Artificial Intelligence and Responsible Innovation

Miles Brundage

Abstract

Researchers in AI often highlight the importance of socially responsible research, but the current literature on the social impacts of AI tends to focus on particular application domains and provides little guidance to researchers working in other areas. Additionally, such social impact analysis tends to be done in a one-off fashion, proposing or problematizing a particular aspect of AI at a time, rather than being deeply integrated into innovation processes across the field. This paper argues that work on the societal dimensions of AI can be enriched by engagement with the literature on “responsible innovation,” which has up until now focused on technical domains like nanotechnology, synthetic biology, and geoengineering. Drawing on this literature, the paper describes and justifies three interrelated aspects of what a more deeply integrated, ongoing practice of responsibility in AI would look like: consideration of the social contexts and consequences of decisions in the AI design space; reflectiveness about one's emphasis on theoretical vs. applied work and choice of application domains; and engagement with the public about what they desire from AI and what they need to know about it. Mapping out these three issues, it is argued, can both describe and theorize existing work in a more systematic light and identify future opportunities for research and practice on the societal dimensions of AI. Finally, the paper describes how philosophical and theoretical aspects of AI connect to issues of responsibility and technological governance.

Introduction

Thought leaders in AI often highlight the potentially transformative social impacts of their field and the need for researchers to proactively engage with the public about these impacts (Norvig and Russell 2009; Horvitz and Selman 2009; Lin et al. 2011; Anderson and Anderson 2011). Arguments for such reflection and engagement are often motivated by the anticipation of substantial and potentially disruptive societal effects of AI and robotics (Nourbakhsh 2013) as well as by a recognition that such behavior is part of the responsibility of scientists qua scientists (Douglas 2009). Yet, as described in more detail below, the current literature on AI and its societal dimensions is often done in a somewhat ad-hoc fashion, gives little practical guidance to researchers in their day-to-day activities as scientists, and pertains only to certain sub-topics in the field.

Simultaneous to this work on AI's societal dimensions, a rapidly growing field of research has emerged over the past decade or so known variously as “responsible research and innovation” or simply “responsible innovation.” Some have argued that the complexity and rapid pace of change in modern technoscience merits a systematic approach to connecting science and technology to their broader social context. Several definitions and frameworks of responsible innovation have been put forth (e.g., Von Schomberg 2011; Stilgoe et al. 2013). Here, responsible innovation is taken to mean “taking care of the future through collective stewardship of science and innovation in the present,” which in turn can be disaggregated into four dimensions: anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al. 2013). A growing number of techniques corresponding to one or more of these dimensions have been piloted, scrutinized, and institutionalized in recent years.

The thesis of this paper is that this framework of responsible innovation provides a useful lens for embedding reflection on the societal dimensions of AI more deeply in the innovation ecosystem. In addition to providing a coherent theoretical framework for retroactively understanding the societal reflection in AI that has already begun, it also serves to highlight gaps in literature and practice that can be addressed in the future. Specifically, drawing on work done on AI and the responsible innovation literature, the paper will describe and justify three interrelated aspects of responsible innovation in AI that, jointly, satisfy the definition above and can drive future work: consideration of the social contexts and consequences of decisions in the AI design space; reflectiveness about one's emphasis on theoretical vs. applied work and choice of application domains; and engagement with the public about what they desire from AI and what they need to know about it.

After reviewing existing work on the societal dimensions of AI, and its limitations, the paper will outline a case for embedding reflection on and practice of responsibility in AI in an ongoing manner. Three proposed aspects of responsible innovation in AI will be described and justified. They will be framed in terms of questions, intended to stimulate reflection and enactment of responsibility in an ongoing rather than ad-hoc fashion. These three sections of the paper will provide examples of what they could mean in practice going forward as well how they have already been enacted, though typically without explicit reference to the terms and frameworks used here such as responsible innovation. Finally, the paper will connect notions of responsible innovation and technological governance more generally to the philosophy and theory of AI, and note open questions that remain in theorizing and practicing responsible innovation in AI.

Limitations of Previous Work

Several literatures have touched on aspects of the societal dimensions of AI in recent years. Perhaps the most closely related literature to the present paper is robot ethics, which analyzes the societal dimensions of intelligent robots (Lin et al. 2011). Additionally, fields such as information and computer ethics (Floridi 2010) have bearing on the societal dimensions of AI, but analysis there tends to focus on how to best think about and shape the social uptake and regulation of technologies that have already been developed, whereas the conception of responsible innovation invoked here also encompasses the processes of innovation themselves and the decisions by scientists, policy-makers, and society that include and even precede innovation itself, such as private and public sector funding. There is a rich literature on machine ethics, which attempts to develop computational models of morality to advise humans or guide the actions of robots (Anderson and Anderson 2011). Making robots ethical can be distinguished, to some extent, from how roboticists and AI researchers themselves can be ethical with respect to their research and its societal dimensions, although the two are related (Winfield 2013).

There have also been attempts to enumerate the ethical responsibilities of roboticists (Murphy and Woods 2009; Parry et al. 2011), but like the robot ethics literature more broadly, these efforts have tended to focus on discrete examples of what roboticists should not do—such as deceive people about the intelligence of their creations—rather than what they can do to bring about positive social good in an ongoing fashion over time and across many possible decision spaces as technologies evolve and new risks and opportunities present themselves. They also tend to rely on rule-based ethical reasoning, rather than on integrating reflection on the societal dimensions of one's ongoing practices in a flexible and productive way. Furthermore, much of the literature on the societal dimensions of AI to date has focused on analyzing particular applications or sub-disciplines of AI (e.g., accountability issues involving military robots or privacy concerns raised by data mining), a practice that fails to yield much practical guidance for AI researchers in other areas. Finally, as implied by the names of fields such as "robot ethics," several of these literatures are specifically about robotics and do not claim to apply to responsible innovation across AI more generally, including, for example, development of disembodied agents. Similarly, much has been written on the societal dimensions of AI, but these literatures tend to focus on discrete sub-topics of or social issues raised by AI one at a time, and to be oriented towards particular envisioned risks or opportunities stemming from AI, rather than a need for a systemic approach to build capacity throughout the emerging science and innovation ecosystem.

The Need for Systematic Responsible Innovation in AI

There are at least two major reasons the aforementioned limitations of existing work on the societal dimensions of AI ought to be addressed and a more comprehensive approach to responsible innovation in AI is needed. First, the nature of AI research will evolve over time, as will its plausible social consequences. Thus, embedding anticipation, reflexiveness, and other aspects of responsibility deeply into the practice of research itself is essential to taking care of the future in Stilgoe et al.'s (2013) sense. For example, the AAAI Presidential Panel on Long-Term AI Futures (Horvitz and Selman 2009) which issued a report on particular risks and the need for responsibility, may indeed have had significant value by legitimizing the call for further attention to responsible innovation by researchers and policy-makers, but it represents only one among many possible models for technological governance. As the responsible innovation literature and the history of technological development attests, the most important social issues raised by a technology may not be the ones anticipated by those working in the field at one particular point in time. A more spatially and temporally distributed model of technological governance would draw on a wider range of insights and enable reflective decision-making across a wider range of actors than would be possible in a model of technological governance in which a small set of particularly motivated researchers do the majority of the work.

Second, a clearly articulated (but flexible) framework for responsible innovation in AI can serve to identify gaps in existing efforts, and thereby catalyze productive future work on the societal dimensions of AI. As noted in the prior section, much existing work focuses on particular risks or opportunities, sub-fields, or applications of AI in isolation. While such work is valuable, there is no reason to assume that existing work exhausts the range of questions to be asked. Indeed, there is reason to think otherwise. To give merely one example to be illustrated in more detail in the next section: there has been work on ethical decision-making by artificial agents, but this represents only one dimension among a potentially very large number in the design space of AI. Framing the question more broadly, as done below, may serve to identify novel affordances and path dependencies that would not otherwise be noticed by researchers working in a particular sub-field of AI. Thus, by characterizing responsible innovation in AI at a level of abstraction that is sufficiently broad to cover the entirety of the field, but which lends itself to second and third-order sub-questions and lines of inquiry, the framework presented here seeks relevance beyond any one specific risk, opportunity, application, or sub-field of AI.

Such an approach would facilitate the building of capacities, distributed across space and time, that allow the AI ecosystem to respond not only to the types of societal concerns already expressed in the literature, but also to help prepare it for future, currently unanticipated decision points, where issues of distributed (social) responsibility may arise.

The next three sections of the paper will motivate and describe the three proposed aspects of responsible innovation in AI, which are:

1. How could different choices in the design space of AIs connect to social outcomes, and what near and long term goals are motivating research?
2. What domains should AI technology be applied to?
3. What does the public want from AI, and what do they need to know?

Goals Matter: Motivation and Description

“How could different choices in the design space of AIs connect to social outcomes, and what near and long term goals are motivating research?”

Reflection on the diversity of options in the space of possible AI and robot designs and research goals is a cornerstone of responsible innovation in AI, since different choices in that space will have different real world consequences. Scientific curiosity and rigor, as well as extant funding regimes, underconstrain the development of AI in any particular way, opening space for explicit normative considerations by experts and the public. To illustrate: “what is the nature of human intelligence?” and “what is the nature of the space of possible intelligences?” and “what is the nature of animal intelligence?” are all potentially very important scientific questions in their own right. Yet, by orienting one's inquiry toward one or the other, different technological artifacts may become more or less likely – mere curiosity does not dictate a single choice – and so on for many other, more specific research questions. This aspect of responsible innovation in AI maps closely onto the dimensions of anticipation and reflexiveness in the Stilgoe et al. (2013) framework, as it suggests the need for a reflective orientation toward how choices in the present influence the future. Douglas (2009) also notes the need for such reflection in any coherent conception of responsibility, writing, “Minimally, we are morally responsible for those things we intend to bring about. … While this is widely accepted, it is a difficult question under which circumstances and to what extent we should be responsible for unintended consequences.” While perfect foresight is unattainable and thus not a moral responsibility of researchers, and the appropriate long-term vision(s) for AI may vary across domains, creatively anticipating the possible impacts of long-term goals for nearer term social outcomes, and acting on such reflections, is a critical element of what it means to be a responsible innovator in AI.

As explained in Norvig and Russell (2009) and elsewhere, different conceptions of AI will yield different research priorities and benchmarks, yet little research has been done on what the social and economic implications of realizing these diverse long-term goals could be. Some such work has already been done, however, and much of this work can be constructively viewed through the lens of the question posed above. For example, Hoffman et al. (2012) argue that by complementing rather than substituting the strengths of human intelligence with computers, orienting research towards the long-term goal of human-centered computing will yield better social outcomes than work towards traditional conceptions of AI. Others (e.g., Nilsson 2005) call for a greater focus on developing integrated intelligent systems rather than making continued incremental progress in particular sub-disciplines, reasoning that there are certain tasks we want machines to do that require human-level intelligence. How AI researchers orient themselves with regards to such long-term considerations may have significant implications. Likewise, short-term goals—the consequences of which an individual research may have relatively more influence over—can greatly influence the development of subsequent technological artifacts and, by extension, their social impact.

A comprehensive list of the possible axes/spectra in the design and goal space of AI is beyond the scope of this paper, but some illustrative examples of points on these axes/spectra that have previously been proposed are: acting rationally (as opposed to thinking rationally, acting humanly, or thinking humanly; Russell and Norvig 2009); being comprehensible and predictable to users and having a high degree of robustness against manipulation as opposed to being easily hacked (Bostrom and Yudkowsky 2014); responding appropriately to authorized users and providing for smooth transfer of authority to and from other agents (Murphy and Woods 2009); choosing actions ethically as opposed to unethically or amorally (Wallach and Allen 2010); not being designed primarily to kill, being designed with safety in mind, and being transparent with respect to the non-human nature of the artifact (Parry et al. 2011). Further consideration of the contours of this multi-dimensional space, the interactions among different aspects of AI design, and their social consequences would help inform more responsible decision-making by innovators as well as by society at large and those with disproportionate influence over AI innovation such as funding agencies and corporations.

While attaining a higher level of reflexiveness and anticipation in AI could be achieved in multiple ways, one responsible innovation methodology that has shown early success in other technical domains is Socio-Technical Integration Research (STIR), which embeds social scientists and humanists in laboratories to serve as a catalyst for novel conversations and considerations of the social context of research (Fisher et al. 2010). By introducing an outside perspective to the laboratory and stimulating interdisciplinary conversations about the goals and visions of laboratory work over a period of time, STIR can be helpful in both stimulating scientific creativity and identifying potential societal concerns and desires related to emerging technologies.

Applications Matter: Motivation and Description

“What domains should AI technology be applied to?”

AI research is often fairly general across application domains, in that a computational technique could be repurposed for a technology application that was not envisioned or desired by the original AI researcher (Horvitz and Selman 2009). At the same time, domain applications can involve substantial time and resources, and the attention of academic and industry researchers is scarce. Thus, there are direct effects of researchers choosing to develop a technology in a particular domain, in that either they catalyze something that might never have existed otherwise or they make it happen sooner than it otherwise would have. Additionally, there can be indirect effects of the choice of an application domain for developing an AI approach – path dependence may be created with regards to the details of the approach, causing long-term effects and establishing public expectations such that, for example, an AI approach becomes associated with its initial application domain, creating further demand in the market.

Before considering ways in which researchers have evaluated and could evaluate the desirability of applying AI to particular domains, two points are worth noting. First, the relationship between this aspect of responsible innovation and the design considerations in the prior section is ambiguous and context-dependent. AI approaches vary in the extent to which they are domain-general, and likewise, there are various domains in which specific approaches as opposed to others make more sense. Second, the attention to application domains here should not be seen as downplaying the importance of basic research or implying that it is necessarily irresponsible. Rather, the choice of whether to devote time to basic vs. applied research (or somewhere in between these misleadingly discrete labels) may invoke different, but related, normative considerations. As previously noted, AI techniques can be applied to various domains, which could be seen as either a risk or an opportunity, depending on one's higher level attitudes about the capacity of society to govern and adopt technologies responsibly.

With regard to which applications ought to be preferentially developed or avoided, many considerations are potentially relevant. For example, Nourbakhsh (2013) notes the mismatch between funding patterns and the practical needs of communities, and he explains various ways in which community-centered robotics development can yield improved social outcomes. Likewise, Gomes (2009) argues for greater attention to sustainability-related applications of computer science research. Fasola and Matarić (2013), Reddy (2006), and others argue for the potential benefits of robotic applications in elder care, and Reddy (2006) also highlights AI applications in search and rescue operations. DARPA (2014) is currently running a competition to develop humanoid robots for disaster response applications, motivated in part by the inability of current robots to respond to the Fukushima nuclear incident and similar situations. Finally, arguments have been put forward by AI researchers and others (e.g., Docherty 2012, Arkin 2009) both for and against the development of lethal autonomous robots for military operations. These possible applications clearly do not exhaust the space of possible AI technologies in society, and each would likely impact people's lives in very different ways. Thus, responsible innovation in AI involves, in part, reflecting on one's role in the broader innovation ecosystem and what role one wants their research and artifacts based on it to play (or not play) in society.

Although compelling arguments have been made for AI technology being applied to various domains, there is likely no way to determine a socially optimal distribution of AI work across various applications, or between basic and applied research. This does not, of course, imply that the current distribution of funding and research time is anywhere near the best that it can be. Engagement with the public, as described in the next section, may help elicit useful feedback about societal priorities and concerns with respect to AI-based technologies.

Engagement Matters: Motivation and Description

"What does the public want from AI, and what do they need to know?"

The final aspect of responsible innovation in AI to be explored in this paper is engagement with the public in general and particular subsets thereof. There are both intrinsic reasons for such engagement (including the ethical premise that those who are affected by systems should have a say in those systems, and the fact that much AI research is publicly funded) as well as instrumental reasons (such as heading off negative reactions to AI in advance and getting useful user feedback that could help facilitate the adoption of technologies). Brown (2007) explores the ways in which technologies can be said to represent the public, and other areas of research such as participatory design (Chen et al. 2013) suggest that those affected by technologies can and should be involved in their development.

One aspect of responsible public engagement in AI involves communicating to the public aspects of AI science and engineering that could be relevant to their welfare. In many cases, while a particular research project's impacts are deeply uncertain, the broad contours of a field's plausible consequences are relatively clear and socially important. For example, further progress in AI and robotics seems likely to reduce the need for routine manual and cognitive work in the coming decades (McAfee and Brynjolfsson 2014). This anticipation carries with it wide-ranging ramifications for the labor force and the need for different forms of education in the future. Thus, AI researchers working on enabling machines to perform tasks in a particular domain ought to engage with those in the relevant industry or industries about what composition of human and machine labor is both feasible and desirable, and educators need access to the detailed knowledge of AI experts if they are to adapt education to the changing needs in the market. Both the choice of communities that one engages about the implications of one's research and the way in which one engages them (for example, characterizing the uncertainty of future progress) are components of what it means to be a responsible AI researcher. Nourbakhsh (2010) highlights the moral relevance of the rhetoric used by robotics researchers, and much the same can be said of AI researchers – hype matters, and the way one engages the public can have myriad consequences.

A different but related aspect of public engagement about AI involves listening to what the public wants (or doesn't want) from AI technologies, the research for which, as previously noted, is often funded publicly. Surveys of segments of the public (European Commission 2012; Takayama et al. 2008) reveal complex and diverse public attitudes towards the desirability of different uses of robots. These results complicate the often mentioned "dull, dirty, and dangerous" vision of what AI and robots should do, and point to the desirability of further public engagement about how to collaboratively envision and govern the future of AI. Moon et al. (2012) propose the Open Roboethics Initiative, an effort to build an open online community in which stakeholder conversations about robot ethics can directly inform technology designs. With regard to public expectations of (and, sometimes, fears about) AI, it should be acknowledged that science fiction already has a key role as one of the de facto means of technology assessment for the masses (Miller and Bennett 2008), and it represents one possible modality of public engagement, through, e.g., close collaborations between scientists and fiction writers, with the Hieroglyph Project being one example of such efforts (Stephenson 2011).

Limitations of the Framework and Open Questions

As noted earlier, there are already several literatures—including machine ethics, robot ethics, and computer ethics, for example—that bear on the issue of AI and its social impact. Additionally, researchers' responsibilities are but one component of a larger ecosystem of collective responsibility for technological governance. Thus, this section will detail some of the ways in which the framework outlined in this paper does not suffice to resolve the complex question of AI and its social impact.

First, the paper so far has been deliberately ambiguous about who is responsible for which aspects of responsible innovation in AI. For example, some aspects of public engagement may be done more efficiently by the leaders of professional societies than all individual researchers in AI (this is already done to some extent; see, e.g., Buchanan and Smith 2013). Also, funding regimes don't fully constrain choices in the AI design space, but they do constrain it to some extent, highlighting the critical role of funding agencies in responsible innovation. Like other scientific work, AI research occurs within the context of a complex ecosystem of funding agencies, educational curricula, sub-disciplinary communities, conference practices, tenure expectations, diverse specialties, etc., and thus it is not always clear who is "responsible" for a particular innovation outcome (Fisher et al. 2006). In many cases, a researcher may be incentivized to act in a way contrary to responsible innovation, and this paper does not attempt to analyze here the extent to which researchers may be obligated (or not) to prioritize social responsibility over practical or professional considerations. Nourbakhsh (2009) notes the prominent role of military funding in AI and robotics research and suggests that different levels of commitment can be envisioned, with exemplars that go above and beyond what is minimally required at one end of a spectrum. Rather than resolving such questions, this paper describes an ideal toward which various actors can strive in their own ways, and for which responsibility is likely distributed across many of the parties in the aforementioned ecosystem, not just individual researchers.

Additionally, responsibility on the part of individual innovators or even entire innovation systems does not exhaust society's responsibilities in technological governance more broadly. Many technologies have their biggest social impact well beyond the time at which they are technically mature, and even a thoughtfully designed product can be used in socially harmful ways. Thus, a focus on innovation processes should not be seen as absolving, for example, policy-makers of the need to regulate technologies, or to ensure they are equitably distributed.

Furthermore, choices made by researches doing technical work in AI may, and indeed should, be influenced by the work of social scientists and philosophers working on AI-related issues. An appropriate choice in the design space of AI may depend on the outcome of a philosophical debate: for example, it might be the case that it is more acceptable to design a robot to claim that it is sentient if the balance of philosophical theory suggests such an attribution is justified, and not otherwise. Philosophical and theoretical work could also bear on the plausibility of an intelligence explosion (Horvitz and Selman 2009) and the plausible sequence, if not timeline, of jobs being possible to automate (McAfee and Brynjolfsson 2014). As a final example of the societal desirability of interdisciplinary dialogue, philosophy and theory of AI could help inform the selection of short and long term research goals by, for example, confirming or discrediting particular hypotheses about the space of possible minds and how AI could develop given the achievement of particular milestones. Thus, just as there is a normative case for bidirectional engagement between AI researchers and the public, there is likewise a case for close engagement between philosophers, theorists, and technical specialists in AI.

Conclusion

This paper has argued that, while much fruitful work on the societal dimensions of AI has been carried out, it is limited in comprehensiveness and flexibility to apply to AI as a whole. Responsible innovation in AI encompasses three interrelated aspects, which in turn satisfy the demands of the responsible innovation framework in Stilgoe et al. (2013) and help theorize and categorize much existing work done by AI researchers on related topics. This three-part framework helps illustrate how different debates, which have proceeded in partial isolation (such as disparate debates about particular design decisions), are connected to one another, and it highlights opportunities for further work. Throughout, examples have been given of researchers already reflecting on these aspects of responsible innovation in AI, but these are merely illustrative; this paper merely suggests three important questions, rather than how to answer them or what further second or third order questions they may give rise to. Finally, the paper has highlighted areas of overlap and differences between literatures such as robot ethics, machine ethics, responsible innovation, philosophy, and theory of AI, as well as discussions of technological governance more broadly, that may help to identify future opportunities for interdisciplinary work at the intersection of AI and other fields.

Acknowledgments

The author would like to acknowledge helpful comments on earlier versions of this paper from David Guston, Erik Fisher, Clark Miller, Illah Nourbakhsh, Stuart Russell, David Atkinson, participants in the Human and Social Dimensions of Science and Technology colloquia at Arizona State University, and two anonymous reviewers. This work was supported by the the National Science Foundation under award #1257246 through the Virtual Institute for Responsible Innovation (VIRI). The findings and observations contained in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

Bibliography

Anderson, M. & Anderson, S. L. (2011). *Machine Ethics*. New York: Cambridge University Press.

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots.* London: Chapman & Hall/CRC.

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Ed.), *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press.

Brown, M.. (2007). Can technologies represent their publics?. *Technology in Society*, 29, 327-338.

Buchanan, B. & Smith, R. (2013). Meeting the Responsibility to Explain AI. Slides presented at the Twenty-Seventh AAAI Conference. Association for the Advancement of Artificial Intelligence.

http://aitopics.org/sites/default/files/articles-columns/Meeting%20the%20Responsibility%20to%20Explain%20AI%20-%20AAAI%20-%2020130718.pdf Accessed 15 January 2014.

Chen, T. et al. (2013). Robots for Humanity: A Case Study in Assistive Mobile Manipulation. *IEEE Robotics & Automation Magazine*, Special issue on Assistive Robotics, 20(1).

DARPA. (2014). About the Challenge. Informational website. http://www.theroboticschallenge.org/about Accessed 15 January 2014.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal.* Pittsburgh: University of Pittsburgh Press.

Parry, V. et al. (2011). Principles of robotics: Regulating robots in the real world. EPSRC document. http://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx Accessed 15 January 2014.

European Commission. (2012). Public attitudes towards robots (report). Special Eurobarometer 382. http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf Accessed 15 January 2014.

Fasola, J. & Matarić, M. (2013). A Socially Assistive Robot Exercise Coach for the Elderly. *Journal of Human-Robot Interaction*, 2(2), 3-32.

Fisher, E. et al. (2006). Midstream Modulation of Technology: Governance From Within. *Bulletin of Science, Technology, & Society*, 26 (6), 485-496.

Fisher, E. et al. (2010). Research thrives on integration of natural and social sciences. *Nature*, 463 (1018).

Floridi, L. (Ed.). (2010). *The Cambridge Handbook of Information and Computer Ethics*. Cambridge: Cambridge University Press.

Gomes, C. (2009). Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society. *The Bridge*, Winter 2009.

Hoffman et al. (Ed.). (2012). *Collected Essays on Human-Centered Computing, 2001-2011*. Washington, DC: IEEE Computer Society Press.

Horvitz, E. & Selman, B. (2009). Interim Report from the AAAI Presidential Panel on Long- Term AI Futures. Online document. Association for the Advancement of Artificial Intelligence. http://www.aaai.org/Organization/presidential-panel.php Accessed 15 January 2014.

Docherty, B. (2012). Losing Humanity: The Case Against Killer Robots. Human Rights Watch report. http://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf Accessed 15 January 2014.

Lin, P. et al. (2011). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: The MIT Press.

McAfee, A. & Brynjolfsson, E. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* New York: W. W. Norton & Company.

Miller, C. & Bennett, I. (2008). Thinking longer term about technology: is there value in science fiction-inspired approaches to constructing futures?. *Science and Public Policy*, 35(8), 597-606.

Moon, A. et al. (2012). Open Roboethics: Establishing an Online Community for Accelerated Policy and Design Change. Presented at *We Robot 2012.http://robots.law.miami.edu/wp-content/uploads/2012/01/Moon_et_al_Open-Roboethics-2012.pdf* Accessed 19 January 2014.

Murphy, R. & Woods, D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems*, 25(4), 14-20.

Nilsson, N. (2005). Human-Level Artificial Intelligence? Be Serious!. *AI Magazine*, Winter 2005.

Norvig, P. & Russell, S. (2009). *Artificial Intelligence: A Modern Approach*. Third edition. Upper Saddle River: Prentice Hall.

Nourbakhsh, I. (2009). Ethics in Robotics. Lecture at Carnegie Mellon University. http://www.youtube.com/watch?v=giKT8PkCCv4 Accessed 15 January 2014.

Nourbakhsh, I. (2013). *Robot Futures.* Cambridge: MIT Press.

Reddy, R. (2006). Robotics and Intelligence Systems in Support of Society. *IEEE Intelligent Systems*, 21(3), 24-31.

Stephenson, N. (2011). Innovation Starvation. *World Policy Journal*, 28, 11-16.

Stilgoe, J. et al. (2013). Developing a framework for responsible innovation. *Research Policy*, dx.doi.org/10.1016/j.respol.2013.05.008

Takayama, L. et al. (2008). Beyond Dirty, Dangerous, and Dull: What Everyday People Think Robots Should Do. *HRI* '08, Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, 25-32.

Von Schomberg, R. (2011). Prospects for Technology Assessment in a framework of responsible research and innovation. In *Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methode.* Wiesbaden: Vs Verlag.

Wallach, W. & Allen, C. (2010). *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press.

Winfield, A. (2013). Ethical Robots: some technical and ethical challenges. Description and slides of a presentation at EUCog meeting, “Social and Ethical Aspects of Cognitive Systems.” http://alanwinfield.blogspot.com.es/2013/10/ethical-robots-some-technical-and.html Accessed 15 January 2014.