard deontic logic, to a variety of application-specific theories. However, the adequate handling of normative concepts such as obligation, permission, prohibition, and moral commitment is challenging, as illustrated by the notorious paradoxes of deontic logic. In this article we introduce an approach to design and evaluate theories of normative reasoning. In particular, we present a formal framework based on higher-order logic, a design methodology, and we discuss tool support. Moreover, we illustrate the approach using an example of an implementation, we demonstrate different ways of using it, and we discuss how the design of normative theories is now made accessible to non-specialist users and developers.

5 Working groups

5.1 Change of Trust – Challenges and Methods to Foster and Repair Trust

Myrthe Tielman (TU Delft, NL), Clare Dixon (University of Liverpool, GB), Marc Hanheide (University of Lincoln, GB), Felix Lindner (Universität Freiburg, DE), Suzanne Tolmeijer (Universität Zürich, CH), Astrid Weiss (TU Wien, AT)

The major discussion topic of this group was: *Change of trust – Challenges and methods to foster and repair trust.* In the first breakout session four main topics of interest for further discussion were identified: (1) identification of failure and generation of explanation for failures; (2) likability vs. trustworthiness; (3) empowerment and putting users in control; and (4) trustworthiness of humans (from a robot perspective).

(1) Regarding failures and explanations, it was discussed that firstly different types of failures need to be distinguished: (1) misunderstood/unexpected behaviour (unexpected communicational effect) and (2) actual system failures (crashes) or wrongly taken actions (e.g. the robot being stuck). In both cases, it was agreed that *explanations* for the end user are key to restore trust. Subsequently, the generation of explanations was discussed. Methods, such as plan-based explanations related to previous decisions were suggested, but questions came up about the correct level of detail of abstractions and human-comprehensible explanations. It was agreed that explanations to end users however do not necessarily need to be in natural language, but can use cues such as closed eyes, blinking lights, nodding head etc. Overall, the aim of explanations should be to increase transparency and understandability in order to repair trust in a failure situation. Other relevant aspects with respect to failures and explanations that were discussed were that repetition should be avoided and reduced. In long runs, robots must not do the same mistakes again. It rather must form a model of the individual user's beliefs (beliefs of beliefs). In general individualisation was also considered key for maintaining trust in HRI. However, one of the big challenges is to understand/recognize when and where users' expectations are violated. The idea came up if a classification of failures and their risk impact for trust (potentially even with a mitigation) could be developed. This idea was later followed up in the subsequent breakout sessions and a preliminary *Failure Taxonomy* was developed. It is planned to elaborate this further as a publication for the 2020 HRI conference.

(2) With respect to likability vs. trustworthiness, the discussion revealed that so far most of the HRI research focuses on the fact that transparency-through-explanation increases the trustworthiness of the system, but through that not necessarily the system's likability [3]. In other words, the relation between transparency through explainability, trustworthiness, and likability are not necessarily positively correlated. Here a potential for significant future research was identified.

(3) As a third topic it was discussed how putting users in control can be achieved through explanations and mitigation strategies in failure situations [2]. Our working hypothesis was that trust can be improved if users are involved in fixing the failure, e.g. pushing the robot out of a problem zone or putting it back into a charging station. However, related research already showed that there are potential cultural differences; e.g. that Japanese think that only experts should fix a robot, but not layman [1]. Similarly, the aspect was discussed if little failures might deliberately foster engagement and subsequently trust (research has already shown that the imperfect robot is more likable [5]).

(4) Finally, the issue of how far humans are trustworthy in Human-Robot Interaction was discussed. Aspects such as hostility (e.g. factory workers who fear replacement) and curiosity (e.g. kids in a museum "playing" with the robot). But how to make humans more compliant in the interaction? We discussed options such as justification for their actions, call/involvement of an authority and mimicking emotions. When presenting these thoughts in the plenary an interesting discussion on *"joint-human-robot-failure-recovery-and-trust-repair"* evolved which also identified novel research directions.

References

1 Markus Bajones, Astrid Weiss, and Markus Vincze. Investigating the influence of culture on helping behavior towards service robots. In *Proceedings of the Companion of the 2017*

ACM/IEEE International Conference on Human-Robot Interaction, HRI '17, pages 75–76, New York, NY, USA, 2017. ACM.
2 Markus Bajones, Astrid Weiss, and Markus Vincze. Help, anyone? a user study for modeling robotic behavior to mitigate malfunctions with the help of the user. *arXiv preprint arXiv:1606.02547*, 2016.
3 Shuyin Li and Britta Wrede. Why and how to model multi-modal interaction for a mobile robot companion. In *AAAI spring symposium: interaction challenges for intelligent assistants*, pages 72–79, 2007.
4 Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–148. ACM, 2015.
5 Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4:21, 2017.

5.2 Towards Artificial Moral Agency

Kai Spiekermann (London School of Economics, GB), Jan M. Broersen (Utrecht University, NL), Einar Duenger Bøhn (University of Agder, NO), Kerstin I. Eder (University of Bristol, GB), Christian List (London School of Economics, GB), Andreas Matthias (Lingnan University – Hong Kong, HK), Marcus Pivato (University of Cergy-Pontoise, FR), Thomas Michael Powers (University of Delaware – Newark, US), Teresa Scantamburlo (University of Venice, IT), Marija Slavkovik (University of Bergen, NO), and Leon van der Torre (University of Luxembourg, LU)

Starting questions: Can AI systems be genuine moral decision makers?
Observation: Some seminar participants think the answer is clearly 'yes', some think the answer is clearly 'no'. Where does this disagreement come from?

A minimal condition of agency (not necessarily moral). A basic (necessary but not sufficient) prerequisite for agency is that we can coherently take an intentional stance towards the system. For example, if a dog is after the sausage in my pocket, we observe a behaviour (following me, trying to get close to my pocket, responding to changes of sausage position, perhaps signalling to me that it wants the sausage) that is best explained by assuming that the dog has intentions, specifically the intention to get hold of and eat the sausage. By contrast, an intentional stance is not plausible towards a raindrop falling to earth (intentions are not necessary for a plausible explanation of the raindrop behaviour). Similarly, chairs and tomatoes do not warrant taking an intentional stance.

A taxonomy of agents (thin vs. thick)

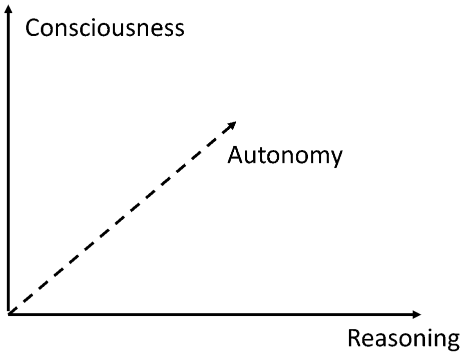

Figure 1 A taxonomy of (not yet necessarily moral) agents: from thin to thick on three dimensions.

Here are some examples of agents that can be classified according to this scheme:

Agent	**Reasoning Capacity**	**Autonomy**	**Consciousness**
Humans	High	High	High
Chimps	Low	High	High
Self-driving car	Intermediate?	Quite high	None
Thermostat	None/Low	None/Low	None

The Moral Turing Test. The basic idea of the standard Turing test is to check whether an observer can distinguish between a black-boxed human and a black-boxed artificial agent (specifically, a Turing machine), purely on the basis of symbolic input and output. The standard Turing test is usually construed as a diagnostic test for human-like intelligence. It is not normally construed as a test for moral behaviour. The idea behind the moral Turing test is to amend the standard Turing test so as to turn it into a diagnostic test for moral capacities. The moral Turing test exists in different versions: (1) Restrict conversation to moral issues and (2) VR setup.

The moral Turing Test tests for observable moral speech or action behaviour. One concern that was raised was that the test, if taken as a test for moral competence, might also invite abuse, e.g. by setting (morally indefensible) standards for human moral agents.

The Traditional and the Moral Chinese Room Argument. We set out the standard argument as presented by Searle, and possible charitable interpretations of it. In its basic setting, there is a human operator in a room who receives written messages in Chinese. The operator does not understand Chinese but has an operating manual that instructs her to respond to those messages in appropriate ways. Technically, we can think of the operator as executing a suitably programmed Turing machine or algorithm that symbolically converts Chinese inputs into adequate Chinese responses, but does so in an entirely syntactic manner. John Searle's claim is that there is no understanding of Chinese going on anywhere here, and therefore that syntactic processing alone is insufficient to generate semantic understanding.

One influential response is to concede that the operator does not understand Chinese, but to argue that the system as a whole does, where this is defined as the composite, consisting of the operator, the manual, and any storage shelves serving as memory in the room. This response is called the "systems response".

The group discussed the systems response in detail and critically investigated its plausibility. After discussing the standard Chinese room argument the group turned to the moral version of the argument. This is basically like the standard argument, except that the communication is restricted to or focused on morally relevant content. The question is whether this argument might in any way establish that purely syntactic machines are not capable of genuine moral understanding and moral agency. Many members of the group expressed the view that the moral Chinese room argument is less compelling than the standard one. One question to be asked is whether semantic understanding or full-blown intentionality in the sense discussed by Searle is necessary for moral agency. Those in the group who do not adhere to a very thick understanding of moral agency tend to think that the answer to this question is negative. (Note that this might not carry over to views about moral patiency.) While the group did not discuss moral patiency in great detail, several members of the group agreed that phenomenal consciousness is a necessary condition for moral patiency, but not for moral agency.

Conceptions of Autonomy. The group noted that there is not one single canonical definition of autonomy, even in debates about agency, but that there are a variety of surprisingly different definitions. These include, among others, definitions of autonomy as:

- unpredictability (though might this render a random walker autonomous?);
- choosing their own preference (might this lead to an infinite regress?);
- self-legislating (though what does this mean precisely?);
- having free choices (though there are many different notions of freedom out there);
- not being (too) influenced by the environment.

We focused, in particular, on the following three different notions of autonomous systems:

- As goal/preference revising / self-legislating systems;
- As systems that pursue set goals without direct intervention (IEEE);
- As systems that can only/best be predicted by running the system.

It was noted that verification (i.e. the process used to gain confidence in the correctness of a system with respect to its specification) requires the specification of the behaviour to be verified, and that this was problematic / challenging for some of these notions of autonomy.

5.3 How Do We Build Practical Systems Involving Ethics and Trust?

Louise A. Dennis (University of Liverpool, GB), Andrea Aler Tubella (University of Umeå, SE), Raja Chatila (Sorbonne University – Paris, FR), Hein Duijf (Free University Amsterdam, NL), Abeer Dyoub (University of L'Aquila, IT), Kerstin I. Eder (University of Bristol, GB), John F. Horty (University of Maryland – College Park, US), Maximilian Köhl (Universität des Saarlandes, DE), Robert Lieck (EPFL – Lausanne, CH), and Munindar P. Singh (North Carolina State University – Raleigh, US)

There is clearly no one unique way to approach the construction of artificial systems that are both ethical and trustworthy. Our working group considered both the variety of ideas, methods and techniques that might contribute to the construction of such systems and, via the consideration of two case studies, attempted to identify the gaps in our understanding which needed to be filled before such systems were possible.

Pathways. We identified a number of ideas that were necessary for the construction of such systems and categorised these into stages in a pathway that leads from the abstract to the concrete. Although we used the term pathway we did not, by this, intend that an ethical and trustworthy system should be constructed first by selecting a philosophical standpoint and then moving towards the ever more concrete, just that an ethical and trustworthy system must involve concepts from all stages in a pathway but may have been designed and constructed in an iterative process where choices at all stages interacted with each other. The stages in the pathway we identified and some of the possible choices within that stage are set out below:

- philosophy: ethics, law, sociology, psychology, politics
- ontology: stakeholder, autonomy, norms, reasons, intentions, plans, values
- theories: deontic logic, rights & duties, norms & obligations, agent theory
- design techniques: model architecture, data selection
- implementation techniques: programming languages, NN structure, machine learning, synthesis
- analysis techniques: data analysis, theorem proving, testing, simulation, (code) review

One possible such path to ethical and trustworthy artificial systems goes from Philosophy and Law through Reasons/Intentions/Plans/Values to Deontic Logic and Agent Theory terminating in declarative or normative programming frameworks of various flavours (see [2]) to which a variety of analysis techniques can be applied in order to demonstrate the trustworthiness of the final implementation.

Desiderata for Analysis of "AI" Systems which help build justifiable trust. A number of missing tools and techniques were identified for the later stages in the pathways, those at the more concrete and computational end. In particular we identified the need for novel or better techniques and tools to support the development of sub-symbolic systems (typified by deep neural networks) which do not manipulate explicit human-understandable representations in order to make decisions. However, related to this we argue that for a system to genuinely embody ethics it will be necessary to have architectures which combine symbolic and sub-symbolic reasoning and tools for developing, verifying and validating such systems. We will want symbolic representations because ethical reasoning is generally about how users believe something should behave which may be easy to state symbolically but difficult to express sub-symbolically. We may nevertheless want sub-symbolic reasoning to handle other aspects of control and decision making including the implementation of the situational awareness that will trigger ethical reasoning. While not exactly a combination of symbolic and sub-symbolic reasoning, the GenEth system [1] is an example of a system that uses machine learning to construct and explicit representation of ethical rules. A variety of techniques will be required in order to justify confidence in a system's trustworthiness. These include, but are not limited to,

- explanation mechanisms for symbolic and sub-symbolic reasoning – systems that allow us an insight into how they make decisions, explain why they act in a certain way or how they use resources become understandable and thus trustworthy; and
- verification and validation techniques – rigorous proof complemented by simulation and end-user testing can provide convincing evidence of a system's trustworthiness.

In order to explore these ideas in greater depth, the group focused on two case studies.

Case Study: Complex System of Norms. We consider a system in which a hierarchy of explicit norms are used to control ethical reasoning and identified a number of issues relating to the construction of such a system. A key issue in such a system will be handling conflicts between norms. These may involve problems that are traditionally considered ethical dilemmas (such as trolley problems) but may also involve other kinds of conflicts.

A simple example of a dilemma-style conflict is when some action is both obligated and prohibited by the norms. There are techniques that can be used to detect potential conflicts at design (e.g. using techniques like those in [4]) time but these inevitably involve some way of capturing the contexts the system may find itself in and so it may also be necessary to use runtime techniques to detect when a conflict has arisen.

A complex hierarchy of norms is likely to arise because norms are being sourced from a variety of places. For instance, some norms may be legal, some may be related to professional practice (e.g., in healthcare situations) and some may be social. Tracking the sources of norms may be a key to discovering inconsistencies and preventing conflicts.

It may be, however, that conflicts arise not because of explicit dilemmas but because of conflicts between norms and lower level processes. For instance, many robotic systems are engineered with low-level obstacle avoidance processes that take precedence over explicit reasoning. Such behaviour might cause a norm to be violated. In a complex system, norms are also likely to be context sensitive, determining whether a norm applies will depend upon the system's situational awareness which is likely to depend upon sub-symbolic processes for, for instance, image classification. The way probabilistic and possibly faulty assessments of situations interact with normative reasoning was both theoretically and practically unclear.

When a conflict arises it is then necessary to decide what to do. During system design, there is time for design processes to determine this, but at runtime this may not be possible. In some systems it may be possible to implement an "ethical fail safe", but where such an option does not exist, other methods might be needed such as selecting one of the possible actions at random, or having some kind of sub-symbolic "shadow" (possibly in a similar way to [6]) of the explicit norms which is used to make decisions when the explicit system is unable to. Explicit reasoning about potential sanctions for norm violation might also assist in the resolution of conflicts [7]. Ideally, once conflicts are resolved, norms are updated to reflect the resolution.

Algorithmic Bias. We observed that discussion of algorithmic bias, its definition, causes and mitigation was currently a topic of much active research which the working group unfortunately had little expertise in (Some preliminare references are [3, 5]). We also noted that the term "bias" was overloaded in the communities involved. For the purposes of the discussion we agreed to consider bias to be when information related to, for instance, a protected characteristic such as gender, race, religion or sexual orientation, was used as part of a decision-making process when the information was in reality irrelevant to the outcome - for instance taking gender into account when predicting success at an office job. We noted that discussion of algorithmic bias tended to focus on sub-symbolic systems but that bias is possible even with explicitly engineered norms:

- Biases in context detection and classification where a sub-symbolic system attributed characteristics to people based on stereotypes could lead to an explicitly normative system making biased decisions based upon that classification.
- Prioritisation between norms (or possibly other interactions) can create bias, for instance norms around parenting tend to affect women more than men and so the priority given to such norms might disadvantage one of these groups – this may depend upon the deployment context.
- Norms themselves can be biased (for instance the norm that women and children should be evacuated first in an emergency).

Techniques for defining and detecting bias therefore have potential application to both symbolic and sub-symbolic systems if appropriately constructed.

Case Study: Sub-symbolic Algorithmic Bias. We considered the possibility of sub-symbolic algorithmic bias in the construction of ethical and trustworthy systems. Assuming we have an adequate definition of bias, the two key problems become how to detect if a system is biased and how to fix a biased system.

(1) How to detect if a system is biased? At design time the provenance of the data can be analysed in order to assess the likelihood of bias being present in the data. Where particular groups who may be biased against can be identified, it is possible to perform statistical analyses of the performance of the system in order to identify potential bias and to have experts inspect the system's output on specific examples in order to determine if it is making appropriate decisions. A particular research challenge in this context is the definition and validation of fairness metrics that can be used to identify bias automatically.

(2) How can biased training data and biased systems be fixed? Once bias is detected, then the reasons for the bias need to be analysed. In general, this requires explanation techniques for analysing decision making in sub-symbolic systems. If non-biased data (or a non-biased subset of data) or missing data can be identified, then the system can be retrained and then re-evaluated for bias. However, techniques may be required to fix biased data or to explicitly screen decisions for bias, where no unbiased data is available. This may itself involve the use of sub-symbolic techniques. The group was aware that there was active research in these areas but was not familiar with the literature.

Conclusion. We considered a number of issues relating to the construction of ethical and trustworthy systems both at the symbolic and sub-symbolic level. We identified a lack of theories, techniques and tools at the more concrete end of the pathways to constructing such systems, in particular the need to combine symbolic and sub-symbolic reasoning to allow ethics/norms to be analyzed and manipulated at both levels. We considered two particular examples of systems that presented challenges to ethics and trustworthiness: systems that use complex sets of explicit norms and the problem of algorithmic bias. The problem of conflicts and dilemmas was of major concern for systems of explicit norms but we also concluded that algorithmic bias could arise not only from biased data, but also from explicit norms and the interaction of explicit norms with biased systems. As a result we identified the following important research questions:

- How can we combine symbolic and sub-symbolic reasoning to enable both flexible reasoning and explicit representations?
- What techniques can we develop to enable us to better analyse and modify the behaviour of sub-symbolic systems (e.g., debuggers, profilers).
- What techniques can we develop to verify and validate the behaviour of sub-symbolic systems (including formal methods and simulation/test-based approaches).
- How can we monitor sub-symbolic systems to detect and contain undesirable behaviour?
- How can/should autonomous systems explain their behaviour?
- How should an autonomous system resolve an ethical/normative conflict in situations where no "ethical fail safe" exists?
- How can we detect "algorithmic bias" in both symbolic and sub-symbolic systems, related to this how do we adequately define algorithmic bias?
- If no unbiased training data exists, how do we correct for bias in sub-symbolic systems?

References

1 M. Anderson and S. Leigh Anderson. Geneth: A general ethical dilemma analyzer. In *Proceedings of the 28th AAAI Conference on AI, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 253–261, 2014.

2 Amit K. Chopra and Munindar P. SIngh. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 48–53, New York, NY, USA, 2018. ACM.

3 David Danks and Alex John London. Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4691–4697. AAAI Press, 2017.

4 Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matthew P. Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

5 Keith Kirkpatrick. Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Commun. ACM*, 59(10):16–17, September 2016.

6 Robert Lieck. *Learning Structured Models for Active Planning: Beyond the Markov Paradigm Towards Adaptable Abstractions.* PhD thesis, Universität Stuttgart, Stuttgart, 2018.

7 Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. Classifying sanctions and designing a conceptual sanctioning process for socio-technical systems. *The Knowledge Engineering Review*, 31(2):142–166, March 2016.

5.4 The Broader Context of Trust in HRI: Discrepancy between Expectations and Capabilities of Autonomous Machines

Emily Collins (University of Liverpool, GB), Kerstin Fischer (University of Southern Denmark – Sonderborg, DK), Bertram F. Malle (Brown University – Providence, US), AJung Moon (Open Roboethics Institute – Vancouver, CA), and James E. Young (University of Manitoba – Winnipeg, CA)

One of the first steps to discussing ethical and trustworthy robots is to take stock of the complex human responses that autonomous machines trigger. Certain robot features may cause human trust, but they may be superficial triggers of trusting feelings or properties that actually justify such trust. However, the challenge is even broader. We need to build a systematic understanding of how design choices affect a complex variety of human responses—including not only trust but other cognitive, emotional, and relational ones. Importantly, these responses often do not reflect the real capabilities of the designed robot, causing discrepancies between what humans perceive the robot to be and what it actually is. We summarize here some of these discrepancies and ways to mitigate them.

The Multidimensionality of Human Reactions. Human responses to robots comprise a wider range of dimensions and are caused by a wide array of factors. Humanoid robots can elicit in-group bias [3], cheater detection [9], spontaneous visual perspective taking [15], and gaze following in infants [10]. Such responses are influenced by the robot's social role [7], people's expectations about robots e.g. [11], their expertise e.g. [4], and even psychosocial predispositions such as loneliness [8]. Humanlike appearance is a particularly powerful cause, leading people to see robots as more intelligent, more autonomous, and as having more mind ([1]; [2]; [14]). But people treat even disembodied technologies similar to human beings [12] and respond to them with behavior that is conventionally appropriate [5]; [16].

If trust is only one response within a manifold of interrelated responses, it becomes unclear which properties of a machine superficially trigger trust and which ones justify trust. Moreover, recent studies indicate that the concept of trust itself is multidimensional. That is, one could trust another human (or perhaps robot) owing to different kinds of evidence—their reliability, competence, sincerity, or ethical integrity ([13]; see http://bit.ly/MDMT_Scale).

Discrepancies Between Human Perceptions and Actual Robot Capacities. Current robot design tends to integrate a large number of social cues into robots' behavior and appearance. However, when interacting with humans, social cues are symptoms of true underlying processes, but robots that show these same cues usually do not have these underlying processes. For example, robots using gaze cues are seen as indicating joint attention and an understanding of a speaker's instructions [6], but robots can produce these behaviors without actually understanding the speaker's communication at all. Equipping a robot with such cues is therefore confusing, if not deceptive, because it creates the impression that the robot has capabilities it does not actually have. Mismatches between expected and real capabilities pose manifest risks. Users may entrust the robot with tasks that the robot is not equipped to do and will be disappointed, frustrated, or distressed when they discover the robot's limited capabilities. In turn, such users will no longer use the product, write scathing public reviews, or even sue the manufacturer.

Discrepancies between perceived and actual capacities of robots have multiple sources. Public media and its frequent exaggerations of technical realities is one source. Deceptive advertisement of robotic products, especially those for social robots intended for consumers, is another. Researchers using Wizard-of-Oz methods can also contribute to spreading false beliefs, because they create an illusion of capacities of the robotic platform, and thorough debriefing after such experiments is often lacking. Finally, since humans acquire capabilities in a particular order such that more basic capabilities provide the basis for more complex ones, they find it hard to imagine that a robot can have a sophisticated ability without having acquired all the more basic capabilities [4].

How to Combat the Discrepancies. How can people recover from mismatches between perception and reality? Currently we do not know. It would take a serious research agenda to understand the conditions of recovery and correction, and it would take multiple approaches. First, because we as yet have no systematic mapping between the specific robot features that elicit specific affective and cognitive responses in humans, we need carefully controlled experiments to establish these causal relations. Second, to better separate deeply ingrained and unchangeable responses from culturally learned and correctable ones, we need to compare response patterns of young children and adults, as well as of people from different cultures. Third, to truly understand how human responses to robots can change we need longitudinal studies that consider the full array of multi-dimensional responses and measure how they change as a result of interacting with robots over time.

High-quality longitudinal research faces numerous obstacles: from cost, time, and required management efforts to participant attrition and ethical concerns of their privacy, from the familiar high rate of mechanical robot failures to their unforeseen effects on daily living. Smaller initial steps are possible, however, to study temporal dynamics that will advance knowledge but also provide a launching pad for genuine longitudinal research. For example, experiments can compare people's responses to a robot with or without information about its true capacities and assess whether people are able to adjust their perceptions. Other experiments can present a robot twice and track people's changing representations from the first to the second encounter, perhaps unfolding differently depending on the specific response dimension. Short-term longitudinal studies could also bring participants back to the laboratory more than once and distinguish people's adjustments to the specific robot (if they encounter it again) from adjustments of general beliefs about robots (if they encounter a different robot).

Another path to handling mismatches between perceived and real robot capabilities is to prevent such discrepancies in the first place. One strategy is incremental robot design—the commitment to advance robot capacities in small steps, each of which is well grounded in user studies and eases people into a changing reality of capacities. Another is to build users' understanding of the robot's behavior by revealing its actual causes and also explicate the robot's limitations. Designers and manufacturers may be reluctant or unable to offer effective explanations of the machine's real capacities (e.g., because of communicative distance between manufacturer and user or because of user suspicion), so the machine might be in the best position to explain its own behavior and limitations. People's perceptions may be stubborn, but explanations that arise in the immediate context of human-robot interaction and in repeated communications might break through people's expectations and inferences and, over time, alleviate discrepancies between perceived and actual robot capacities.

References

1 Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics*, 1(2):195–204, 2009.

2 Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q Stafford, Bruce A MacDonald, and Daniel M Wegner. Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PloS one*, 8(8):e72589, 2013.

3 Friederike Eyssel and Dieta Kuchenbrandt. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4):724–731, 2012.

4 Kerstin Fischer. What computer talk is and isn't. *Human-Computer Conversation as Intercultural Communication*, 17, 2006.

5 Kerstin Fischer. *Designing speech for a recipient: the roles of partner modeling, alignment and feedback in so-called'simplified registers'*, volume 270. John Benjamins Publishing Company, 2016.

6 Kerstin Fischer, Katrin Lohan, Joe Saunders, Chrystopher Nehaniv, Britta Wrede, and Katharina Rohlfing. The impact of the contingency of robot feedback on hri. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 210–217. IEEE, 2013.

7 Jennifer Goetz, Sara Kiesler, and Aaron Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60. Ieee, 2003.

8 Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human-computer studies*, 64(10):962–973, 2006.

9 Alexandru Litoiu, Daniel Ullman, Jason Kim, and Brian Scassellati. Evidence that robots trigger a cheating detector in humans. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 165–172. ACM, 2015.

10 Andrew N Meltzoff, Rechele Brooks, Aaron P Shon, and Rajesh PN Rao. “Social” robots are psychological agents for infants: A test of gaze following. *Neural networks*, 23(8-9):966–972, 2010.

11 Steffi Paepcke and Leila Takayama. Judging a bot by its cover: an experiment on expectation setting for personal robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 45–52. IEEE, 2010.

12 Byron Reeves and Clifford Ivar Nass. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.

13 Daniel Ullman and Bertram F Malle. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 618–619. IEEE, 2019.

14 Michael L Walters, Kheng Lee Koay, Dag Sverre Syrdal, Kerstin Dautenhahn, and René Te Boekhorst. Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials. *Procs of New Frontiers in Human-Robot Interaction*, 2009.

15 Xuan Zhao, Corey Cusimano, and Bertram F Malle. Do people spontaneously take a robot’s visual perspective? In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 335–342. IEEE, 2016.

16 Clifford Nass. Etiquette equality: exhibitions and expectations of computer politeness. *Communications of the ACM*, 47(4):35–37, 2004.

Participants

- Andrea Aler Tubella
University of Umeå, SE
- Jan M. Broersen
Utrecht University, NL
- Einar Duenger Bøhn
University of Agder, NO
- Raja Chatila
Sorbonne University – Paris, FR
- Emily Collins
University of Liverpool, GB
- Louise A. Dennis
University of Liverpool, GB
- Franz Dietrich
Paris School of Economics & CNRS, FR
- Clare Dixon
University of Liverpool, GB
- Hein Duijf
Free University Amsterdam, NL
- Abeer Dyoub
University of L'Aquila, IT
- Sjur K. Dyrkolbotn
West. Norway Univ. of Applied Sciences – Bergen, NO
- Kerstin I. Eder
University of Bristol, GB
- Kerstin Fischer
University of Southern Denmark – Sonderborg, DK
- Michael Fisher
University of Liverpool, GB
- Marc Hanheide
University of Lincoln, GB
- Holger Hermanns
Universität des Saarlandes, DE
- John F. Horty
University of Maryland – College Park, US
- Maximilian Köhl
Universität des Saarlandes, DE
- Robert Lieck
EPFL – Lausanne, CH
- Felix Lindner
Universität Freiburg, DE
- Christian List
London School of Economics, GB
- Bertram F. Malle
Brown University – Providence, US
- Andreas Matthias
Lingnan University – Hong Kong, HK
- AJung Moon
Open Roboethics Institute – Vancouver, CA
- Marcus Pivato
University of Cergy-Pontoise, FR
- Thomas Michael Powers
University of Delaware – Newark, US
- Teresa Scantamburlo
University of Venice, IT
- Munindar P. Singh
North Carolina State University – Raleigh, US
- Marija Slavkovik
University of Bergen, NO
- Kai Spiekermann
London School of Economics, GB
- Myrthe Tielman
TU Delft, NL
- Suzanne Tolmeijer
Universität Zürich, CH
- Leon van der Torre
University of Luxembourg, LU
- Astrid Weiss
TU Wien, AT
- James E. Young
University of Manitoba – Winnipeg, CA