IEEE Robotics & Automation Magazine (RAM) paper, presented at ICRA 2026, Vienna, Austria. Cite as RAM paper.

A Roadmap for Responsible Robotics

Dejanira Araiza-Illan, Kevin Baum, Helen Beebee, Raja Chatila, *Senior Member, IEEE,* Sarah Moth-Lund Christensen, Simon Coghlan, Emily Collins, S. Kate Conroy, Alcino Cunha, Anna Dobrosovestnova, Hein Duijf, Vanessa Evers, Michael Fisher, *Corresponding Author*, *Member, IEEE,* Nico Hochgeschwender, Nadin Kökciyan, Séverin Lemaignan, Francisco Rodriguez-Lera, *Member, IEEE,* Sara Ljungblad, Martin Magnusson, *Member, IEEE,* Masoumeh Mansouri, Michael Milford, *Senior Member, IEEE*, AJung Moon, *Member, IEEE,* Thomas M. Powers, Pericle Salvini, Teresa Scantamburlo, Nick Schuster, Marija Slavkovik, Ufuk Topcu, Daniel Vanegas, Andrzej Wąsowski, *Member, IEEE,* Yi Yang, *Member, IEEE*

Abstract

This document presents the outcomes of the Dagstuhl Seminar *Roadmap for Responsible Robotics*, held in September 2023 at the Leibniz Centre for Informatics, Schloss Dagstuhl, Germany. The seminar brought together researchers from Robotics, Computer Science, Social and Cognitive Sciences, and Philosophy with the aim of charting a path towards improving responsibility in robotic systems. Through intensive interdisciplinary discussions centered on the various values at stake as robotics increasingly integrates into human life, the participants identified key priorities to guide future research and regulatory efforts. The resulting roadmap outlines actionable steps to ensure that robotic systems co-evolve with human societies, promoting human agency and humane values rather than undermining them. Designed for diverse stakeholders—researchers, policymakers, industry leaders, practitioners, NGOs, and civil society groups—this roadmap provides a foundation for collaborative efforts toward responsible robotics.

I. Introduction

Just as Artificial Intelligence (AI) systems have now influentially entered the public arena, robotic systems, too, are set to become increasingly relevant in society. It would, however, be easy to misunderstand robots as merely physical embodiments of AI. Consequently, "responsible robotics" could be misconceived as a straightforward extension of the current lively debates around "responsible AI." In this roadmap, we aim to complement these rich discussions by highlighting the sociotechnical challenges unique to robotic systems. We begin by situating robotics-specific issues within the broader context of AI. Then, we underscore the importance of the robotics community proactively engaging with responsible robotics to help shape a positive future with robots. We present the roadmap in Sec. V in the form of the main gaps identified and summarized in three tables. We map them onto the challenges discussed in Secs. II–III and attempt to identify which stakeholders might fill the gaps.

A. Robotics

Robotics refers to a diverse set of products, technologies, and sub-disciplines, many of which are not powered by AI or sophisticated, intelligent software systems (though some are or will be). A robot is defined as "a programmed actuated mechanism with a degree of autonomy to perform locomotion, manipulation or positioning" [1]. Robotic systems, unlike purely algorithmic systems such as Machine Learning (ML) models, span a multitude of physical morphologies, a wide range of mechanical designs, the spectrum of autonomy (i.e., from teleoperated to fully autonomous), as well as varieties of "intelligence." While some sophisticated systems such as autonomous vehicles make use of AI (for instance for pedestrian detection and vehicle steering), many other robotic systems, such as a robotic eating aid device or industrial robots in factories, may not. All systems require us to consider issues unique to their appearance and movement in addition to issues of use and other experiences, which may include responsible AI, data ethics, and other domain-specific ethics considerations.

D. Araiza-Illan is at Johnson & Johnson, Belgium; daraizai@its.jnj.com. K. Baum is at DFKI, Germany; kevin.baum@dfki.de. H. Beebee is at University of Leeds, UK; h.beebee@leeds.ac.uk. R. Chatila is at ISIR, France; Raja.Chatila@isir.upmc.fr. S. Moth-Lund Christensen is at University of Sheffield, UK; s.m.l.christensen@sheffield.ac.uk. S. Coghlan is at University of Melbourne, Australia; simon.coghlan@unimelb.edu.au. E. Collins is at University of Manchester, UK; e.c.collins@manchester.ac.uk. S. K.Conroy is at Queensland University of Technology, Australia; skateconroy@gmail.com. A. Cunha is at Universidade do Minho, Portugal; alcino@di.uminho.pt. A. Dobrosovestnova is at IT:U, Austria; anna.dobrosovestnova@it-u.at. H. Duijf is at Utrecht University, Netherlands; h.w.a.duijf@uu.nl. V. Evers is at CWI, Netherlands; Vanessa.Evers@cwi.nl. M. Fisher is at University of Manchester, UK; michael.fisher@manchester.ac.uk. N. Hochgeschwender is at University of Bremen, Germany; nico.hochgeschwender@uni-bremen.de. N. Kökciyan is at University of Edinburgh, UK; nadin.kokciyan@ed.ac.uk. S. Lemaignan is at PAL Robotics, Spain; severin.lemaignan@pal-robotics.com. F. Rodriguez-Lera is at Universidad de León, Spain; fjrodl@unileon.es. S. Ljungblad is at Chalmers University of Technology, Sweden; Sara.ljungblad@chalmers.se. M. Magnusson is at Orebro University, Sweden; martin.magnusson@oru.se. M. Mansouriis at University of Birmingham, UK; m.mansouri@bham.ac.uk. M. Milford is at Queensland University of Technology, Australia; michael.milford@qut.edu.au. A. Moon is at McGill University, Canada; ajung.moon@mcgill.ca. T. M. Powers is at University of Delaware, USA; tpowers@udel.edu. P. Salvini is at University of Oxford, UK; salvini.pericle@gmail.com. T. Scantamburlo is at University of Trieste, Italy; teresa.scantamburlo@units.it. N. Schuster is at University of Georgia, USA; nick.schuster@uga.edu. M. Slavkovik is at University of Bergen, Norway; Marija.Slavkovik@infomedia.uib.no. U. Topcu is at University of Texas, USA; utopcu@utexas.edu. D. Vanegas is at VU Amsterdam, Netherlands; d.f.preciadovanegas@vu.nl. A. Wąsowski is at University of Copenhagen, Denmark; wasowski@itu.dk. Y. Yang is at KU Leuven, Belgium; yi.yang@kuleuven.be.

'Robotics' can also be understood in different ways and with varying scope. We see it as a very broad field, encompassing researchers in various disciplines, businesses, engineers, regulators, and end users. Responsible robotics should also consider the interest of a range of stakeholders, including whole societies, the natural environment, as well as animals [2], [3].

B. Responsibility

Considering the social, ethical, and legal implications of the field is not new to the robotics community. The term "roboethics" was coined two decades ago, and the European Robotics Research Network published a roboethics roadmap in 2007 [4]. The community continues to make steady progress discussing ethical dimensions of robotics. These topics were regularly included in leading robotics conferences, such as the *International Conference on Robotics and Automation*, and many articles on normative issues have appeared in robotics publications, including the *Handbook of Robotics* [5]. Over time, this discourse has become increasingly interdisciplinary, not only through a number of EU and UK projects,[1] but also through moving towards standards and guidelines. For example, the IEEE published "Ethically Aligned Design", a guideline that consolidates global expertise of over 250 experts, and the British Standards Institution published its standard, BS8611: "Guide to the ethical design and application of robots and robotic systems" [6], [7]. The present work aims to contribute to this progress, drawing on multidisciplinary research expertise. It focuses primarily on elucidating open research questions relevant to ensuring responsibility in robotic systems given the current state of the art.

C. Responsible Robotics

What are *responsible robotics*? The term *responsibility* is ambiguous both in philosophy and in law. According to Santoni de Sio and Mecacci, this complexity is rarely reflected in the debates on responsibility for the behavior of AI and robotic systems [8]. They point towards several responsibility gaps: the culpability gap, the moral accountability gap, the public accountability gap, and the active responsibility gap. Such gaps arise from different kinds of sources, including but not limited to technical, social and legal ones. Thus, responsibility gaps often require complex solutions.

Responsible robotics is the idea that various parties involved in development, deployment, integration, usage and maintenance of robots need to act in a responsible manner toward all stakeholders. This involves behaving ethically in their roles, making ethically sensitive design and deployment decisions, and ultimately taking responsibility for how robotics as a field progresses and how robots are used. It is generally clear that responsible robotics is really about *human* responsibility in this field, and not about a possible attribution of responsibility to the machines themselves [9]. It is also crucial to consider various interconnected dimensions of responsibility, including *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; and *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 and animals. These dimensions interact to form a comprehensive framework of responsibility in robotics, ensuring that each party involved, from designers to end-users, upholds their respective duties in promoting good robotic practices. This document identifies a core (but not exhaustive) set of features of responsible robotics.

D. Limitations and Outlook

We acknowledge that the current group of authors predominantly represents perspectives from the Global North and WEIRD (Western, Educated, Industrialized, Rich, and Democratic) research contexts. The contributors to this report are researchers from robotics, philosophy, human-robot interaction, software engineering, and artificial intelligence, among others. While we have an interdisciplinary outlook, some aspects, e.g., the legal aspect are not covered. This expertise provides valuable insights, but it also introduces the risk of a perspective skewed toward specific research interests. We highlight the urgent need to incorporate a broader range of viewpoints and greater diversity in future iterations or reformulations of similar roadmaps. This roadmap is not the end, but rather the beginning. Our goal is to spotlight and exemplify some critical issues, foster the advancement of responsible robotics, and facilitate our collective journey through the intricate sociotechnical landscape ahead.

II. Principles

A. Contrasting Robotics with AI

There is no universally agreed-upon definition of what a robot is (and is not). We establish a pragmatic working definition that is followed in the rest of this article. The ISO robotics vocabulary [1] distinguishes between robots and other automated systems, describing a robot as 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 specifically implies that a robot is first and foremost a physical piece of machinery, which excludes software-only systems as not robots; e.g. purely-software bots, voice assistants, large language

[1] EthicBots: https://cordis.europa.eu/project/id/17759/de; REELER: https://responsiblerobotics.eu; SIENNA: https://www.sienna-project.eu/robotics; TechEthos: https://www.techethos.eu; REMARO: https://www.remaro.eu, RoboTIPS: https://www.robotips.co.uk; Verifiable Autonomy: https://doi.org/10.1109/JPROC.2019.2898267

models, or image recognition. Robots are embodied in the real physical world [10]. Furthermore, robots and robotic devices have some degree of autonomy. Mobile machinery that only follows pre-programmed instructions without coupling to the environment (e.g., 3D printers) does not qualify as a robot.

While many ethical and responsibility implications overlap between robots and AI, the embodied and autonomous nature of robots brings a host of new considerations. An obvious issue arises from the physicality of a robotic system: the physical safety of people and animals. In the case of a mobile pizza delivery robot, this includes the safety of the natural environment (see *Sustainability* below), as well as the safety of any bystanders and encountering people (e.g., pedestrians including people with strollers, children on their own, and people with mobility issues, sharing the sidewalk with the mobile delivery robot). Some of the issues are directly related to the programmable and scalable use of kinetic force, for example in cobots. Further concerns arise from how humans react to moving artifacts in our physical environment. The fields of human-robot interaction (HRI) and social robotics (SR) are rich with examples of how different designs (e.g., morphology, nonverbal and verbal behaviors) affect human perception, beliefs, and decisions. For instance, robot embodiment coupled with autonomous movement and perceived goal-orientedness is known to elicit a tendency to respond and treat robots as quasi-social actors; for example by adopting an intentional stance towards them [11].

This tendency is further strengthened by so-called social robots being designed to increasingly look and behave like humans. The potential dangers of designed and perceived robot sociality have already been discussed in roboethics and related fields [12], [13], [14]. Deception, unilateral bonds, and the ability of robots to reshape affect- and relationality-laden practices have been named as concerns (e.g., when robots are deployed in service sectors). Robot physicality also implies that we cannot simply borrow design assumptions and practices from software science. Deciding to terminate interactions with a robot is not as simple as uninstalling a mobile app, and turning off a robot does not mean that it no longer affects the space in which it is present. This means that ethics analysis of a robotic system, regardless of the amount of AI integrated into it, must go beyond simply assessing the software onboard the robot to also include broader considerations.

Because robots are situated in the real world, they are subject to its real-time dynamics and constraints, as well as its uncertainties, for their perception, action, decision-making and reactivity. While they may benefit from prior training on large datasets, they must also build new understandings and capabilities during action. Adaptable robots should be able to improve their actions over time, for example through reinforcement learning. They are subject to computational complexity limits and related challenges from power and computer hardware limitations, which are stricter than those in a typical machine learning system.

B. Foundations

Roles and Agents: From companies to university and government agencies, responsibility for the ethical and effective use of robotic products and services rests on multiple actors. An important step to ensure responsible robotics is to explore the diverse roles and responsibilities of key stakeholders affecting or being affected by robots.

Universities (and our specific stories of robotic ideas) play a crucial role in shaping professionals who design, engineer, and operate robotic systems. Without ethics education and taking responsibility, we are at risk. By ensuring that future engineers feel a sense of responsibility for the systems they design, universities can significantly contribute to the development of ethically grounded technology professionals. Engineering and design curricula should include studies of stakeholder needs, alternative solutions, responsible design and innovation, safety standards, and potential consequences of misuse and abuse. This could be done by intensifying the dialogue and collaborations with other disciplines, following promising initiatives such as Embedded EthiCS [15].

To align robotics products and services with ethical standards and societal well-being, companies must conduct thorough risk assessments, addressing potential misuses and abuses and implementing safeguards in their products. For example, for 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 [16]. Additionally, they should develop 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.

Governments play a pivotal role in creating and enforcing regulations for robotic products and AI services [17]. 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 alternative solutions and their effects, mitigating the potential for misuse or misunderstanding.

Values: When people plan for robotic systems to become more capable and autonomous, and integral in human life, there is a risk that the process becomes technology-driven, and that these systems — and the people and organizations developing, deploying, regulating, and using them — operate outside appropriate ethical constraints [18]. We understand "responsible robotics" as an effort to capture such perspectives and share it within the community and in society. Taking responsibility for one's actions is central to being a morally good agent, as is holding others responsible for theirs where appropriate. This

involves more than the attribution of causal responsibility; it has to do with what one should and should not do or cause to happen. Our primary focus is on human moral responsibility for the behavior of robotic systems. A robot is, after all, an embodied being to which we can attribute direct causal responsibility for its behavior in the physical world. By "responsible robotics," then, we refer to this entire nexus, ideas and dreams of people and society being supported by robotic products in various ways. If responsible robotics means designing systems that respect human rights and core humanistic values, a suggestion of what these rights may include are:

Dignity: The inherent worth of each and every member of society who stands to be impacted by robotic systems must be respected. (All have the right to be treated with dignity.)

Autonomy: Supporting human beings to act in accordance with their own interests and aspirations, both individually and collectively.

Privacy: Respecting that children and adults need to protect sensitive information about themselves, and share it only with certain other individuals or organizations as they see fit.

Safety: Not exposing people to robotic systems that pose serious threats to their lives, health, and well-being.

Responsible robotics will also promote certain values, including:

Trust: Those who stand to be impacted by robotic systems should have good reason to believe that these systems are aligned with their own legitimate interests.

Justice/Fairness: As competing moral interests exist, broadly impactful robotic systems must behave in ways that all affected parties have sufficient reason to accept, even when their own interests are overridden.

Accountability: The relevant agents can be held accountable for adverse outcomes of robotic behavior, especially when justice/fairness requires a robotic system to override the legitimate interests of some in favor of those of others.

Sustainability: Among the most pressing adverse outcomes of robotic systems is the degradation of life-sustaining ecosystems. This includes environmental, social, and economical sustainability. Most societies also recognize ethical limits to what may be done to sentient animals.

The above values imply these further requirements:

Transparency: Any available information about robotic systems (and their behavior) that bears on human assessment of them should be accessible to all relevant parties.

Understandability: Information about why these systems behave as they do in particular situations should be available and presented in ways stakeholders can understand.

Predictability: Any information necessary for anticipating how robots will behave in immediate and future situations should be made available to all relevant parties.

While our main objective in the Dagstuhl Seminar was not to produce a comprehensive list of values and principles for achieving responsible robotics, there was a rich discussion on what ethical values and other requirements are important and should be considered by the community. The values listed above receive widespread assent within the group, without the effort to distinguish them as exclusive of one another, or to acknowledge priority or hierarchical ordering of the values. We have selected only a few of the many possible values, which are those we discussed in depth. Fairness and trustworthiness, in particular, are considered to be core values in responsible AI. We highlighted how consideration of the same values can lead to sometimes drastically different, additive, and unique insights for responsible robotics. The values discussed here should not be considered as exhaustively covering the full range of responsible robotics issues. Rather, they are intended to provide a general framework that is subject to ongoing discussion. We discuss key values in more detail below.

III. Responsibilities

Responsibilities affecting the research, design, deployment, and use of robots are held by people in different roles, such as governments, funding agencies, researchers, industry and citizens. Responsibility is relevant across the whole product life cycle of a robot, from design and development to deployment and subsequent operation, and to robot recycling or waste. This is related to social, environmental, and economical sustainability. The responsibility values, relevant at deployment time, should be considered from the design stage onwards, with continual reflection on future events, and consideration of how to mitigate future risks based on what those reflections propose. For example, a responsible deployer will reflect on responsibility values at the design stage to mitigate issues that could arise at deployment and beyond.[2] Developers should consider ethical aspects, such as trust, justice, fairness, and accountability, during development. The formalization of these concepts during development can be beneficial to implementing responsible robotic applications.

[2]For a more comprehensive discussion of the relationship between terms related to trust as they mediate various stakeholders at different stages of the life cycle of a robot, see [19].

A. Trust

There is a risk of people overly trusting robotic and other technological systems, whether explicitly or subconsciously. Researchers should contribute a realistic, nuanced perspective of the concept of trust, and how it is related to the design, adoption, and use of robots. Trust can be said to overlap with many of the responsibilities that are considered below. We can trust that a robot system is *just*, *fair*, or respects *privacy*. Trust is also related to *reliability* (trusting that the robot will 'do its job') and *understandability* (knowing why the robot is doing such-and-such). Finally, trust can be analyzed through the lens of autonomy, and we recognize that there is a growing interest in *trusted autonomy* and the requirements for ensuring trust when robots act autonomously [20]. These conceptual and technical issues will all be furthered by work on responsible robotics. We focus now on dimensions of trust that are orthogonal to these.

Should we trust that a robot *works in our interests*? The answer depends not only on the robot but also on whom we ask. For example, factory owners and assembly workers may have starkly different opinions about trust in an industrial robot system. For autonomous road vehicles, the answer may vary among owners, passengers, fellow drivers, and pedestrians; whether it concerns the efficient routing of traffic, or reacting to a dangerous situation. A continuous *hard minimization framework* may be necessary to balance these competing factors. For social or service robots, such as in hospital settings, trust may vary between doctors, administrators, and patients. It is important to operationalize trust in a variety of contexts. How does one measure trust? Are direct measurements of trust possible? If measurements of trust are to be merely indirect, will we have sufficient confidence in them to know that modifications in an operational context are indeed increasing trust?

B. Justice / Fairness

Like other technologies, robots can create issues of justice and fairness. The questions regarding robotics, fairness, and justice often incorporate aspects seen in the debates on AI ethics and algorithmic injustice. But again, the physicality of robots plays a role.

Design and Production: For a responsibly created robot system, not only must the technical challenges be addressed, but also the potential fairness implications. Many possible questions arise: What are the quality and functionality of the parts chosen for the system and how do they impact a range of different people or groups? Are there unjust working practices in corporations that develop robots? Are the materials chosen for the design ethically sourced, or does it rely on, for example, exploitative mining practices that affect disadvantaged communities?

Deployment, Upkeep and De-commissioning: Later phases of the robot system lifecyle also raise justice questions. For example, during deployment, various stakeholders might be subject to unfair outcomes—not just consenting users of the system, but potentially others too. Intentions for just and fair systems can easily be derailed by corporate and social structures. Consider a system developed to relieve its users of having to deal with menial tasks, where purchase and upkeep costs mean that the system is only accessible to a certain group. Or consider how a cleaning robot intended to perform menial tasks displaces jobs vital to the well-being of some individuals. Such imbalances must also be considered, especially when design and development of a given robotic system primarily occurs in one — a Western, educated, industrialized, rich and democratic — context, while the deployment and the affected stakeholders are from outside this social and cultural setting.

Another key question is whether a given system is morally or socially sustainable in its upkeep and de-commissioning environment. For an average robotic system, the above process is not merely a one-and-done activity. It is rather an ongoing repeated cycle, where experiences from deployment push the vision of the design, leading to continuous alterations and updates to the system. Incorporating fairness and justice into a system design is, therefore, not a single check list completed at the start of the development phase, but *requires continuous reflection upon the system's design, deployment and interaction with its environment, and its subsequently accompanying action through any updates or enhancements to the system.*

C. Accountability

Ensuring accountability throughout the entire lifecycle of a robot is essential to guarantee responsible design, deployment, and long-term operation. Accountability, in this context, refers to the capacity to trace decisions, actions, and changes back to specific actors, and to ensure that each phase of the robot's life is governed by clear responsibilities and verifiable processes. This lifecycle can be structured into five key phases: (1) design and manufacturing by the original manufacturer; (2) preparation for deployment in real environments, possibly involving third parties; (3) hardware and software updates; (4) real-world operation; and (5) post-deployment maintenance. This section aligns with these phases and describes how accountability must be established and maintained across technical, procedural, and ethical dimensions, as illustrated in Fig.1.

In the early stages (1) the main actors – the designers, developers and integrators – are accountable for design and ethical issues not directly related to technical challenges, for example the contextual or political issues. Why, where, and by whom should a robot prototype be created at all? They are also accountable for thoroughly understanding the requirements and objectives set by the client that will be involved with the robot. Moreover, they must adhere to the principles of relevant ISO standards or certifications, such as ISO 12100 and ISO 13482:2014, for all hazards identified in the application, ensuring: a) an inherently safe design; b) the necessary protective measures; and c) any required information for responsible use.

They are responsible for correct hardware design, the right selection of materials, and simulation-based evaluations of mechanical design and dynamics. They must also define the software architecture, including selection of local and cloud

assets and services, including their provenance (what is the source, creator, provider and its long term sustainability). They must ensure the documentation of hardware, software, manuals, goals, and both predicted and emerging functionality. Finally, they are accountable for task-based risk assessment definition under expected scenarios and contexts.

It is recommended to perform a verification of all processes involved during the prototype release by an accountable independent organization, assessing both internally developed and third-party hardware and software components (2). This verification should include at least these seven structured procedures: *Visual and acoustic inspection* should be conducted using senses such as sight and hearing, without specialized inspection tools, both when the robot is turned off and in operation. *Practical testing* of the robot prototype and the associated equipment should be carried out under normal and abnormal conditions, ensuring that the test data is stored and made available. *Measurement comparison* involves comparing actual values of the robot prototype characteristics with the specified limits from the outset, while also recording the process and the logs. *Diagram examination* entails a structured review or walk-through of circuit diagram designs and layout drawing designs, including electrical, pneumatic, and hydraulic systems, along with the related specifications. *Software examination* consists of a structured review or walk-through of the software code and related specifications, which may include code inspection and software testing. A *task-based risk assessment review* should also be performed, including a structured walk-through of the risk analysis, risk estimation, and proper documentation. Finally, a *document examination* must be carried out, consisting of a structured review or walk-through of relevant documents related to the robot prototype. These seven verification procedures collectively contribute to ensuring that the robot prototype meets the required safety, functionality, and reliability standards before deployment in a public space.

It is necessary to consider whether the robot hardware will be updated or enhanced after its release, or even beforehand (3). If so, by whom – only by the manufacturer, or also by third parties, such as integrators? Will final users have the ability or authorization to make modifications? Similarly, it must be clarified whether the robot software will be updated or enhanced, and again, by whom. In the case of (autonomous) vehicles, for instance, over-the-air updates can significantly change the vehicle's driving behavior.

Furthermore, some questions should be answered in order to establish that the emerging robot behavior is accountable (4 and 5). Any hardware and software change should be accountable. Any change in remote software should be accountable and verifiable. It should be defined in design steps and avoided if it implies risks.

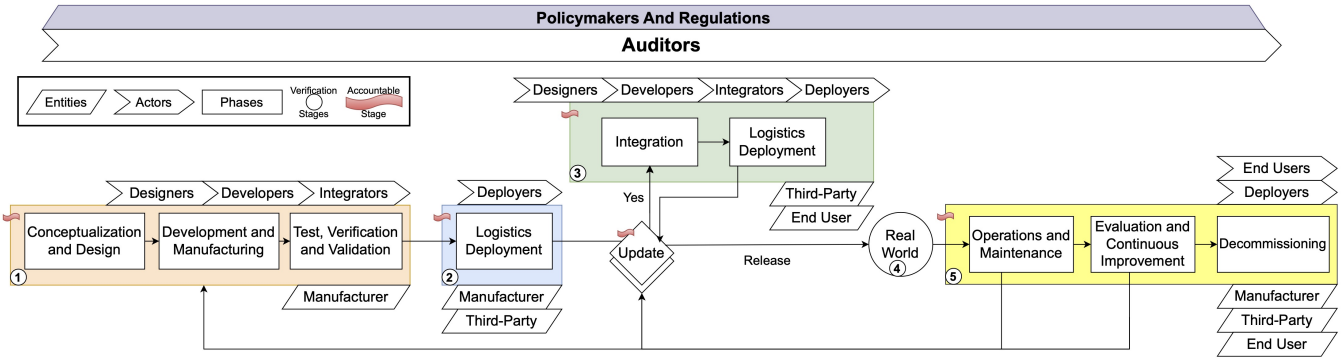

Fig. 1. Phases of Robotics Development and Deployment Lifecycle

D. Environmental Sustainability

Following the outline of the British Standards Institution's Guide to Sustainable Robotics, we consider the environmental sustainability of robotic systems in different life cycles of a robot.

Production and Deployment: How environmentally sustainable are the materials used to build our robots? Where do they come from, how resource intensive are the fabrication processes that are used? What are the environmental aspects of deployment, such as transporting the robot to its target area, and environmental damage of the robot deployment? For example, what if an underwater robot crashes and spreads a large amount of battery cells in a fragile biological ecosystem?

Energy, Software, and Communication: Sustainability issues are particularly important once the robot has been deployed and is working normally. Where does the energy for continued operation come from? And is the energy consumption environmentally justified? Robot software is a significant energy consumer. Is any use of resource-intensive data-driven machine learning justified? Where does this usage occur, at the "edge" or within a cloud server? Is the amount of data collected, stored, and transferred justified? Will this storage be required to continue growing forever? Issues of green AI [21] are relevant here.

Waste, Repair, and Decommissioning: Even without failures, the normal operation of robots might lead to large amounts of waste. When a failure occurs this waste can increase dramatically. An example is the debris of "dead" satellites littering low-Earth orbits, and, consequently, the new measures ensuring that newly proposed satellites have a de-orbit procedure

factored into their design. What if the robot fails? Is it repairable? Issues such as the "right to repair" and the environmental impact of maintenance and repair are important, as is considering "end of life" plans and ways to extend the life of the robot as early as possible. Modularity, interoperability, and standardization are crucial to many of these elements. Different sustainability issues might come to the fore depending on the length of life of the robot. Long-living robots need to be robust and their materials potentially recyclable or interchangeable. Short-living robots can be made of sustainable materials (e.g., cardboard).

E. Privacy

Technology used in robots can easily become active trackers. Indeed, the intended application of many robots is surveillance. Examples of actively tracking products are not only surveillance drones, but also delivery robots, cars, aerial drones, cobots in a factory setting, and sex robots that can monitor people's actions meticulously. Crucially, robots could affect or actively track bystanders, not only their deployers (e.g., cars that automatically film people). It is important to consider diverse people and their needs — including those of children and members of vulnerable groups — both when these people are intended as potential users as well as when they are not (see [22], [23]). Understanding the legal protection for users and bystanders as well as practical ways of obtaining consent remain challenges for these new technologies.

Responsible design should recognize that robots are machines equipped with perception mechanisms — data collection devices that are connected to networks and data storage systems. For example, there has been an incidents of footage of a person using the bathroom taken by a robot vacuum being spread on social media by people at the robot company. People using vacuum cleaning robots, or robot toys, may not be aware that the robot may be collecting data in their home. Thus, for any robotic system deployed in environments where humans are, its responsible design should follow the privacy-by-design principles [24]. This involves proactive rather than remedial design of data collection and handling, making privacy the default choice, inventing design strategies that simultaneously achieve the desired functionality and respect privacy, ensuring end-to-end security, maintaining transparency, and prioritizing user-centered design. Designers need to prototype and test: Are there alternative solutions? What kind of data collection can be avoided? How may it be used? What sensors are not necessary? Is the choice of sensors privacy neutral? Eick and Anton recommend conducting privacy impact assessments over the robot life cycle, documenting what data are collected and shared and how they are secured [25]. Taras and colleagues propose a system whereby optical and analog processing of camera data precedes any digital processing so that private image data can never be captured by the system while it can still be used for visual localization [26].

The physical presence of a robot in the same space as humans nuances the above requirements compared to other AI systems. Being embodied opens tracking possibilities at many new situations beyond what static devices, mobile phones, or web-browsers do. The actual privacy and the perceived levels of privacy protection will often not be the same, as subjects may be watched. The violation of privacy may happen already in situ, not just via data collection, but also through the robot's physical presence in real time. This is unlike for most other AI systems that often deal with post-collection data, non-real-time privacy treatments. Robots pose similar problems to other physical devices such as smartphones, smart TVs, and other IoT systems. However, robots — especially if displaying autonomy and agency — exacerbate the problem further as they not only can collect data more actively and purposefully, but also can endanger the spatiotemporal experience of privacy. For this reasons, active real-life privacy related signaling is important and remains an open challenge.

F. Safety

The broader AI research and policy communities have raised concerns under the umbrella of AI Safety, both in relation to the potential development of Artificial General Intelligence (AGI) [27] and the more immediate implications of deploying General Purpose AI (GPAI) systems [28], [29]. AGI refers to the idea of highly autonomous systems with some cognitive abilities comparable to or exceeding those of humans, raising questions of long-term risks such as value misalignment, loss of control, or unintended behaviors. GPAI, by contrast, designates adaptable AI systems—like large language models—that can be repurposed across diverse tasks, and whose widespread deployment poses regulatory challenges related to transparency, accountability, and societal impact.[3]

While these concerns differ from the physical and operational safety traditionally addressed in robotics [30], [31], [32], [33], they share a common foundation: ensuring that intelligent systems, whether embodied or disembodied, act predictably, align with human values, and operate safely in complex, real-world environments shared with humans.

We can define safety as *no harm being made to human subjects, to the environment and certain animals, and valuable or critical infrastructure*. In this respect, responsible robotics must adhere to the rigorous safety standards long established in traditional engineering domains, such as industrial automation and machinery design. However, safety can concern different matters for different types of robots and their contexts. For example, flying robots such as drones or unmanned aerial vehicles (UAV) involve a different type of safety concern than a lawn mower robot. Therefore, safety not only means reliable behavior and trustworthiness (doing what it is expected to), but the whole process of designing, programming and deploying specific use with specific expectations. It requires constant assessment to understand how safety translates to the context of ever emerging robotics application domains (e.g., domestic robots, care robots, field robots, service robots, etc.).

[3] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

If these robots become increasingly autonomous, connected, and based on GPAI, it is equally important to address cybersecurity as a core component of responsible design. Cybersecurity breaches can lead to indirect but severe safety consequences, including loss of control, unauthorized behavior, or data leaks that compromise human dignity and privacy. Recent research highlights the need to consider both safety and cybersecurity in an integrated manner, particularly in cognitive social robots that operate in close interaction with people [34].

Although many safety discussions around intelligent systems have focused on areas such as language technologies or decision-making software, similar concerns arise when these systems are used in robots. Problems such as a mobile robot not doing or being used as its designers intended, causing side effects during task execution, or behaving unpredictably in unfamiliar environments can all have serious consequences once the system interacts with the physical world [35]. These challenges are often more visible and immediate in robotics, where mistakes may affect people directly. At the same time, recent research emphasizes the need to look beyond technical performance and consider how safety is shaped by social expectations, everyday use, and the perspectives of those affected by the system [36]. This broader understanding is especially important for robots that operate alongside people in public or private settings.

G. Predictability

There has been work on defining predictability in the context of robotics, and exploring its connection to other relevant properties, namely understandability. The common idea of all these definitions is that predictability is about matching the expectations of the user or observer. Furthermore, the predictability lies in a continuum; given a goal, a robot is predictable if its chosen plan matches the expectations of the user, observer for that goal. It may be less predictable, if the goal is unknown to the observer.

As predictability requires the user to know the goal, it becomes a design objective to clarify what is the goal and how it will be achieved. This can be achieved by designing the robot to also be legible, building single purpose robots, and educating the users. Not all users have the same expectations of what is the best plan, so responsible design for predictability should incorporate mechanisms for the robot to adapt to the individual users, which allows the predictability to improve over time.

Full predictability might not always be a desirable property for all different users and contexts. For a robot operating in public spaces, a fully predictable behavior might open opportunities for observers to abuse of the robot, so a key responsibility at design time is precisely to identify the level of predictability that is adequate for each stakeholder.

Predictability is also a technical concept that can compensate for lower performance. Regardless of the observer and the robot platform, task and domain, the extent and specificity with which a robot's actions can be predicted also vary. 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. The truck will likely continue on its current trajectory in the next moment but may increase or decrease its velocity and its heading may change.

For many systems predicting the future (or near-future) performance is essential. For autonomous vehicles, localization—knowing where the vehicle or robot is located in space—is a key estimation task that enables safe navigation and higher-level behaviors. 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 90% of the time but is able to predict when it is performing poorly 99% of the time it is actually failing. An autonomous vehicle using the first system will unknowingly navigate using incorrect localization information 1% of the time. When using the second system, this percentage drops to approximately 0.1%, a seemingly minor but in reality a major difference for such a safety critical application. Unfortunately, the former system is much more likely to yield a top tier publication in the current robotics research publication landscape, despite the second system having far more utility for many end-user applications.

H. Understandability

Related to predictability, there has been some work on defining understandability in the context of robotics. These definitions vary slightly, but the key common idea is that understandability (or legibility) is about conveying the intent of an embodied artificial actor. Like predictability, understandability lies in a continuum: a robot is as legible as its chosen plan enables the user to confidently infer its goal. As with predictability, understandability might not always be a desirable property for all the different users, so a key responsibility at design time is precisely to identify the level of understandability that is appropriate for each stakeholder. Similarly, not all users have the same mental model of how a robot operates, so responsible design for understandability should incorporate some mechanism for the robot to adapt to the individual users, so that (at least) understandability improves over time. Finally, over-engineering of technical solutions to convey intent should be avoided. Often, simple solutions like making the robot signal or verbalize what it is going to do next are better than trying to convey intent indirectly by choosing a specific motion plan.

I. Transparency

Transparency in the design and development process and transparency in the robots themselves are both crucial to the responsible design and development of robotics. Concerning the former, we need to know what the designers intended, how

they developed and trained their software, and what safeguards were put in place. Concerning the latter, the transparency here is less about "seeing the code" and more about transparency of behavior and transparency of intent in physically embodied systems. This second element is a core part of the IEEE P7001 standard [37] especially as we wish to be able to ask (autonomous) robots questions such as "why did you do that?" [38] and expect a clear and truthful answer. This is particularly important for the trustworthiness of any robot.

J. Dignity

Concepts such as social dignity provide essential guidelines for ethical care giving [39], [40]. There is a need to recognize individuals' dignity in human-robot interactions in caregiving contexts. Caregiving robots must respect dignity by acknowledging humans as vulnerable beings with needs, autonomy, and rationality [41]. Furthermore, the formation of relationships between robots and humans is considered crucial, relying on reliability and social trust [42], [43]. Unlike other technological artifacts, social robots can establish quasi-social relationships with users, invoking social recognition and empathy to foster meaningful interactions [44]. These interactions are vital for preserving the sense of agency and dignity in those receiving robotic care.

Dignity is also related to universal economic rights and justice, building on peaceful and respectful interaction between people [45]. The idea of using robots in different contexts and for different purposes raises questions of their use in different practices. There are several destructive industries, such as the sex industry and war. Sex robots have been regarded to perpetuate humans as commodity, with the risk of amplifying and increasing the trafficking industry [45]. Concerns about the impact of autonomous weapons systems on human dignity have led to calls for their prohibition [46]. The debate on the morality of autonomous robots also encompasses military applications, where some advocates highlight their tactical advantages, suggesting reduced risk to human lives [47]. However, it is all our shared responsibility, to prevent and handle conflicts and stop war. War is a breakdown of law and order, and a destruction of civilized society and dignity. Industry plays an important role in amplifying moral decisions and their effects towards peace. It is questionable that some actors in the robotic industry are selling robotic products with privacy issues in civil society (e.g., robotic vacuum cleaners with video streaming) and robotic weapons used in war. The world needs sustainable peaceful societies and sufficient moral rather than moral blindness, also in the area of robotics.

IV. Application areas

Based on the responsibilities listed in Sec. III, we now discuss some examples of how they apply in a set of common use cases of robotics. Self-driving robots with various degrees of autonomy are already used heavily in *logistics and transportation* in warehouses and factories, and there is a lot of development in academia as well as industry striving to make such robots more agile and flexible, safe and understandable, robust and dependable. Trust in this setting comes down to workers trusting that the robots are working in their interests (on one hand) and trusting that they perform their tasks efficiently and reliably (on the other hand). Accountability issues are easier to handle in these mostly controlled industrial settings than in many other applications of robots, as well-established procedures and legislation are already in place for assigning responsibility in case of equipment failure, accidents or disruptions, whether they arise from human error, system malfunction, or negligence. Interestingly, in some cases, privacy concerns can relate more to sharing information within the company than to third parties. For example, workers may be more wary of their managers accessing people tracking data — perhaps those used to improve safety of the system — than sharing it with a robot supplier. Safety is paramount whenever these robots coexist with human workers, and shared operation can be a way to increase efficiency, making the best use of both human and automated work force. Since safety in part depends on human awareness, it should be noted that compromises may be required between privacy and safety. Predictability is an important factor, to the extent that end users typically prefer robots that follow predefined paths and merely stop for obstacles, even if they are able to plan and move freely. Understandability in terms of communication of intent with visual or verbal cues may help to improve predictability; and at the same time lead to better safety and more trust.

V. Advancing Responsible Robotics

We list the gaps identified as part of our discussions, categorizing these in terms of their urgency. With urgent issues, we attempt to identify who might fill this gap, and when we might envisage progress being made. For medium-term issues, we only describe who might fill the gap; and for much longer-term issues we do neither.

A. Gaps that need filling more urgently

GAP	Who should fill it or solve it?	Expected?
Identifying progress indicators for Responsible Robotics	Researchers (social science), Standards committees with input from industry	Short-term
Requirements Engineering for Responsible Robotics	Researchers (inter-disciplinary teams, Software Engineering)	Short-term
Teaching Responsible Robotics	Educators, Researchers	Short-term
Operationalizing explainability, predictability, and understandability in Robotics	Engineers, Philosophers	Short-term
Reporting irresponsible incidents/practices in Robotics	Legislation, Regulation, Professionalization	Medium-term
Responsible technological intervention in systemic problems, given ongoing resource constraints.	Policy experts and regulators (Engineers, Social Scientists, Philosophers, Citizens)	Medium-term
Planning with interacting values in uncertain and dynamic situations	Researchers (Computer Science, AI)	Medium-term
Testbeds for assessing interaction-based/ethical harms	Researchers (HRI, Psychology, Ethics), Users	Medium-term
Enabling the *second-hand* robot market	Insurance, Regulators, Business, Roboticists	Medium-term
Educating users to live with robots	Researchers (HRI, Education, Psychology)	Medium-term
Creation and curation of appropriate, and agreed, ML training datasets	Engineers (policy experts, philosophers, citizens)	Medium-term
Understanding, operationalizing, and arbitrating tensions and trade-offs between different goals and values for robotic systems.	Policy experts and regulators (Engineers, Social Scientists, Philosophers, Citizens)	Long-term

B. Gaps that need filling in the less immediate term

GAP	Who should fill it or solve it?
What would an international "Coordination on Responsible Robotics" agency do?	Legislators, International NGOs, Researchers (Law, ...),
Easier/Simpler/Cheaper insurance/liability for "responsibly produced" robots	Insurance, Legal, Regulators, Business
Clear specification of requirements for robotic systems and tool chain for design and verification	Engineers
Evaluation of robotic systems: safety, efficacy, effectiveness for target population, broader effects on society.	Policy Experts, Regulators, Researchers (Engineers, Social Scientists, Philosophers), Citizens
Norms for robots acting in the "real world"	Researchers (Social Sciences, Design)
Challenges for broader societal trust in robotics	Policy Experts and Regulators, Researchers (Social Scientists, Engineers, Philosophers, Citizens)
Varieties of robotic personality for different applications	Engineers, Social Scientists, Citizens
Clarifying relation between robotic agency and (causal, legal, moral) responsibility	Philosophers, Engineers, Policymakers, Regulators

C. Questions Deserving More Attention

GAP
Clarifying agency, autonomy, and responsibility spectra for robotic agents
Modeling and prediction of human behavior, in a fair way, noting that the human behavior can evolve in response to robots
Is the increasingly human-to-human-like nature of interaction between humans and robots a benefit or a loss, not just in individual cases, but for society as a whole?
Understanding and operationalizing predictability when an embodied robot is reacting with the world, while the world (including humans) is itself unpredictable and complicated
New methods, languages, and principles to talk about autonomy (autonomous robots fall into the "grey area" between human agents and mere tools) rather than borrowing terminology and analogies from human psychology and philosophy aimed at the human case
Domestic robots are likely on their way and so we must tackle questions around these as the potential societal consequences are huge. Do we wish to live in a world where our children are treating robots as their best friends, etc.? Or is this already happening on social media?
Understanding the methods and processes involved in robotics: building robots is not a science and the product is not research papers. It is a design process involving researchers, designers, end-users etc. Addressing the disconnect between the lab-focused research of academics and the fact that many robots are already being used in the field.
Clarifying autonomous (also moral) agency for robotic agents.

VI. Concluding Remarks

Robot systems raise some similar ethical issues as other products like AI systems, but their physicality must also be taken into account. This roadmap outlines key values and questions that need to be addressed to advance responsible robotics. We acknowledge that, in the abstract, the values and concepts of responsible robotics are broad and sometimes slippery, and people are bound to disagree with different aspects discussed above. We stress that our list is not meant to be either unique or exhaustive. Our choice of values and how we construe them was motivated by the current state of research and development in robotics, reflecting the expertise and experience of the multidisciplinary group participating in the seminar. As we noted, these choices are also informed by the distinctive challenges robots present as artificial embodied agents (as opposed to human and animal agents or disembodied/algorithmic agents like most AI systems). With this article, we hope to continue and grow the conversation about responsible robotics in the wider research community as well as in society at large.

Acknowledgments

This work was made possible through Dagstuhl Seminar on *Roadmap for Responsible Robotics* (23371). In addition, we acknowledge the following sources of support for individual authors: Fisher is supported by the UK Royal Academy of Engineering's *Chair in Emerging Technologies* scheme and the work is partially funded by EPSRC in the UK, through the Computational Agent Responsibility project (EP/W01081X/1). Rodríguez-Lera by Grant PID2021-126592OB-C21 funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; Ljungblad by Wallenberg AI, Autonomous Systems and Software Program, Humanity and Society; Moon by the Natural Sciences and Engineering Research Council of Canada. Baum by the German Research Foundation (DFG) under grant No. 389792660, as part of TRR 248, see https://perspicuous-computing.science, by the German Federal Ministry of Education and Research (BMBF) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007), and by the European Regional Development Fund (ERDF) and the Saarland within the scope of (To)CERTAIN. The authors would like to thank Laura Stenzel for additional comments.

References

[1] I. S. Organization, "8373:2021 standard: "robots — vocabulary"," 2012, https://www.iso.org/obp/ui/#iso:std:iso:8373:ed-3:v1:en.
[2] S. Coghlan and C. Parker, *Harm to nonhuman animals from AI: A systematic account and framework*. Philosophy & Technology: Springer, 2023, vol. 36, no. 2.
[3] R. Sparrow and M. Howard, "Robots in agriculture: Prospects, impacts, ethics, and policy," in *Precision Agriculture*. Springer, 2021, doi:10.1007/s11119-020-09757-9.
[4] G. Veruggio, "The euron roboethics roadmap," in *Proc. 6th IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 612–617, doi:10.1109/ICHR.2006.32133.
[5] B. Siciliano and O. Khatib, Eds., *Springer Handbook of Robotics*. Springer, 2016.
[6] Ethically Aligned Design (version, *2)*. IEEE, 2019.
[7] BS8611, *"Guide to the Ethical Design and Application of Robots and Robotic Systems"*. British Standards Institution, 2016.
[8] F. S. de Sio and G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philosophy and Technology*, vol. 34, pp. 1057–1084, 2021.
[9] S. Vallor, "Edinburgh declaration on responsibility for responsible AI," 2023, https://medium.com/@svallor_10030/edinburgh-declaration-on-responsibility-for-responsible-ai-1a98ed2e328b.
[10] C. Bartneck, C. Lutge, A. Wagner, and S. Welsh, *An Introduction to Ethics in Robotics and AI*. Springer, 2020.
[11] T. Ziemke, "Understanding social robots: Attribution of intentional agency to artificial and biological bodies," *Artificial Life*, vol. 29, no. 3, pp. 351–366, 2023.
[12] M. Coeckelbergh, *Robot Ethics*. The MIT Press, 2022.
[13] D. Feil-Seifer and M. Mataric, "Socially assistive robotics," *IEEE Robotics and Automation Magazine*, vol. 18, no. 1, pp. 24–31, 2011.
[14] A. Langer, R. Feingold-Polak, O. Mueller, P. Kellmeyer, and S. Levy-Tzedek, "Trust in socially assistive robots: Considerations for use in rehabilitation," *Neuroscience and Biobehavioral Reviews*, vol. 104, pp. 231–239, 2019.
[15] B. J. Grosz, D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, and J. Waldo, "Embedded ethics: integrating ethics across cs education," *Commun. ACM*, vol. 62, no. 8, p. 54–61, July 2019. [Online]. Available: https://doi.org/10.1145/3330794
[16] National Institute of Standards and Technology, "Risk management framework for information systems and organizations: a system life cycle approach for security and privacy," December 2018, doi:10.6028/NIST.SP.800-37r2.
[17] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer, 2019, vol. 2156.
[18] D. Cawthorne and A. Robbins-Van Wynsberghe, "From healthdrone to frugaldrone: Value-sensitive design of a blood sample transportation drone," in *2019 IEEE international symposium on technology and society (ISTAS)*. IEEE, 2019, pp. 1–7.
[19] D. Cameron, E. Collins, S. de Saille, and I. Eimontaite, "The social triad model: considering the deployer in a novel approach to trust in human–robot interaction," *International Journal of Social Robotics*, 2023.
[20] H. A. S. J. Abbass and D. J. Reid, *Foundations of Trusted Autonomy*. Incorporated, 1rst edition: Springer Publishing Company, 2018.
[21] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, 2020.
[22] H. R. Pelikan, B. Mutlu, and S. Reeves, "Making sense of public space for robot design," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2025, pp. 152–162.
[23] M. Gamboa, "My body, my baby, and everything else: An autoethnographic illustrated portfolio of intra-actions in pregnancy and childbirth," in *Procs. of 7th.TEI*, 2023, pp. 1–14, https://doi.org/10.1145/3569009.3572797.
[24] A. Cavoukian, "Privacy by design. the 7 foundational principles. implementation and mapping of fair information practices," *Ph. D. Information & Privacy Commissioner, Ontario, Canada*, 2011.
[25] S. Eick and A. I. Anton, "Enhancing privacy in robotics via judicious sensor selection," in *Proc. ICRA*, 2020, pp. 7156–7165, doi:10.1109/ICRA40945.2020.9196983.
[26] A. K. Taras, N. Sunderhauf, P. Corke, and D. G. Dansereau, "Inherently privacy-preserving vision for trustworthy autonomous systems: Needs and solutions," *Journal of Responsible Technology*, vol. 17, 2024.
[27] T. Everitt, G. Lea, and M. Hutter, "Agi safety literature review," in *Proc. IJCAI'18*, 2018, pp. 5441–5449, https://www.ijcai.org/proceedings/2018/5441.

[28] E. Commission, "Third draft of the general-purpose ai code of practice published, written by independent experts," March 2025, https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts.

[29] J. McDermid, Y. Jia, and I. Habli, "Upstream and downstream ai safety: Both on the same river?" 2024, https://arxiv.org/abs/2501.05455.

[30] F. Baowei, S. N. Wan, C. Sunita, and K. K. Chee, "The safety issues of medical robotics," *Reliability Engineering and System Safety*, vol. 73, no. 2, pp. 183–192, 2001.

[31] A. Bicchi, M. A. Peshkin, and J. E. Colgate, "Safety for physical human–robot interaction," in *Springer Handbook of Robotics. Springer, Berlin, Heidelberg*, B. Siciliano and O. Khatib, Eds. Springer, 2008.

[32] S. Braganca, E. Costa, I. Castellucci, and P. M. Arezes, "A brief overview of the use of collaborative robots in industry 4.0: Human role and safety," in *Studies in Systems, Decision and Control, vol 202. Springer, Cham*, P. Arezes *et al.*, Eds. Springer, 2019.

[33] J. Guiochet, M. Machin, and H. Waeselynck, "Safety-critical advanced robots: A survey," *Robotics and Autonomous Systems*, vol. 94, pp. 43–52, 2017.

[34] F. Martín, E. Soriano-Salvador, J. M. Guerrero, G. G. Múzquiz, J. C. Manzanares, and F. J. Rodríguez, "Towards a robotic intrusion prevention system: Combining security and safety in cognitive social robots," *Robotics and Autonomous Systems*, vol. 190, p. 104959, 2025.

[35] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. ManÃ©, "Concrete problems in ai safety," 2016, arXiv preprint.

[36] B. Gyevnar and A. Kasirzadeh, "Ai safety for everyone," *Nature Machine Intelligence*, pp. 1–12, 2025.

[37] A. F. T. Winfield, S. Booth, L. A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R. Muttram, J. I. Olszewska, F. Rajabiyazdi, A. Theodorou, M. Underwood, R. H. Wortham, and E. Watson, "IEEE P7001: A new standard on transparency," *Frontiers in Robotics and AI, section Ethics in Robotics and Artificial Intelligence*, 2021.

[38] V. J. Koeman, L. A. Dennis, M. Webster, M. Fisher, and K. V. Hindriks, "The "why did you do that?" button: Answering why-questions for end users of robotic systems," in *Proc. EMAS*, 2019, pp. 152–172, doi:10.1007/978-3-030-51417-4_8.

[39] N. Felber, F. Pageau, A. McLean, and T. Wangmo, "The concept of social dignity as a yardstick to delimit ethical use of robotic assistance in the care of older persons," *Medicine, Health Care and Philosophy*, vol. 25, no. 1, pp. 99–110, 2021.

[40] L. Zardiashvili and E. Fosch-Villaronga, "'Oh, dignity too?' said the robot: human dignity as the basis for the governance of robotics," *Minds and Machines*, vol. 30, no. 1, pp. 121–143, 2020.

[41] J. P. Arto Laitinen, Marketta Niemelä, "Demands of dignity in robotic care," *Techné: Research in Philosophy and Technology*, vol. 23, no. 3, pp. 366–401, 2019.

[42] J. Hardy, "Ethical algorithms in human-robot-interaction: A proposal," 2023, doi:10.5121/csit.2023.130214.

[43] S. Coghlan, J. Waycott, A. Lazar, and B. Barbosa Neves, "Dignity, autonomy, and style of company: Dimensions older adults consider for robot companions," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, 2021.

[44] M. Cappuccio, A. Peeters, and W. McDonald, "Sympathy for dolores: Moral consideration for robots based on virtue and recognition," *Philosophy & Technology*, vol. 33, no. 1, pp. 9–31, 2019.

[45] K. Richardson, "The asymmetrical'relationship' parallels between prostitution and the development of sex robots," *Acm Sigcas Computers and Society*, vol. 45, no. 3, pp. 290–293, 2016.

[46] A. Sharkey, "Autonomous weapons systems, killer robots and human dignity," *Ethics and Information Technology*, vol. 21, no. 2, pp. 75–87, 2018.

[47] A. Johnson and S. Axinn, "The morality of autonomous robots," *Journal of Military Ethics*, vol. 12, no. 2, pp. 129–141, 2013.