Responsible AI:
Requirements and Challenges

Malik Ghallab
LAAS-CNRS, University of Toulouse,
malik.ghallab@laas.fr

May 2019*

Abstract

This position paper discusses the requirements and challenges for responsible AI with respect to two interdependent objectives: (i) how to foster research and development efforts toward socially beneficial applications, and (ii) how to take into account and mitigate the human and social risks of AI systems.

1 Introduction

AI significantly contributes to and benefits from the accelerated momentum of technology development, which is opening a wealth of opportunities and has already brought numerous social and human benefits, as assessed for example by the evolution of the Human Development Index throughout the world. AI technologies help medical professionals improve prevention, diagnosis and care procedures. They are of benefit in environment preservation and monitoring programs, in agricultural projects, and in the modeling and management of cities, infrastructures and industries. They contribute to safer and more efficient mobility and transportation systems. They offer effective tools for multi-modal and multi-lingual interaction and information querying. However, these rapid technology developments are also the matter of legitimate concerns about risks, disruptive effects and social strains that need to be properly understood and addressed.

The concerns about AI are expressed in various forums and programs seeking to leverage AI developments toward social good, to mitigate the risks and investigate ethical issues. This is notably illustrated through the initiatives taken by international organizations, such as the United Nations and its specialized agencies,[1] the European Union,[2] or the Organisation for Economic Cooperation and Development.[3] The G7 political leadership has recently announced the future setup of an *International Panel on AI*, akin to the IPCC for the climate change. Other initiatives have been taken by technical societies,[4] NGOs, foundations, corporations, and academic organizations.[5]

The requirements and challenges regarding responsible AI developments can be analyzed with respect to two interdependent purposes: *(i)* how to foster research and development efforts toward socially beneficial applications, and *(ii)* how to take into account and mitigate the risks of AI systems. These objectives correspond to technical as well as legal and social challenges, which are briefly summarized in this position paper.

AI Perspectives (to appear)

[1] E.g., ITU studies or UNESCO initiatives.

[2] E.g., EU High level Expert Group on AI.

[3] E.g., AIGO and the forthcoming OECD AI Policy Observatory.

[4] E.g, IEEE Ethically Aligned Design.

[5] E.g., HUMANe AI, International Observatory on the Societal Impacts of AI, AI4People, Human-Centered AI, AI Now, Center for the governance of AI, AI for Good Foundation.

2 AI for the social good

AI technologies, as most digital technologies, have become ubiquitous. Learning, reasoning, heuristic search and problem solving algorithms are found in a very wide range of applications, directly integrated into artifacts or indirectly via cloud connections. Most industrial and economic sectors are deploying these techniques in their engineering methods and products. Even culture and arts are experimenting with AI in their creative tools.

The needs for socially beneficial AI applications are tremendous and raise numerous challenges. Several initiatives are attempting to address some of these needs. For example, the AI for Global Good Summit of the ITU is concerned with encouraging R&D in AI to actively contribute to the 17 Sustainable Development Goals (SDGs) of the UN. The last edition of the Summit considered a few development areas such as:

- the interpretation and processing of satellite images in food and agronomic applications (SDG 2), and in environment preservation programs (SDG 6, 13 et 17);
- the collection, treatment and open dissemination of medical data and knowledge related to epidemics and various health conditions (SDG 3); and
- the simulation of urban environments for the management and decision-making support in smart cities (SDG 11).

AI techniques can contribute to other UN sustainable development objectives, such as in education (SDG 4), water resource management (SDG 6) and industrial production (SDG 9 and 12).[6]

The challenges for fostering AI toward social good fit in two main categories: *incentives* and *integrative research*.

Incentives. The usual market incentives tend to focus on high and rapid return on investment. They may not provide research funding and investments meeting the significant needs of socially beneficial developments, specially in their initial and risky phases. A few non-profit foundations are to be commended for funding exemplary projects.[7] However, more support is needed from international cooperation and public funding, which should bring significant and concentrated resources on key objectives. Although all OECD countries (and many developing countries) have an AI development plan, their funding remains modest, as compared to the R&D investments of the few main industrial players of the field. Public incentives need to be scaled-up on socially beneficial programs.

Integrative research. The usual organization and granularity of academic research in many fields, including in AI, tend to favor focused analytical methods and disciplinary targets. They promote investigations within the useful but often narrow assumptions of each community in order to bring further in-depth and well formalized knowledge. This is certainly needed and essential for the progress of science. But it is not sufficient for driving and amplifying AI contributions toward social good. The developments required for contributing to the SDGs mentioned above and similar projects, are not just "a matter of application". They raise rich integrative research problems, within AI, as well as with other fields.

Integrative research within AI is demanded for addressing heterogenous tasks, which are inherent to socially beneficial applications. Such tasks require multiple cognitive functions, e.g., sensing, data association, as well as extraction and reasoning on the underlying ontology of a domain, in order to better actively perceive, organize, explain and rationalize a perceived field. The challenges require integrating data-based modeling and model-based reasoning. They demand combining bottom-up learning and correlation with

[6] The IAP 2019 Conference and General Assembly of Inter-Academy Partnership is devoted to these issues.

[7] E.g., Thorn Spotlight project to fight human trafficking; Allen Philanthropies with the Planet project for the conservation of coral reefs; the Rainforest Connection NGO for forests and environment conservation.

top-down causal rationalization. They also require fusing a diversity of input sources, and integrating consistently multiple knowledge representations and processing approaches that are mathematically heterogeneous.

Integrative research problems between AI and other fields are clearly at the core of most socially beneficial developments of AI. They correspond to targeted interdisciplinary projects, as well as to long term transdisciplinary programs. They also require the involvement of non-academic contributors, social actors and stakeholders within investigations and developments. These integrations are usually more complicated because of the diversity of cultural and methodological backgrounds. But they are needed in order to ground the work into real issues and to develop relevant contributions, which have to be assessed mainly from their effective field success than from their formal computational properties.

Integrative research is intrinsically difficult. It requires a long time span, due in particular to the overhead of collaborations and field tests. Given the usual criteria and bibliometrics indicators used for the funding and assessment of academic work and careers, integrative research appears risky. Furthermore, the view that science is "neutral" with respect to its possible uses is still appealing. Many researchers perceive their role as mainly to contribute to knowledge, and to leave it up to society to make use of it. But the intricacy and high pace of technosciences, particularly in AI, no longer support such a view. Today, a significant part of the AI community is concerned with promoting a research agenda that anticipates and takes into account the social utility of its investigation (see, for example, the widely supported Open Letter, and the subsequent agenda of [22]). However, a shift in the academic cultural and organizational paradigm may be needed to amplify integrative research in AI. In this regard, studies in epistemology (e.g., [17]), and examples from other domains such as the earth and climate science community [16] can be very informative.

3 Mitigating AI risks

AI scientists belong to a highly enthusiastic and positive community, supportive of social and humanistic values. Most AI publications highlight good motivations and excellent possible effects of their contributions. But not many do investigate their inherent risks. Every AI development involves particular risks that demand to be studied and addressed specifically. There are a few general categories of risks that are common to many applications. These are notably: *(i)* the *safety* of critical AI applications, *(ii)* the *security and privacy* for individual users, and *(iii)* the *social risks*. The issues in these three categories are not independent; many of them may not be exclusive to AI. They entail distinct scientific, technical, political and legal challenges, with different time horizons.

3.1 Safety critical AI applications

AI techniques are frequently integrated within artifacts and systems endowed with sensory-motor capabilities and increasing levels of autonomy. These are robots, drones, cyber-physical components, automated plants, networks and infrastructures. These techniques are more and more being deployed in safety critical applications and areas that can have very high economic or environmental costs, such as for example in:

- health: stimulators, prostheses, monitors, surgical devices, drug processes;
- transportation: autonomous vehicles, traffic control;
- network management: energy, logistics, hydraulics, various infrastructures; and
- surveillance and defense systems.

Relatively few industrial sectors have to comply with very strict certification procedures, as in aeronautics or intrusive medical devices. Procedures requiring informal technical descriptions and declarations of conformity to standards may not be sufficient given:

- the complexity and opacity of many AI models and techniques; and

- the intricate traceability of the hardware and software components within systems that are becoming larger and more complex.

The risks in human lives and social and environmental costs are not sufficiently studied and assessed. Comparisons to human-controlled systems (without AI) often raise hopes that are still difficult to quantify, e.g., reduction in road accidents or in medical errors. These comparisons are not always convincing given the public expectation and acceptance: a victim of an autonomous system is naturally much less accepted than one due to a human error.

The technical challenges here are about the extension of *Verification and Validation* (V&V) methods to AI and their industrial deployment. It is essential to be able to accurately analyze and qualify the safety properties of components and systems using AI. Formal methods (deterministic or stochastic), and/or simulation and testing methods, should in particular allow:

- to state formally the assumptions about the environment of a system, which are required for its correct functioning;
- to specify its expected functionalities and limitations; and
- to determine its essential characteristics: correction, reliability, probability of errors, false positives, sensitivity to uncertainty of data and parameters.

V&V is a very active field in Computer Science. It is well advanced for closed, well-modeled functioning environments. AI brings to the V&V field a rich set of challenges to handle software, robots, and cyber-physical systems that interact with open, partially known and imperfectly modeled environments. Among these challenges, the following issues are outstanding:

- how to formally quantify the uncertainty of a system while taking into account the nature of the data and models used, e.g., in medical diagnosis [3]?
- how can a system monitor online its environment and own state with respect to the assumptions that are needed for its correct functioning, and adapt when these assumptions are not met?
- how to assess the V&V properties of a complex system integrating AI techniques from the V&V properties of its components (compositional properties) ? how about blackbox-type components?
- what are the possible V&V approaches for a system that learns and evolves continually in interaction with its environment?

These issues, and others, are major research challenges, of concern to a large community (see for example [1], [24], and [13]). However, many deployments will certainly take place before all these challenges are solved. Furthermore, theoretical restrictions in computational complexity and decidability have been known for decades, or recently uncovered (e.g., learning undecidability [4]). Nonetheless, it remains essential to raise the awareness of designers and users of critical applications about open issues and limitations of current techniques, about mitigating methods and the required vigilance in rapid deployments.

3.2 Security and privacy for individual users

AI techniques have become the mediator between the users and the digital world. Access to online data produced by the billions of people and connected systems, and, beyond data, to knowledge relevant to each user, is increasingly based on semantic content. A vocal assistant must correctly perceive oral requests in natural language. An associated querying engine must interpret each request it its context and in relation to the user's profile, which is constantly learned, refined and evolving. Images, videos and data from various physical, chemical, or physiological sensors, are to be interpreted and indexed with respect to their semantic content. Increasingly, a person's interactions with her environment, with machines and systems (at home, in stores and public equipments), or even her interactions with other persons, are performed digitally and mediated via AI. Each person generates a growing and potentially indelible "digital trace" of her behavior. Even

without direct use of digital interfaces, it is difficult to avoid leaving such a trace (e.g., walking in areas with video surveillance and facial recognition, or making purchases).

The mediation role of AI with the digital world has become so important that, for many, AI is undistinguishable from digital technologies. Studies about opinions and attitudes regarding AI can be highly instructive (e.g., [29]).[8] They can provide insight about where research and education efforts should concentrate. The general public has often ambivalent perceptions of the field, sometime mixing:

- *uncritical expectations*: algorithms and computations are accurate and correct, decisions recommended by a machine are "rational";
- *legitimate concerns* about the security and confidentiality of a user's interactions, the exploitation of personal and aggregated data, and opinion manipulation capabilities; and
- *unfounded fears* about the "singularity", or the currently improbable perspective of machines with intentions, emotions, consciousness, that may take control of human.

AI mediated interactions raise social risks (covered in subsection 3.3), as well as individual risks. The latter correspond to real and subjective vulnerabilities, frustrations and the possible rejection of digital technologies by a part of the population, which can feel marginalized.

The needs at this level are technical, but also educational, institutional and legal. The technical problems concern in particular the following points:

- *Security* of digital interactions: the state of the art is well advanced but the deployment of known techniques is clearly insufficient, specially in portable applications and connected objects. Security vulnerabilities frequently make the news headlines, e.g., in vacuum cleaner robots or vocal assistants [6]. There are also hard open problems that need to be addressed, e.g., the susceptibility of neural network techniques to attacks and adversarial examples [9].
- *Confidentiality, privacy* and use of personal data: here also there is an insufficient deployment of the state of the art.
- *Intelligibility and transparency*: these issues raise challenging scientific and technical problems. A decision support system should be able to explain its assumptions, limitations, and criteria. The important issue of the decision criterion is often overlooked: a rational decision is almost always relative to some criterion, which does not necessarily meet a user's inclination and priorities. A decision support system must be able to explain and justify the response to a request. All this must be done in terms that are understandable to the user.

The insufficient deployment of known security and confidentiality techniques is generally due to weak economic incentives and regulatory constraints. The recent EUGDPR measures reinforce confidentiality and respect for privacy. However, these and other similar measures are criticized as addressing the problems in partial and insufficient manners. The contractual relationship between a user and a platform is unbalanced. The imbalance highlights the user's vulnerability to platforms deployed by a small number of corporations that have huge economic and legal support potentials. It is natural for these corporations to pursue their own interests, including by harvesting profitable behavior data, as long as this is legal. They offer services regarded as essential to everyone for a modern social life, but at a largely hidden cost. Furthermore, a user may decide (in theory) about the use of her personal data, but she has not much to say about the aggregated data and the resulting models to which she contributes. These models represent an important source of revenues, as well as risks. In some cases, a user may not agree to the elaboration of a behavior model, or she may view it as a public resource to be used solely for open research. Additional legal and technical studies are needed, e.g., for the development of accountable *data trusts*, which can play an intermediary role between users and platforms to better balance

[8] The MIT Tech Review presentation of this study is untitled: "*Americans want to regulate AI but don't trust anyone to do it*".

contractual relationships [8].

Guidelines (e.g., the UN Guiding Principles or the EU AI Ethics Guidelines) and ethical commitments of companies are certainly useful and needed, but not sufficient. The urgent requirements here are more in regulations and public policies than in ethics [27]. Legal studies and possibly social experiments are needed to raise awareness, support deliberations, and foster international cooperations regarding AI and digital regulations.

3.3 Social risks

The acceptability of a technology is often interpreted in terms of customers, i.e., the existence of a sufficiently broad public that adopts and uses the technology. But social acceptability is much more demanding than individual acceptance. Among other things, social acceptability needs:

- to take into account the long term, including possible impacts on future generations;
- to worry about social cohesion, in particular in terms of employment, resource sharing, inclusion and social recognition;
- to integrate the imperatives of human rights, as well as the historical, social, cultural and ethical values of a community; and
- to consider global constraints affecting the environment or international relations.

Biases. Decision support tools can be biased. In some cases, systems are intentionally designed as unbalanced, e.g., for a recommender system integrating propaganda or commercial goals. Users should be explicitly warned about the underlying objectives of systems that may distort their outcomes. More problematic are the hidden and non intensional biases of systems required to be neutral and fair. Numerous cases of gender, ethnical or seniority biases have been reported in decision support systems for health, banking, insurance, recrutement, career assessment, or even in public services such as legal assessment and city surveillance applications [14, 20, 25]. This is generally the case because these systems lacks transparency, intelligibility and rely on training data which is biased in hidden ways difficult to uncover and mitigate. There is a need for further research in techniques for auditing the fairness of a system, and in regulations requiring their use for certification mechanisms.

Behavior manipulation. It has been known for ages that individuals can be manipulated. AI technologies augment their vulnerability, in particular with the worldwide deployment of ergonomic and playful devices that implement powerful communication, sensing, processing and decision making functions. Manipulation capabilities are illustrated by the increasingly more effective techniques for social monitoring, text and audio-visual "optimization", debate steering, behavior modeling and shaping, and market driving [30]. The incentives for using available techniques toward profitable purposes are very high. Dubious practices with high social, political and economic risks will remain in use as long as they are unregulated. In addition to regulations, and for supporting them, further research in AI may contribute to methods for detecting manipulation attempts.

Democracy. The political risks, illustrated by the Cambridge Analytica scandal, are analyzed by several authors as a threat to democracy [19], [31]. Studies show that AI presents opportunities as well as risks on the full range of human rights, with already observed impacts [21].

Economy. Economic risks correspond to several AI deployments, for example in High Frequency Trading (HFT), or in algorithmic pricing. The possible destabilization effects of HFTs are far from being well understood [26]. Algorithmic pricing using learning, profit optimization and indirect interactions between computational agents can lead, even without any explicit agreement, to artificially higher prices, as with the illegal price cartel mechanisms [11]. The main assumption of the liberal economy postulates a supposedly neutral free market, considered as a vir-

tuous "unknowable and uncontrollable" information processor, which should remain unregulated. The real time observation, learning, modeling and feedback control capabilities permitted with AI tools are in clear contradiction with this assumption. Regulations to mitigate the corresponding risks are urgently needed.

Employment. AI contributes to the increasing automation of services, industry and agriculture, which brings progress, as well as important social risks for employment. There is no general consensus on this risk (nor is there one on global warming). However, available studies, which remain insufficient, converge toward a substantial reduction of jobs in the short to medium term. According to an OECD study for its 21 countries [2], 9% of jobs have a high risk of automation; a higher percentage of 20 to 25% of jobs have a medium risk (other studies conclude to more alarming risk levels, e.g., [15]). Furthermore, technology developments are strongly suspected to be a contributing factor for the observed increase in social inequalities [5], which reduce social involvement.

It is clear to most observers that the existing social measures for handling temporary fluctuations (e.g., unemployment benefits) are inadequate for a long-standing, continuing change. Several laudable studies and initiatives are undertaken to mitigate the unemployment risks, in terms of training and job creation (e.g., Innovation for Jobs), resource sharing, social recognition and integration. The challenge here is to further develop these initiatives in order to respond in time to the undesirable consequences of numerous technology deployments.

Military systems. AI in weapons and military systems correspond to another area of worry, which raises ethical concerns, as well as risks of international instability and increased conflicts. AI technologies greatly enhance the military capabilities of perception, surveillance, intelligence, fighting, and intervention. AI is naturally a dual-use technology, easily transposed from the commercial to the military domain. This makes impracticable control procedures such as those used for nuclear weapons containment [7]. This also makes weapons and devices with integrated AI relatively more "affordable" than other heavy military technologies.[9] These weapons may be more easily accessible to rogue groups. In addition, international arm trade agreements, including the recent Arms Trade Treaty, do not cover digital weapons, such as drones, robots, ROV and AUV. The widely supported Open Letter for a ban on autonomous weapons is an excellent initiative which needs to be pursued into studies and regulations.

Can we mitigate technically the above social risks by extending the problem solving and reasoning competences of our tools with moral appraisal capabilities? We certainly need machines which are, by design, provably safer, more secure, intelligible, unbiased, respectful of privacy, and meeting in their functioning the constraints and rules demanded by society. These and similar properties can be reasonably well understood, formalized, and machine implementable [10, 23]. Technical standards for meeting them in AI systems should be developed and deployed, as for other artifacts. However it is unclear what might be the specification of an automated weapon, or an automated trader, capable of resolving ethical choices on the basis of moral principles. Several approaches to the notion of an "artificial moral agent" in a general sense (i.e., levels 3 and 4 of [18]), are criticized as being philosophically illegitimate (e.g., [12, 28]). They can be quite misleading. We should strive to clarify and disseminate widely the knowledge about the capabilities and limitations of our tools, and to integrate the social involvement and assessment of their potential uses as an essential component in our research and design methodology.

The needs for responsible AI developments with respect to the social risks correspond in particular to political and legal measures and to in-

[9] E.g, the SGR-A1 autonomous Sentry Gun, capable of covering a radius of several kilometers, is said to cost about 200K$.

ternational agreements. However, the required measures are part of the regulatory mechanisms of society. These mechanisms have a quite long response time: decades are needed to better understand, educate, spread the awareness and build up the social forces required to impose regulations. But the momentum of technology has become much faster. The discrepancy between the two dynamics demands for *proactive approaches*. However, no predictive models of the possible social and economic effects of a given technical deployment are readily available. A proactive approach must rely on *social experiments*, and integrative research about social risks and mitigation measures. Here too, a change of paradigm is required to fund and develop joint investigations between AI and social scientists, to give a better understanding of AI to the former, and of social and economic mechanisms to the latter. More involvement of AI within relatively recent areas such as "Science, Technology and Society" (e.g., at Stanford or MIT) should provide opportunities to complement the usual empirical observation methodology of social sciences with significant experimentation, modeling and even simulation. It should be noted that simulation, based on elementary models, is emerging in a few areas of social sciences. AI can actively contribute to its development and effectiveness. Finally, let us remark that social experimentation before a technical deployment reduces the discrepancy between the technology momentum and the social regulation dynamics.

4 Conclusion

AI, like any other technology, can have virtuous effects, as well as much less desirable consequences. AI as a research field cannot be blamed for the latter. The specific historical, social and economic context of a deployment can make an AI machine "a Dr Jekyll or a Mr Hide". The discrepancy between the slow social and legal mechanisms and the fast technology momentum renders the steering of the deployments and uses of AI more challenging.

AI scientists and professionals do not have, obviously, the full steering control. But neither are they powerless nor irresponsible. They are accountable for and capable of raising the social awareness about the current limitations and risks of their field. Up to some point, they can choose or at least influence their research agenda. They can engage into integrative research and work toward the needed paradigm shift in order to foster socially beneficial developments and address the human and social risks of AI. The initiatives and projects referred to here illustrate many of these engagements which are going on and gaining strength. The growing effectiveness of AI is simply commensurate with its social responsibility. The technical and organizational challenges are tremendous, but the AI scientific community has to face them.

List of abbreviations.

- AI: Artificial Intelligence
- EU: European Union
- G7: Group of Seven
- IAP: Inter-Academy Partnership
- IPCC: International Panel on Climate Change
- ITU: International Telecommunication Union
- NGO: Non-Governmental Organization
- OECD: Organisation for Economic Cooperation and Development
- UN: United Nations
- UNESCO: United Nations Educational, Scientific and Cultural Organization

Declarations.

- Availability of data and material: not applicable.
- Competing interests: the author has no competing interest.
- Authors' contributions: the author wrote the paper.
- Acknowledgements: many thanks to the reviewers whose comments and suggestions have helped improve this paper.

- Funding: not applicable
- Authors' information: not applicable

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. *Computing Research Repository*.

[2] Arntz, M., Gregory, T., and Zierahn, U. (2016). The Risk of Automation for Jobs in OECD Countries. *OECD Social, Employment and Migration Working Papers*, (189).

[3] Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

[4] Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A., and Yehudayoff, A. (2019). Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48.

[5] Brynjolfsson, E. and McAfee, A. (2016). *The Second Machine Age*. Norton.

[6] Carlini, N. and Wagner, D. A. (2018). Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. *Computing Research Repository*.

[7] Chien, A. A. (2019). Open Collaboration in an Age of Distrust. *Communications of the ACM*, 62(1).

[8] Delacroix, S. and Lawrence, N. D. (2018). Disturbing the 'one size fits all', feudal approach to data governance: bottom-up data Trusts. *SSRN*, pages 1–30.

[9] Edwards, C. (2019). Hidden Messages Fool AI. *Communications of the ACM*, 62(1).

[10] Etzione, A. and Etzioni, O. (2016). Designing AI Systems that Obey Our Laws and Values. *Communications of the ACM*, 9(59):29–31.

[11] Gal, M. S. (2019). Illegal Pricing Algorithms. *Communications of the ACM*, 62(1).

[12] Hunyadi, M. (2019). Artificial moral agents. really? In Laumond, J., Danblon, E., and Pieters, C., editors, *Wording Robotics*, Tracts in Advanced Robotics. Springer.

[13] Ingrand, F. (2019). Recent Trends in Formal Validation and Verification of Autonomous Robots Software. In *IEEE International Conference on Robotic Computing*.

[14] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *Propublica*.

[15] Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., and Sanghvi, S. (2017). Jobs lost, jobs gained: Workforce transition in a time of automation. Technical report, McKinsey Global Institute.

[16] Mauser, W., Klepper, G., Rice, M., Schmalzbauer, B. S., Hackmann, H., Leemans, R., and Moore, H. (2013). Transdisciplinary global change research: the co-creation of knowledge for sustainability. *Current Opinion in Environmental Sustainability*, 5(3-4):420–431.

[17] Mittelstrass, J. (2011). On transdisciplinarity. *Trames. Journal of the Humanities and Social Sciences*, 15(4).

[18] Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now*, 72:12–14.

[19] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133).

[20] O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Random House.

[21] Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., and Kim, L. Y. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks. *SSRN*.

[22] Russell, S. J., Dewey, D., and Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine*.

[23] Sandewall, E. (2019). Ethics, Human Rights, the Intelligent Robot, and its Subsystem for Moral Beliefs. *International Journal of Social Robotics*, pages 1–11.

[24] Seshia, S. A., Sadigh, D., and Sastry, S. (2016). Towards Verified Artificial Intelligence. *Computing Research Repository*.

[25] Skeem, J. L. and Lowenkamp, C. (2016). Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. *SSRN*.

[26] Sornette, D. and von der Becke, S. (2011). Crashes and High Frequency Trading. *SSRN*.

[27] Vardi, M. Y. (2019). Are We Having An Ethical Crisis in Computing? *Communications of the ACM*, 62(1).

[28] Yampolsky, R. Y. (2013). Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach. In Muller, V., editor, *Philosophy and Theory of Artificial Intelligence*, pages 389–396. Springer Verlag.

[29] Zhang, B. and Dafoe, A. (2019). Artificial Intelligence: American Attitudes and Trends. Technical report, Center for the Governance of AI, University of Oxford.

[30] Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89.

[31] Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.