Report from Dagstuhl Seminar 23371

Roadmap for Responsible Robotics

Michael Fisher[*1], **Marija Slavkovik**[*2], **Anna Dobrosovestnova**[*3], **and Nick Schuster**[*4]

1 **University of Manchester, GB.** michael.fisher@manchester.ac.uk
2 **University of Bergen, NO.** marija.slavkovik@uib.no
3 **TU Wien, AT.** anna.dobrosovestnova@tuwien.ac.at
4 **Australian National University – Canberra, AU.** nick.j.schuster@gmail.com

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23371 "Roadmap for Responsible Robotics". The seminar was concerned with robots across all their forms, particularly autonomous robots capable of making their own decisions and taking their own actions without direct human oversight. The seminar brought together experts in computer science, robotics, engineering, philosophy, cognitive science, human-robot interactions, as well as representatives of the industry, with the aim of contributing to the steps towards ethical and responsible robotic systems as initiated by actors such as the European Robotics Research Network (EURON), the European Union's REELER, and others. We discussed topics including: "Why do autonomous robots warrant distinct normative considerations?", "Which stakeholders are, or should be, involved in the development and deployment of robotic systems, and how do we configure their responsibilities?", "What are the principal tenets of responsible robotics beyond commonly associated themes, namely trust, fairness, predictability and understandability?". Through intensive discussions of these and other related questions, motivated by the various values at stake as robotic systems become increasingly present and impactful in human life, this interdisciplinary group identified a set of interrelated priorities to guide future research and regulatory efforts. The resulting roadmap aimed to ensure that robotic systems co-evolve with human societies so as to advance, rather than undermine, human agency and humane values.

Seminar September 10–15, 2023 – https://www.dagstuhl.de/23371

2012 ACM Subject Classification Computer systems organization → Robotic autonomy; Social and professional topics → Codes of ethics; Security and privacy → Trust frameworks; Software and its engineering → Software verification and validation

Keywords and phrases Robotics, Responsibility, Trust, Fairness, Predictability, Understandability, Ethics

Digital Object Identifier 10.4230/DagRep.13.9.103

1 Executive Summary

Michael Fisher
Marija Slavkovik
Anna Dobrosovestnova
Nick Schuster

The ISO 8373 standard ("Robots and Robotic Devices – Vocabulary") defines a robot as "an actuated mechanism programmable in two or more axes moving within its environment, to perform intended tasks". Aligned with this definition, we consider "robotics" to cover a wide

[*] Editor / Organizer

range of devices – e.g. vehicles, probes, drones, industrial devices, and personal robots – as well as the complex sociotechnical processes surrounding the development and deployment of such systems. Given that robotic systems are increasingly capable of acting without direct human oversight, and that they're being deployed in an increasing variety of contexts, a range of concerns beyond technical reliability emerge. Many authors, across a variety of disciplines, have pointed to the need for "responsibility" in robotic systems. However, while it is popular to highlight this as a target, there is no agreed route to achieving *responsible robotics*. In addition, there is sometimes even little agreement on what *responsibility* here comprises.

The aim of this Dagstuhl Seminar was to identify the key components of responsibility in this context and then, crucially, provide a roadmap for achieving responsible robotics in practice. By doing so, the seminar contributed to the ongoing efforts established with the Roboethics Roadmap put forth in January 2007 by the European Robotics Research Network (EURON), the European Union's REELER, SIENNA, and TECHETHOS projects, and the UK's RoboTIPS project, among others.

In the original proposal of the seminar, four themes commonly associated with responsible robotics were emphasized: *trust*, *fairness*, *reliability*, and *understandability*. In the course of the seminar, however, the participants – comprising philosophers, engineers, roboticists, cognitive scientists, and industry representatives – identified a broader range of concerns. Firstly, some discussions focused on what responsibility means from different disciplinary perspectives and how these apply to the development, deployment, use, and disposal of robots. In these discussions, it was emphasized that the very term "responsibility" is ambiguous in philosophy and law. The ambiguity and the complexity of the term is, however, rarely reflected in the debates on responsibility in the context of AI and robotics. Referring to [1], responsibility gaps in sociotechnical systems were discussed. We converged on an understanding of responsible robotics as broadly capturing the idea that various parties involved in development, deployment, integration, and maintenance of robots need to be acting in a responsible manner. This involves behaving ethically in their various roles, building ethically sensitive robots, and ultimately taking responsibility for how robotics as a field progresses and how robots are used. This includes "role responsibility", relating to specific functions in robotics; "professional responsibility", which covers obligations in the robotics profession; "moral responsibility", involving ethical decision-making and anticipation of consequences; "legal responsibility", pertaining to compliance with relevant laws and regulations; "social responsibility," regarding the broader impacts of robotic systems on human societies; and "environmental responsibility," regarding their impacts on the natural environment.

As an important step to ensure responsible robotics, discussions considered the diverse roles and responsibilities of key stakeholders, including businesses, universities, governments, users, and others who stand to affect, or be affected by, robotic systems. Specifically, it was noted that universities play a crucial role in shaping the professionals who design, engineer, and operate robotic systems. Engineering and design curricula should thus include modules on responsible innovation, safety standards, and the potential consequences of misuse. This could be done by intensifying the dialogue and collaborations with other disciplines, in particular humanities and social sciences, following promising initiatives such as Embedded EthiCS. To align robotics with ethical standards, businesses in turn must conduct thorough risk assessments, addressing potential misuses and implementing safeguards in their products. For example, in the case of AI-based robotic systems, providers may rely on existing risk management frameworks such as the one recently developed by the

National Institute of Standards and Technology for AI system (`https://www.nist.gov/itl/ai-risk-management-framework`). Additionally, they should provide comprehensive user manuals, conduct user training programs, and actively collaborate with regulatory bodies to establish industry-wide standards. Transparent communication about the capabilities and limitations of their products is essential to ensure that users have a clear understanding of how to responsibly engage with robotic technologies. Furthermore, governments play a pivotal role in creating and enforcing regulations that govern the use of robotic products and services. They must collaborate with industry experts to establish ethical guidelines, safety standards, and legal frameworks. Regulatory bodies should continuously update these frameworks to keep pace with technological advancements. Furthermore, governments should invest in public awareness campaigns to educate citizens about the benefits and risks of robots, mitigating the potential for misuse or misunderstanding.

Discussions also emphasized that an extended definition of responsibility, encompassing not only technical but also social and political considerations, requires a similarly expansive understanding of trust, fairness, reliability, and understandability as well as the addition of other normative concepts. To address this, other potentially relevant concepts were identified through an iterative voting exercise. The final list included: *dignity*, the inherent worth of each member of the moral community who stands to be impacted by robotic systems; *autonomy*, enabling human beings to act in accordance with their own interests and aspirations; *privacy*, empowering people to protect and share sensitive information about themselves as they see fit; *safety*, protecting the various aspects of physical and emotional well-being; *trust*, ensuring that people have good reason to believe that robotic systems are aligned with their legitimate interests; *justice/fairness,* making the impacts of robotic systems acceptable to all who stand to be affected by them; *accountability,* ensuring that the right agents are held to account for adverse outcomes; and *sustainability,* regarding the impacts of robotic systems on the natural world and future generations. It was not our objective to generate an exhaustive list. Rather, the list reflected the principle concerns that emerged from discussion of current and near-future uses and capabilities of robotic systems.

In summary, apart from the group level discussions, 4 working groups were held: These included working groups on:

- Fairness
- Trust
- Why robots require different considerations?
- Predictability

In sum, the main outcome of the seminar was a draft of a document developed collectively and encompassing these and other related topics. The document is intended for a wide range of stakeholders and relevant, affected parties, including researchers, policymakers, industry leaders, practitioners, NGOs, and civil society groups. Recognizing that the current group of authors primarily represents research perspectives (and those coming primarily from the Global North), we are aware of the necessity to incorporate a broader array of viewpoints. Therefore, we are committed to including more diverse perspectives in this discussion going forward, to inform future versions of this roadmap, to better promote the development of responsible robotics, and to help navigate the complex sociotechnical terrain that lies ahead.

References

1 Santoni de Sio, Filippo, and Giulio Mecacci. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34: 1057-1084, 2021.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Fairness in robotics: a philosophical approach

Helen Beebee (University of Leeds, GB)

"Fairness" is a concept that crops up a lot in discussions of AI ethics. However it is, I claim, often used in a much broader sense than in philosophy. In philosophy fairness is a concept that is normally deployed in the context of "distributive justice" – that is, where what is at issue is the "fairest", or most just, allocation of resources, opportunities, and so on across society. Fairness, however, is generally not carefully defined. A straightforward and relatively uncontroversial definition might be: A process or rule – and, by extension, an outcome or action resulting from it – is fair if it treats everyone equally, unless unequal treatment is merited/justified/deserved.

Fairness is specifically about treating people equally, and not about treating people as they deserve to be treated – which is a much broader concept. (Nobody deserves to be burgled, but the fact that X was burgled while other people were not is not inherently unfair.)

In AI, fairness seems to have been singled out as a distinctively important moral concept. This is perhaps merited when it comes to the kinds of bias that can arise in decision-making based on the application of large-scale demographic data to a particular case, as is sometimes the case in machine learning. This does apply to come extent to robotics, and – when it does apply – the relevant considerations are how a given demographic generalisation is being used, whether there is differential treatment, and, if there is, whether that differential treatment is justified. But behaving fairly is just one way of behaving well. Robots operate in local situations, and – just like humans – they need to behave well more generally, and not just fairly. So it is not at all clear that "fairness" is more important than other moral concepts when it comes to robotics.

3.2 Responsibility in Autonomous Systems

Michael Fisher (University of Manchester, GB)

Main reference Louise A. Dennis, Michael Fisher: "Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines", Cambridge University Press, 2023.
URL https://doi.org/10.1017/9781108755023

Responsibility in Robotics comprises two, related aspects:

1. The responsible development of robotic systems; and
2. The responsibility that our robots have, especially once they become autonomous.

Concerning (1) there is already a vast literature on "responsible innovation" and, while we must build on that, we need also take into account the issues relevant to robotics and particularly autonomous robotics. Work on standards in these areas is important, for example the British Standards Institution "Guide to the Ethical Design and Application of Robots and Robotic Systems" (BS8611), published in 2016 and revised in 2023, as well as related work on "Sustainable Robotics" (BS8622). A key aspect here is that strong verification techniques (beyond probabilistic estimates) should be required , especially where robots are to be involved in critical issues. In all this influencing institutions/government/regulators, etc, is vital. To change/update regulatory guidelines, to stimulate changes in government policies, and to change the way robotic systems are developed.

Once we delegate sufficient agency to a robotic system, making it autonomous, then the decisions the robot might make become crucial. Here, the trustworthiness of autonomous robots becomes central. Although trustworthiness in standard systems often equates to "reliability", the move to more autonomous systems expands trustworthiness so that it must incorporate beneficiality – that we believe the robot is making its decisions for our benefit. Such views support a move to more nuanced architectures (e.g. neuro-symbolic) providing better ways to build autonomous robots and making predictability, understandability, fairness, and trustworthiness easier

References

1 British Standards Institution. BS 8611: Guide to the Ethical Design and Application of Robots and Robotic Systems. `https://standardsdevelopment.bsigroup.com/projects/9021-05777`. Revised 2023.

2 Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik„ Stuart Russell, and Karen Yeung. *Trustworthy AI.* In *Reflections on Artificial Intelligence for Humanity*, pages 13–39. Springer, 2021.

3 Louise A. Dennis and Michael Fisher. *Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines.* Cambridge University Press, 2023.

4 Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Bernd-Holger Schlingloff, Michael Winikoff, and Neil Yorke-Smith. Towards a Framework for Certification of reliable autonomous systems. *Autonomous Agents and Multi Agent Systems*, 35(1):8, 2021.

5 M. R. Mousavi, A. Cavalcanti, M. Fisher, L. Dennis, R. Hierons, B. Kaddouh, E. L.-C. Law, R. Richardson, J. O. Ringer, I. Tyukin, and J. Woodcock. Trustworthy Autonomous Systems Through Verifiability. *Computer*, 56(2):40–47, 2023.

6 Interest Group on Neuro-Symbolic AI. Alan Turing Institute. `https://www.turing.ac.uk/research/interest-groups/neuro-symbolic-ai`.

3.3 The Importance for Robots of Knowing When They Don't Know

Michael Milford (Queensland University of Technology – Brisbane, AU)

Joint work of Helen Carson, Jason Ford and Michael Milford
Main reference Helen Carson, Jason J. Ford, Michael Milford: "Unsupervised Quality Prediction for Improved Single-Frame and Weighted Sequential Visual Place Recognition", in Proc. of the IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 – June 2, 2023, pp. 3955–3961, IEEE, 2023.
URL https://doi.org//10.1109/ICRA48891.2023.10160679

As robotics and automation both matures and stalls due to hard deployment challenges, it looks increasingly likely that fully automated, unsupervised systems will only make up a subset of total robot deployments, for both capability and operational concerns. Instead much robotic deployment will occur in collaborative and semi-supervised environments, where robots and people both play a role and interact in rich and meaningful manners beyond simple supervision and oversight. To maximize both the capability of these autonomous systems as well as their collaborative potential with human operators, both present and remote, these autonomous systems will need to "know when they don"t know" – the power of introspection. Introspection enables graceful performance degradation, but also facilitates handover to human operators. From a pragmatic point of view, for any safety or operationally critical activity, a system with a certain level of performance, say in terms of accuracy, but no introspection capability will often be vastly inferior to a system with slightly lower accuracy but good introspection capability. Whilst introspection and related concepts like verification are mature, well practiced areas in domains like aerospace, their consideration and treatment in robotics is relatively early stage or not done. We propose that a substantial research investment and focus in introspection for robotics and autonomous systems will pay dividends, both in terms of advancing knowledge but particularly in enabling promising robotic technologies to successfully make the transition into trusted, enduringly deployed systems.

3.4 Trust and Interactive Robotics

AJung Moon (McGill University – Montreal, CA)

This talk unpacks the many ways in which the word "trust" is and has been used in robotics, human-robot interaction, and AI ethics. Between "trustworthy AI" and user trust in specific capabilities of a robot, there is a large gap and diversity of solutions to address the trust-trustworthiness problem.

Building on the current trends in AI ethics (namely, principle-based approaches toward trustworthy AI and framing of ethics issues as model/product-level fairness/transparency/accountability problems) and the temptation by roboticists to borrow much of AI ethics contents directly for "responsible robotics," I problematize how these trends can fail us in our attempts to build generally good robotic systems.

I describe this as a 'water bottle model of trust'. I argue that responsible robotics is an exercise that should help build trustworthy robotics design norms that focus on considerate forms of design and deployment of all robots, rather than one that narrowly guides the ethical design of a single system/hardware/feature. In this process, we should challenge our existing assumptions/taboos (e.g., those related to anthropomorphization) and think about what our shared vision of the world with robots looks like.

3.5 Encouraging Inferable Behavior for Autonomy: Repeated Bimatrix Stackelberg Games with Observations

Ufuk Topcu (University of Texas – Austin, US)

Joint work of Mustafa O. Karabag, Sophia Smith, David Fridovich-Keil, Ufuk Topcu

When interacting with other non-competitive decision-making agents, it is critical for an autonomous agent to have inferable behavior: Their actions must convey their intention and strategy. For example, an autonomous car's strategy must be inferable by the pedestrians interacting with the car. We model the inferability problem using a repeated bimatrix Stackelberg game with observations where a leader and a follower repeatedly interact. During the interactions, the leader uses a fixed, potentially mixed strategy. The follower, on the other hand, does not know the leader's strategy and dynamically reacts based on observations that are the leader's previous actions. In the setting with observations, the leader may suffer from an inferability loss, i.e, the performance compared to the setting where the follower has perfect information of the leader's strategy. We show that the inferability loss is upper-bounded by a function of the number of interactions and the stochasticity level of the leader's strategy, encouraging the use of inferable strategies with lower stochasticity levels. As a converse result, we also provide a game where the required number of interactions is lower bounded by a function of the desired inferability loss.

4 Working groups

4.1 Are robots different from AI? Definition and (some) Related Considerations

Anna Dobrosovestnova (TU Wien, AT)

Joint work of Kevin Baum, Anna Dobrosovestnova, AJung Moon, SeverinLemaignan, Pericle Salvini, Sara Ljungblad, Kate Devitt Ufuk Topcu, Raja Chatila, Teresa Scantamburlo

According to the definition of robot laid out in in the ISO 8373:2012 standard (International Organization for Standardization [ISO], 2012)): robot is an actuated mechanism programmable in two or more axes with a degree of autonomy (i.e., the ability to perform intended tasks based on current state and sensing, without human intervention), moving within its environment, to perform intended tasks. This definition allows us to distinguish between robots and other automated systems. Specifically, it implies that a robot is, first and foremost, a physical piece of machinery. This already excludes software, e.g. software bots, voice assistants, or image recognition from the broad category of robots. Furthermore, the definition underscores how robots require some degree of autonomy. Sostero (2020) points out certain ambivalence when it comes to anchoring what autonomy means because current state of the art technology and existing regulations allow only for limited autonomy. That said, this means the given definition of robot also excludes mobile machinery that only follows pre-programmed instructions without coupling between the machinery and the environment (e.g. 3D printers). To summarize, robots can be considered intelligent embodied agents situated in the real world, which means their existence and operation occur in the real world.

While many concerns related to ethical implications and responsibility overlap between robots and AI systems, the embodied and (partly) autonomous nature of robots bring with it a host of considerations relevant in the context of the broader conversation about ethics of robotics and responsibility. Firstly, the physicality of the robotic systems mean they can cause physical harm and cause injury. Secondly, the embodied nature of robots, coupled with autonomous movement and perceived goal orientedness is known to elicit in people a tendency to respond and treat robots as (quasi-)social actors. This tendency is further enabled by the fact that the so called social robots are increasingly developed to look and behave like humans. The potential dangers of designed and/or in perceived robot sociality have already been discussed in the robot ethics and related literature in relation to deception, unilateral bond, and how such robots can reshape affect and relationality laden practices e.g., when it comes to robots deployment in service sectors. Beyond these issues, we also identified a host of challenges related to the uncertainty of deployment the robots in the physical world. For instance, the robot physicality also implies that we cannot simply translate design assumptions and practices borrowed from the software industry. For example, deciding to terminate interactions with, or use of a robot, is a different process when compared to uninstalling a software or a mobile app. Turning off a robot does not mean the robot no longer impacts the spaces wherein it is present. This can have various implications, ranging from concerns for data privacy to material sustainability. Likewise, some of the robots provide crucial physical assistance to those who have various impairments. Ceasing service of such system can mean the difference between a person"s ability to conduct activities of daily life by themselves and not.

Based on these, and other concerns stemming from the (physically) embodied and (partly) autonomous character of robots mean, we argue, robots extend the scope of (ethical) and responsibility related considerations beyond what has been addressed in the discourses about ethical and responsible AI.

References

1 Sostero, M. (2020). Automation and Robots in Services: Review of Data and Taxonomy (JRC Working Papers Series on Labour, Education and Technology 2020/14). European Commission, Joint Research Centre (JRC). `http://hdl.handle.net/10419/231346`.

4.2 The Role of Fairness in Responsible Robotics

Sarah Moth-Lund Christensen

Joint work of Kevin Baum, Helen Beebee, Simon Coghlan, Emily Collins, Kate Devitt, Michael Fisher, Masoumeh Mansouri, Sarah Moth-Lund Christensen

In philosophy, fairness has traditionally not been investigated as a moral concept in itself, but instead played second fiddle in relation to political philosophical concerns such as "distributive justice" [1]. In other words "fairness" has been utilised as a concept regarding the allocation of resources, opportunities, and so on across society. As such, it is not a concept easily and clearly defined in the literature. However, in recent years "Fairness" has been heavily deployed as a key term not as part of theoretical discussions on distributive justice, but instead in opposition to the rising concern regarding bias issues in machine learning models [2].

This working group aimed to broaden the terms of the "Fairness" debate, recognising that "Fairness" as a concept is not and should not be reduced to "Algorithmic Fairness". As such, the group set out to investigate what further role "Fairness" might play with regards to responsible robotics, and as such whether the concept of "Fairness" may be used to identify and illuminate issues regarding contemporary robotics design and deployment. The following inter-related points were identified as potential Fairness concerns relevant to Responsible Robotics design and deployment practices:

Algorithmic Fairness: Algorithmic injustice is an ongoing concern regarding the use of machine learning models. As machine learning can also be used in robots, the ongoing algorithmic injustice debate still relevant to discussion of fairness in relation to robotics.

Fairness and Design: Fairness may play a role even on the lowest level of design. Certain components may have poorer performance for certain demographics, while design choices such as language used can affect for whom the technology is useful for.

Fairness and Accessibility: On a larger scale, the deployment of robots in public and private domains can give rise to concerns regarding financial inaccessibility. As an example, consider a robot developed as a disability aid or as a teaching tool that is prohibitively expensive for the individuals or public schools that are in need of it. As such, societal or systemic structures can give rise to fairness concerns for the deployment of robots.

For completeness, it should be noted that the work group participants throughout the session discussed and expressed concerns on the merit of Fairness as a focal key concept in the Responsible Robotics debate, particularly in comparison to either more established political philosophical concepts such as Justice, or in comparison to other motivating concepts of well-established moral nature. Hence, a conceptual issue emerged regarding the notion of "Fairness", as the lack of current in-depth definitions results in ambiguity and lack of clarity when attempting to use "Fairness" as a guiding principle. As such, further formal investigation in the definition and dimensions of "Fairness" is needed in order to fully establish its full role in Responsible Robotics.

References

1 John Rawls. *A Theory of Justice.* Harvard, MA: Harvard University Press, 1971.

2 Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in Proceedings of Machine Learning Research. 81: 149–159., 2018.

4.3 Trust and Responsible Robotics

Nick Schuster (Australian National University – Canberra, AU), Hein Duijf (LMU München, DE), Nadin Kokciyan (University of Edinburgh, GB), and Thomas Michael Powers (University of Delaware – Newark, US)

Responsible robotics requires that robots, as well as the people and organizations who design and deploy them, are trustworthy. This goes beyond getting people to trust robotic systems, including their human elements, which might be accomplished whether or not people have good reason for trust (e.g. through effective advertising). Rather, robotic systems must satisfy certain independent normative standards in order to warrant trust. What are these standards, and what challenges must robotic systems overcome to satisfy them? We think that trust is importantly distinct from transparency, explainability, and predictability. While these qualities can make robotic systems trustworthy, trust seems especially important where robotic systems aren't, or can't be, made fully transparent, explainable, and predictable. Like human-human interactions, human-robot interactions can involve unavoidable uncertainty. Also like human-human interactions, human-robot interactions can take place in physical space (as opposed to cyberspace) and therefore make immediate physical harm a real and sometimes visceral possibility. These factors necessitate a trust that's structurally similar to trust between human actors: people often have to trust each other to behave appropriately without full knowledge of each others' motives, intentions, capacities, propensities, needs, vulnerabilities, etc. But robots are different from humans in relevant respects too. For instance, they don't share humans' basic interests in avoiding pain, injury, and death; they can't communicate with humans as humans can with each other; and they can be controlled by humans in ways that humans can't be controlled by each other. These factors pose distinct challenges for trust in robotic systems. A promising approach to clarifying and addressing such challenges would draw on social epistemology and moral psychology, to better understand what undergirds trust(worthiness) in general, as well as engineering and organizational studies, especially human factors, to explore possibilities for designing and deploying robotic systems such that people have good reason to trust them despite, or perhaps even because of, their peculiarities.

4.4 Predictability

Michael Milford (Queensland University of Technology – Brisbane, AU)

Joint work of Michael Milford, Nico Hochgeschwender, Dejanira Araiza-Illan, Alcino Cunha, Andrzej Wasowski, Yi Yang, Francisco J. Rodriguez Lera, Hein Duijf, Raja Chatila, Martin Magnusson

There has been some work on defining predictability in the context of robotics, and exploring its connection to other relevant properties, namely understandability. Although not entirely consensual, the common idea of all these definitions is that predictability is about matching the expectations of the user/observer and that predictability lies in a continuum (given a goal, a robot is as predictable as its chosen plan matches the expectations of the user/observer for that goal).

Full predictability might not always be a desirable property for all different users/observers (for a robot operating in public spaces, fully predictable behavior might open opportunities for observers to abuse/bully the robot), so a key responsibility at design time is precisely to identify the level of predictability that is adequate for each stakeholder. Predictability requires the user/observer to know the goal – the design should clarify how this will be achieved, either by designing the robot to also be understandable/legible, building single purpose robots, educating the users (this one a responsibility at deployment time...), etc. Not all users have the same expectations of what is the best plan, so responsible design for predictability should incorporate in the robot some mechanism for the robot to adapt to the individual users, so that (at least) predictability improves over time, or again level everyone by educating at deployment time...

Predictability is also a technical concept: regardless of the user/observer and the robot platform, task and domain, the extent and specificity with which a robot"s actions can be predicted also varies. For example, a large robot moving with substantial inertia through the environment – such as an autonomous truck – has a highly predictable set of next step possibilities – it will continue to move in the current direction at near the current velocity, possibly with the application of acceleration or braking changing its velocity. A human observer does not need to know anything about the algorithms or control systems for the robot in order to have broad predictability for the autonomous truck – it will likely continue on its current trajectory in the next moment, but may increase or decrease its velocity and its heading may change (initially not by very much).

Continuing the autonomous truck example, another key aspect of predictability is predicting the performance of the system. For autonomous vehicles, localization – knowing where the vehicle or robot is located – is a key estimation task that enables safe navigation and higher level behaviours. One aspect of the predictability of a localization system is predictability of how well it is performing – also relating to the concept of introspection. Imagine a choice of two localization systems: one that works well 99 % of the time but is unable to predict its failures that remaining 1% of the time, versus a second system that works well 95% of the time, but is able to predict when it is performing badly 95% of the time. An autonomous vehicle using the first system will unknowingly navigate using incorrect localization information 1% of the time: using the second system, this percentage drops to 0.25%, a major different for such a safety critical application. A side note relates to the research culture of robotics and related fields currently: one key issue with research in this domain currently is that the former system is much more likely to yield a top tier publication, despite the second system having far more utility for many end-user applications.

Participants

- Dejanira Araiza-Illan
Johnson & Johnson –
Singapore, SG
- Kevin Baum
DFKI – Saarbrücken, DE
- Helen Beebee
University of Leeds, GB
- Raja Chatila
Sorbonne University – Paris, FR
- Sarah Christensen
University of Leeds, GB
- Simon Coghlan
The University of Melbourne, AU
- Emily Collins
University of Manchester, GB
- Alcino Cunha
University of Minho – Braga, PT
& INESC TEC – Porto, PT
- Kate Devitt
Queensland University of
Technology – Brisbane, AU
- Anna Dobrosovestnova
TU Wien, AT
- Hein Duijf
LMU München, DE
- Vanessa Evers
University of Twente –
Enschede, NL
- Michael Fisher
University of Manchester, GB
- Nico Hochgeschwender
Hochschule Bonn-Rhein-Sieg, DE
- Nadin Kokciyan
University of Edinburgh, GB
- Severin Lemaignan
PAL Robotics – Barcelona, ES
- Sara Ljungblad
University of Gothenburg, SE &
Chalmers University of
Technology – Göteborg, SE
- Martin Magnusson
örebro University, SE
- Masoumeh Mansouri
University of Birmingham, GB
- Michael Milford
Queensland University of
Technology – Brisbane, AU
- AJung Moon
McGill University –
Montreal, CA
- Thomas Michael Powers
University of Delaware –
Newark, US
- Daniel Fernando Preciado
Vanegas
Free University of
Amsterdam, NL
- Francisco Javier Rodríguez
Lera
University of León, ES
- Pericle Salvini
EPFL – Lausanne, CH
- Teresa Scantamburlo
University of Venice, IT
- Nick Schuster
Australian National University –
Canberra, AU
- Marija Slavkovik
University of Bergen, NO
- Ufuk Topcu
University of Texas – Austin, US
- Andrzej Wasowski
IT University of
Copenhagen, DK
- Yi Yang
KU Leuven, BE