Three philosophical perspectives on the relation between technology and society, and how they affect the current debate about artificial intelligence

Poel, Ibo van de

DOI
10.1515/humaff-2020-0042

Publication date
2020

Document Version
Final published version

Published in
Human Affairs

HUMAN AFFAIRS *30*, 499–511, 2020 DOI: 10.1515/humaff-2020-0042

THREE PHILOSOPHICAL PERSPECTIVES ON THE RELATION BETWEEN TECHNOLOGY AND SOCIETY, AND HOW THEY AFFECT THE CURRENT DEBATE ABOUT ARTIFICIAL INTELLIGENCE

IBO VAN DE POEL

Abstract: Three philosophical perspectives on the relation between technology and society are distinguished and discussed: 1) technology as an autonomous force that determines society; 2) technology as a human construct that can be shaped by human values, and 3) a co-evolutionary perspective on technology and society where neither of them determines the other. The historical evolution of the three perspectives is discussed and it is argued that all three are still present in current debates about technological change and how it may affect society. This is illustrated for the case of Artificial Intelligence (AI). It is argued that each of the three perspectives contributes to the debate of AI but that the third has the strongest potential to uncover blind spots in the current debate.

Key words: Technology; society; philosophy; technological determinism; values; co-evolution; artificial intelligence.

Introduction

Philosophical reflection on technology is perhaps as old as humanity. Since the early days of human evolution, humans have made and used tools for survival, and, henceforth, sometimes have been characterized as *homo faber* (tool maker) (e.g. Bergson, 1911). When humans started to reflect philosophically may be harder to trace historically, but philosophical reflections on technology can at least be traced back to antiquity in the Western world, but are likely older (Franssen, Lokhorst, & Van de Poel, 2018). Nevertheless, as a specialized discipline, philosophy of technology is of a much more recent date. Ernst Kapp was probably the first to use the term in 1877.

In this essay, I am particularly interested in philosophical thinking about the relation between technology and society, and about that between technological change and social change. Consequently, in this article my focus is on what Mitcham (1994) has called humanities philosophy of technology rather than engineering philosophy of technology. The latter is more interested in issues in engineering practice, like the nature and evolution of

DE GRUYTER

technological artifacts, design, and the nature of technological knowledge, while the first focuses more on technology as a social, cultural and historical phenomenon, and its relation to society.

While humanities philosophy of technology may have started with writers like Lewis Mumford (Mumford, 1934) and Ortega y Gasset (Ortega y Gasset, 1939/1997), it is often thinkers like Martin Heidegger (Heidegger, 1962) and Jacques Ellul (Ellul, 1964) that are seen as the frontrunners. They expressed a view on the relation between technology and society that conceives of technology as an autonomous force that determines society. While this view has been, and still is, influential, particularly in more popular discussions about technology, it has now been largely surpassed in professional philosophy of technology by a view that has arisen since roughly the 1980s under influence of philosophers like Langdon Winner (e.g. Winner, 1980) and the rise of the field of Science and Technology Studies. According to this second view, technology is basically a human product shaped by human interests and values, and it can also be shaped by these according to human will. In addition to these two views, I will distinguish a third one, which is of a more recent date but also has older roots. This third view stresses the co-evolution of technology and society and recognises explicitly the sometimes self-contained character of technology, and its unexpected and unintended consequences.

In current societal debates, we find elements of all three views. In current popular discourse, for example, about the fear that Artificial Intelligence (AI) will take over from humans, we can clearly recognize the idea of technology as an autonomous and determinate force. However, in these debates, we also witness the articulation of a range of values, which should guide the development of AI (e.g. High-Level Expert Group on AI, 2019), which very much fits the second view.

Therefore, the three perspectives on technology and society that I sketch below also function as tropes, or figures of speech that we have recourse to when we try to understand technological change and how it relates to, or affects, social change. Each perspective comes with certain core assumptions that define certain development as threats, and others as opportunities. What in one mode of thinking may be seen as malleable and open to choice, in another mode may be seen as given and unchangeable. The different modes of thinking about technology and society are therefore not innocent: they help to determine not only how we interpret technology and its relation to society but also what we see as possible and desirable.

Below, I shall briefly discuss each of the three perspectives, and their intellectual history, and will then illustrate for the case of AI that the perspectives are still present in current debates, although the third perspective seems somewhat underrepresented.

Technology as autonomous and determinate force

The idea of technology as an autonomous force that determines society and societal change can be found in early philosophers of technology like Jacques Ellul and Martin Heidegger. Ellul in his book *The Technological Society* (La Technique) describes technology as an autonomous force that develops largely independently from human choices (Ellul, 1964). For Ellul, technology stands for a certain way of relating to reality and for certain values, pre-eminently efficiency.

Like Ellul, Heidegger in his essay *The Question Concerning Technology* (Die Frage nach der Technik) is not so much interested in specific technologies but in Technology, with a big T, as a certain way of relating to, and perceiving reality (Heidegger, 1962). For him, Technology represents, in essence, an instrumental relation to reality, in which everything—nature, fellow humans—appears as a resource or a means to an end.

Somewhat similar ideas can already be found in an earlier book by Karl Jaspers called *Man in the Modern Age* (Die Geistige Situation der Gegenwart) (Jaspers, 1931/1933), and in the work of Günther Anders, who, in his work *Die Antiquiertheit des Menschen* (The outdatedness of humans), of which the first volume appeared in 1956, stresses what he calls a growing a-synchronicity between humans and modern technology. Lewis Mumford in his writings, on what he calls modern monotechnics, also leans to describing technology as an autonomous and determinate force (Mumford, 1967). The idea of technology as a modern ideology is also very much present in the Frankfurt School of Philosophy, for example, in Herbert Marcuse's book *One-dimensional man* (Marcuse, 1964) and Jurgen Habermas' essay on science and technology as ideology (Habermas, 1968).

All these philosophers perceive of technology as a more or less autonomous force, that cannot, or at least not easily, be resisted. Moreover, they associate technology, and in particular modern technology, with certain values and a certain relation to reality that is increasingly becoming dominant due to the autonomous force of technology. In line with these two ideas, the first perspective on technology and society may be characterized by the following two key assumptions:

1, Technology develops autonomously, i.e. according to its own laws, not, or hardly, open to human choice;

2, The impact of technology on society is deterministic.

The combination of these assumptions can not only be found among techno-pessimists like Ellul and Heidegger but also among techno-optimists. While it is hard to find a contemporary philosopher that represents such a view, it is clearly present in popular culture and among non-philosophers.[1] For example, Smith (1994) describes technological determinism in American culture in the late nineteenth and early twentieth century. She writes that the "belief that in some fundamental sense technological developments determine the course of human events had become dogma by the end of the [19th] century" (Smith, 1994, p. 7). This belief was, among others, installed by advertisements, "[f]from the early 1900s onwards, advertising agencies sold the public on the idea that the latest advances in technology brought not only immediate personals gains but also social progress" (Smith, 1994, p. 19).

The ideas of autonomous technology and technological determinism are still very much vivid today, as is witnessed by statements that we hear all too often, such as: "technological progress is inevitable"; "new technologies will eventually be used anyway"; "we cannot un-invent technologies once the genie is out of the bottle"; and "we will need to adapt to new technological realities." While such ideas are often coupled to a faith that technology will bring progress, we certainly also find today the techno-pessimistic view in popular culture and among non-philosophers. An example is the essay *Why the Future Does not Need Us* in

[1] An example is perhaps Francis Bacon's incomplete novel The New Atlantis (1627).

which Bill Joy, then Chief Scientist at Sun Microsystems, voices the concern that advances in robotics, genetic engineering, and nanotechnology will lead to the destruction of mankind (Joy, 2000).

The idea of machines taking over is an important subtheme under the trope of technology as an autonomous and determinate force, and one that keeps returning over time. It can already be found in novels like Mary Shelley's *Frankenstein* (first published in 1818) and Samuel Butler's *Erewhon* (first published anonymously in 1872). It has also been regularly voiced by scientists and engineers. For example, in his book *Engines of Creation* (1986), the molecular nanotechnologist Eric Drexler sketches the doomsday scenario of minuscule nanobots that have run out of control and that keep replicating, eventually eating up all matter so that only 'grey goo' is left.

The main difference between these techno-pessimists and techno-optimists is the values that are associated with technological development and change. For techno-optimists these are positive values like social progress, economic prosperity, freedom and democracy. Techno-pessimists stress negative values like efficiency, instrumentalization, domination of humans, tyranny, alienation, and the end of mankind. Despite these diametrically opposed normative assessments of technological change, they both conceive technological development as an autonomous process that determines societal developments. Consequently, there are few possibilities for human choice in technological development.

For techno-optimists it is probably not a problem that technology seems beyond human control as the inevitable technological changes will eventually bring human progress and other positive values. Techno-pessimists often feel a need to offer some way out, but due to their conception of technology the only way out they usually see is to abandon technology altogether, or at least to abandon 'modern' technology and to revalue older 'more humane' forms of technology and 'non-technological' ways of relating to reality, like religion in Ellul's writings and poetry for Heidegger. It is here that the second perspective on technology and society offers radically different, and much more diverse, options for remedying potential negative effects of technology and technological change.

Technology as a human product shaped by human interests and values

The second perspective on technology stresses the human-made character of technology. Technology is a human construct. Consequently, technologies are shaped by human interests and values, and open to human choice. This view can already be found in some of the works of earlier philosophers of technology like Lewis Mumford (e.g. Mumford, 1934/1963) and Langdon Winner (e.g. Winner, 1977, 1986). Winner, for example, in his book *Autonomous Technology* draws attention to how technology often appears as an autonomous and determinate force (along the lines of the first approach), but he believes this to be rooted in our (mis)conception of technology rather than in the essence of technology, as Ellul and Heidegger held (Winner, 1977).

One of the earliest and strongest expressions of the view of technology as a human and value-laden product is probably Langdon Winner's essay *Do artifacts have politics?* Winner (1980) argues that technological artifacts have political qualities and, hence, are value- and power-laden. Winner makes a distinction between technologies that are, in his view, by their very nature politically-laden, like for example the atomic bomb that according to him requires an authoritarian political structure to control its risks, and technologies that have politics due to their specific design, which could have been chosen differently, for example "concrete buildings and huge plazas constructed on university campuses in the United States during the late 1960s and early 1970 to defuse student demonstrations" (Winner, 1980, p. 124). It is particularly examples of the latter kind that fit the thinking of the second perspective (and that have been mostly referred to by later authors).

The idea of technology as a human construct has also been strongly articulated in Science and Technology Studies (STS), in particular in more constructivist approaches. In line with constructivist approaches to science (e.g. Bloor, 1976; Latour & Woolgar, 1979; Collins, 1985), since the 1980s different models and theories for understanding technology and technological change as human constructs have been developed (see e.g. Bijker, Hughes, & Pinch, 1987).

One early example is the SCOT, Social Construction Of Technology, model developed by Wiebe Bijker (see e.g. Bijker, 1995). According to this model, technological artifacts are interpreted differently by different social groups; such interpretations also typically suggest different paths for further technological development. Depending on what interpretation becomes dominant, technological change will take different paths. What is striking about the SCOT model is that factors such as the state of technological knowledge or what is technically feasible at a certain time do not seem to have an independent place in the model. Technology is just a human construct.

The view of technology as a human product open to, for example, design choices has increasingly been accepted by philosophers of technology and it can, in different degrees, be found also among the second and third generation of philosophers influenced by Heidegger like, for example, Don Ihde, Alfred Borgmann, Andrew Feenberg and Peter-Paul Verbeek (e.g. Ihde, 1993; Borgmann, 1984; Feenberg, 1991; Verbeek 2011).

Although there are, of course, many nuanced differences between thinkers that roughly fit this second perspective, I think one can fairly state that often the following three assumptions are present:

1, Technology is a human product or social construction and, as such, open to human choices;

2, Technology is value-laden, and different products can embed different values, depending on their design;

3, We (can) shape new technologies by our interests and values.

Whereas thinkers in the first approach typically have a monolithic idea about Technology, with a big T, in this second perspective they tend to talk about technologies in the plural, and to stress that the normative assessment of technologies may be very different from one technology to another. Different technologies have different normative qualities depending on choices made by humans, for example during the design process.

We find this second view also in popular media, in policy circles, and among non-philosophers. An example is the science-fiction writer Isaac Asimov who has become well-known for his three laws of robotics (Asimov, 1950):

1, A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2, A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3, A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The idea behind these laws is not that robots will automatically obey them, but that they should be so designed (by humans) that they do. This clearly fits the idea of technology as a human product.

The idea that we can deliberately design values into technology is also key to Value Sensitive Design (Friedman & Kahn, 2003; Friedman & Hendry, 2019), and related approaches like Design for Values (Van den Hoven, Vermaas & Van de Poel, 2015) that aim at incorporating values of ethical importance into the design of new technologies.

This way of thinking has also been taken up in policy circles. Initially, it was particularly new variations to technology assessment that made this approach more widespread. Whereas traditional technology assessment was aimed at predicting and assessing the consequences of new technologies, modern varieties like Constructive Technology Assessment (Rip, Misa & Schot, 1995) and real-time technology assessment (Guston & Sarewitz, 2002) stress the importance of feeding insights from the anticipated possible consequences of new technologies into the design and research and development (R&D) phase of new technologies, so that better technologies can be developed.

These ideas have also been taken up in approaches like Mid-Stream Modulation (Fisher, Mahajan & Mitcham, 2006) and, more recently, in the approach of Responsible Research and Innovation (RRI) (Owen, Bessant & Heintz, 2013). RRI has particularly been taken up in the European Union where it has been defined as “the on-going process of aligning research and innovation to [sic] the values, needs and expectations of society” (European Commission 2014).

Co-evolution of Technology and Society

The second perspective is more nuanced than the first in the sense that it focuses on technologies rather than monolithically on technology, its essence or its inevitable development. In this way, it also opens possibilities for more nuanced normative assessments of technologies and suggests constructive ways for better governing the development of new technologies.

However, the second perspective may also be in danger of overstating the degree to which we can steer or direct technological developments. In addition to the first and second perspective, which stress, respectively, the autonomy of technology and human choice in technology, we might distinguish a third perspective that theoretically starts from the idea of co-evolution of technology and society (Rip & Kemp, 1998).

The idea of co-evolution is certainly not new and many authors I cited for the second perspective may well subscribe to it. What sets the third perspective apart from the second, then, is not just or merely a recognition of the co-evolution of technology and society but rather a recognition of what I would like to call the *non-malleability* of technology, and of the fact that technology brings *novelty* and, therefore, unforeseen and unintended (social) consequences.

Let me explain these terms in some more detail. With non-malleability, I do not mean so much to refer to the laws of physics (although these puts limits on what is technologically feasible), but rather to the fact that technological developments are often hard to govern. In the literature, this has been understood in a variety of ways, for example in terms of technological complexity and scale (Collingridge, 1992), in terms of technological momentum (Hughes, 1994), in terms of path-dependence and lock-in (David, 1985; Arthur, 1989) or in terms of technological regimes (Nelson & Winter, 1977; Rip & Kemp, 1998; Van de Poel, 2003). What all these explanations have in common is that they see the non-malleability of technology not as purely technical in nature, but as, at least partly, social in nature. It is due to organizational complexities, economic considerations, power constellations, social institutions, and the like.

Technological development, according to the third perspective, is not just non-malleable but it also brings novelty (Rip & Kemp, 1998). With novelty, I mean here that it creates something that did not exist before and that we cannot fully capture beforehand. This novelty is an opportunity; it may, for example lead to new options that help to solve social and moral dilemmas (Van den Hoven, Lokhorst & Van de Poel, 2012). However, the novelty is also a potential threat in the sense that it may lead to unintended and undesirable risks or side-effects.

- We can, again, summarize the third perspective in terms of three key assumptions: Technology and society co-evolve; technology does not determine society nor do societal choices fully determine technology;
- Some aspects of technology development are very hard to change or unmalleable, and are hardly (still) open to human choice;
- Technology creates novelty and unexpected (and unintended) consequences.

The combination of the second and third assumption results in the so-called dilemma of technological control that was already formulated in Collingridge (1980). This dilemma states that, in its early phases, a new technology is still malleable, but one lacks sufficient knowledge about its social impact to steer it in the right direction. Later, when this knowledge has become available, the technology is so well-entrenched in society that it has become hard or impossible to change anymore.

The dominant response to this dilemma in the second perspective is to try to proactively steer technology during its early phases, like the R&D and design phase, while addressing the knowledge problem by increased anticipation and deliberation (e.g. through stakeholder involvement). While such an approach is certainly sensible, it also runs the risks of overlooking issues and concerns that are hard, or impossible, to anticipate at these early stages.

An alternative to the anticipatory approach, and more in line with the third perspective, is an approach that addresses the other horn of the Collingridge dilemma by trying to avoid, or at least postpone, the lock-in of a new technology and using this time for an extended period of experimentation and learning about a new technology (Van de Poel, 2016). Such an approach would conceive of the introduction of a new technology into society as a form of social experimentation (Krohn & Weyer, 1994; Felt et al., 2007) and would seek better, and more acceptable, forms of social experimentation with technology. Rather than on anticipation, the emphasis is in such an approach on experimentation, adaptability and learning.

In the light of the third perspective, the RRI approach aimed at better aligning technology with the 'values, needs and expectations of society' is sympatric, but perhaps somewhat naïve, both in the sense that technology may be harder to govern than expected and that even, when such governance efforts are successful, unexpected consequences and unpleasant surprises will occur from time to time, due to the novelty of technology.

Moreover, from a co-evolutionary point of view, the 'values, needs and expectations of society' are not given but evolve, as a result of technological development. They, hence, do *not* provide a normative rock bottom that can guide technological development. While many may subscribe to such a statement where it concerns societal needs and expectations, some would probably believe that values are more stable and unchangeable and have, or at least can have, a solid normative foundation. Ethicists of technology have, however pointed out that technological developments may induce technomoral change (Swierstra, 2013) or value change (Van de Poel, 2018a).

Acknowledging the possibility of value change may, occasionally, lead to a form of moral relativism, but it does not necessarily imply a moral relativist position. One may, as well, argue that new technologies create new realities and new types of moral situations and, hence, new moral problems, that demand new moral values to adequately deal with them. When the idea that technology creates novelty is taken seriously, this is a real possibility. So conceived, the introduction of new technology is also a form of moral experimentation, in which we only along the way find out what the new moral issues created by a new technology are, and, along the way, (re)invent the moral standards and values by which to judge that technology (Van de Poel, 2018b).

AI as an example

I want to end this essay with a brief example of how the three philosophical perspectives on technology and society play out in a concrete case, namely Artificial Intelligence (AI). This helps to see how all three are still present and relevant today, as well as to suggest how my discussion of them might be relevant to deal with the new challenges to society that AI introduces.

The first perspective, technology as an autonomous and determinate force, is clearly visible in techno-optimist as well as techno-pessimist visions on AI. On the one hand, there are scientists, governments and industries sketching AI largely as an inescapable development that will bring economic and social progress. The argument is often that we should free up large amounts of money for AI in order not to be surpassed by competitors who will do the same, so contributing to a self-fulfilling prophecy.

Also, many techno-pessimists have recourse to the first perspective. An example is Stephen Hawking's warning that AI could end mankind (Cellan-Jones, 2014), a fear that is now also voiced it in many popular articles about AI and machine learning. Similar fears are voiced in Nick Bostrom's recent book *Superintelligence* (Bostrom, 2016). While these voices might help to point out some of the potential perils of AI and can, hence, be seen as a form of early warning, their analyses bring relative few insights as to how to improve the development of AI, and its social impacts.

Here the second perspective has proven more useful. The earlier mentioned report of the High-Level Expert Group on AI (2019) is a clear example of the application of the second perspective; it articulates the ethical principles of respect for human autonomy, prevention of harm, fairness and explicability as the ones that should guide the development and design of AI. These ideas are backed by more scholarly work on responsible AI, Humane AI, explainable AI, and meaningful human control (e.g. Floridi et al., 2020; Wachter, Mittelstadt & Russell, 2018; Santoni de Sio & Van den Hoven, 2018). So, the second perspective is now clearly showing its relevance for better governing the development of AI.

However, the second perspective may – in line with my general discussion above – have two important blind spots. One is that the actual control of AI developments may be much harder than expected, or at least hoped. One problem, as also alluded to by Bostrom, is that AI may give countries a competitive advantage not only economically but also in warfare, which might make it particularly hard to control, especially because some governments are clearly much less interested in developing AI in a responsible manner than others.

Another blind spot may be the novelty and unintended consequences brought by AI. These are partly due to the fact that AI allows the design of artificial agents that are autonomous and adaptive, and can hence learn – often in unpredictable ways – from interaction with their environment (cf. Floridi & Sanders, 2004). Moreover, AI may bring social and economic disruption, for example in terms of (un)employment, but also conceptual and moral disruption, as it challenges some key philosophical notions like (human) moral agency and responsibility. The tendency of techno-pessimists, in the first approach, and of those adhering to the second perspective is often to reject such disruptions, and the accompanying AI technologies, either because they are seen as a peril to humanity (first perspective) or because they endanger some of our core human values (like human autonomy, and responsibility) (cf. van Wynsberghe & Robbins, 2019).

Here a somewhat more nuanced view might prove fruitful. On the hand, AI does not only bring threats but also opportunities, and some conceptual and moral changes may be desirable, not because they are triggered by AI but because we have *independent* (philosophical) reasons to consider them good or desirable. What the third perceptive also adds to the other two is a stronger emphasis on the co-evolution of AI and society, and, hence, on developing AI technologies that support humans rather than replace them. Like most other technologies, AI may well improve human capabilities and contribute to a better society.

Given the uncertainties and opportunities as well as threats that surround the development of AI, and in line with the third perspective, one should also aim at a more gradual introduction of AI into society, in which it does not only amount to an uncontrolled de-facto social and moral experiment, but in which we can (first) apply more small-scale and guided forms of social and moral experimentation that allow us to learn and adapt along the way. Such an approach to AI may sound idealistic, but in other realms of technologies, like the medical, we have come to accept over time that new treatments, drugs or vaccines first need to be tested out extensively before they can be safely and responsibly introduced on a larger scale in society.

Conclusion

Three philosophical perspectives on the relation between technology and society can be distinguished. Very roughly, they either interpret technology as the determining force in this relation (first perspective) or view humans and society as the determining force (second perspective) or start from the idea of co-evolution of technology and society (third perspective). As was illustrated for the case of AI, these perspectives are all present in current debates about new technologies. The perspectives can therefore be seen as cultural resources that people have recourse to in debates about such technologies. However, that does not mean that they are equally adequate or desirable from a normative point of view. I believe the third perspective, co-evolution between technology and society, to be preferable for at least two reasons. One is that it able to integrate insights from the other two; the other reason is that I consider it to be more descriptively adequate, although that it admittingly hard to demonstrate. Moreover, as the case of AI suggests, it is able to point out blind spots in current debates in which the first and second perceptive are often still more dominant.

Acknowledgement: This publication is part of the project ValueChange that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 788321.

References

Anders, G. (1956/1980). *Die Antiquiertheit Des Menschen.* [The outdatedness of humans]. Vol. 1. München: Beck.

Arthur, W. B. (1989). Competing technologies, Increasing returns, and lock-in by historical events. *The Economic Journal* 99 (394):116-131. doi: 10.2307/2234208.

Asimov, I. (1950). *I, Robot.* 1st ed. New York: Gnome Press.

Bacon, F. (1627). *New Atlantis: A work unfinished.* London: William Lee.

Bergson, H. (1911). *Creative evolution.* Translated by A. Mitchell. New York: Holt.

Bijker, W. (1995). *Of bicycles, bakelite, and bulbs: Toward a theory of sociotechnical change.* Cambridge (Ma.): MIT Press.

Bijker, W, T. P. Hughes, & T. Pinch, eds. (1987). *The social construction of technological systems: New directions in the sociology and history of technology.* Cambridge (Ma.): MIT Press.

Bloor, D. (1976). *Knowledge and social imagery.* London and Boston: Routledge & Kegan Paul.

Borgmann, A. (1984). *Technology and the character of contemporary life: A philosophical inquiry.* Chicago/London: University of Chicago Press.

Bostrom, N. (2016). *Superintelligence: Paths, dangers, strategies.* Oxford and New York: Oxford University Press.

Butler, S. (1872). *Erewhon; or, over the Range.* London: Trübner.

Cellan-Jones, R. (2014, December 2). *Stephen Hawking warns artificial intelligence could end mankind.* BBC. https://www.bbc.com/news/technology-30290540.

Collingridge, D. (1980). *The social control of technology.* London: Frances Pinter.

Collingridge, D. (1992). *The management of scale: Big organizations, big decisions, big mistakes.* London and New York: Routledge.

Collins, H. (1985). *Changing order: Replication and induction in scientific practice.* London: Sage.

David, P. A. (1985). Clio and the economics of QWERTY. *American Economic Review* 75 (2), 332–337.

Drexler, K. E. (1986). *Engines of creation: The coming era of nanotechnology*. New York: Anchor Books.

Ellul, J. (1964). *The technological society*. (J. Wilkinson, Trans.). New York: Alfred A. Knopf. (Original edition, La Technique).

European Commission. (2014). Rome declaration on responsible research and innovation in Europe. https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf.

Feenberg, A. (1991). *Critical theory of technology*. New York: Oxford University Press.

Felt, U., Wynne, B., Callon, M., Gonçalves, M. E., Jasanoff, S., Jepsen, M., Joly, P.-B., Konopasek, Z., May, S., Neubauer, C., Rip, A., Siune, K., Stirling, A., & Tallacchini, M. (2007). Taking European knowledge society seriously. Report of the Expert group on science and governance to the Science, economy and society directorate, Directorate-general for research, European Commission. Brussels: Directorate-General for Research, Science, Economy and Society.

Fisher, E., Mahajan, R. L., & Mitcham, C. (2006). Midstream modulation of technology: Governance from within. *Bulletin of Science, Technology & Society, 26* (6), 485–496. doi: 10.1177/0270467606295402.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*. doi: 10.1007/s11948-020-00213-5.

Floridi, L., & Sanders J.W. (2004). On the morality of artificial agents. *Minds and Machines* 14(3), 349–379.

Franssen, M., Lokhorst, G.-J., & Van de Poel, I. (2018). Philosophy of technology. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Fall 2018 Edition).*

Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. Cambridge (Ma.): MIT Press.

Friedman, B., & Kahn, P. H., Jr. (2003). Human values, ethics and design. In J. Jacko & A. Sears (Eds.), *Handbook of human-computer interaction*, 1177–1201. Mahwah, NJ: Lawrence Erlbaum Associates.

Guston, D. H., & Sarewitz, D. (2002). Real-time technology assessment. *Technology in society 24* (1–2), 93–109. doi: Doi: 10.1016/s0160-791x(01)00047-1.

Habermas, J. (1968). *Technik und Wissenschaft als "Ideologie"*. [Technology and science as "ideology"]. Frankfurt: Surhrkamp.

Heidegger, M. (1962). *Die Technik und die Kehre*. [Technology and the Turn]. Pfullingen: Neske.

High-Level Expert Group on AI. (2019, April 8). Ethics guidelines for trustworthy AI. Brussels: EC. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Hughes, T. P. (1994). Technological momentum. In M.R. Smith & L. Marx (Eds.), *Does technology drive history? The dilemma of technological determinism* (pp. 115–142). Cambridge (Ma.) and London: MIT Press.

Ihde, D. (1993). *Philosophy of technology: An introduction*. Saint Paul: Paragon.

Jaspers, K. (1931/1933). *Man in the modern age*. (E. Paul & C. Paul, Trans.). London: Routledge.

Joy, B. (2000). Why the future doesn't need us. *Wired,* April 2000, 238–262.

Kapp, E. (1877/2018). *Elements of a Philosophy of technology: On the evolutionary history of culture*. (L.K. Wolfe, Trans.). Minneapolis and London: University of Minnesota Press.

Krohn, W., & Weyer, J. (1994). Society as a laboratory. The social risks of experimental research. *Science and Public Policy, 21*(3), 173–183.

Latour, B. & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills: Sage.

Marcuse, H. (1964). *One-dimensional man; Studies in the ideology of advanced industrial society*. Boston: Beacon Press.

Mitcham, C. (1994). *Thinking through technology. The path between rngineering and philosophy*. Chicago and London: University of Chicago Press.

Mumford, L. (1934/1963). *Technics and civilization*. New York: Harcourt Brace Jovanovich.

Mumford, L. (1967). *The myth of the machine*. 1st ed. 2 vols. New York: Harcourt.

Nelson, R. R. & Winter, S. G. (1977). In search for a useful theory of innovation. *Research Policy,* 6, 36–76.

Ortega y Gasset, J. (1939/1997). *Meditación de la Técnica*, *Filosofía Hoy*. [Meditation on technics]. Madrid: Santillana.

Owen, R., Bessant, J. R., & Heintz, M. (2013). *Responsible innovation: Managing the responsible emergence of science and innovation in society*. Chichester: John Wiley.

Rip, A. & Kemp, R. (1998). Technological change. In S. Rayner & E.L. Malone (Eds.), *Human choice and climate change* (pp. 327–399). Columbus, Ohio: Battelle Press.

Rip, A., Misa, T. J., & Schot, J. (Eds.) (1995). *Managing technology in society: The approach of constructive technology assessment*. London and New York: Pinter.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI 5*(15). doi: 10.3389/frobt.2018.00015.

Shelley, M. W. (1818). *Frankenstein; or, the Modern Prometheus*. London: Lackington, Hughes, Harding, Mavor, & Jones.

Smith, M. R. (1994). Technological determinism in American culture. In M.R. Smith and L. Marx (Eds.), *Does technology drive history? The dilemma of technological determinism* (pp.1–35). Cambridge (Ma.) and London: MIT Press.

Swierstra, T. (2013). Nanotechnology and technomoral change. *Etica & Politica / Ethics & Politics,* XV(1), 200–219.

Van de Poel, I. (2003). The transformation of technological regimes. *Research Policy 32*(1), 49–68.

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics,* 22(3), 667–686. doi: 10.1007/s11948-015-9724-3.

Van de Poel, I. (2018a). Design for value change. *Ethics and Information Technology*. doi: 10.1007/s10676-018-9461-9.

Van de Poel, I. (2018b). Moral experimentation with new technology. In I. Van de Poel, D.C. Mehos & L. Asveld (Eds.), *New perspectives on technology in society: Experimentation beyond the laboratory* (pp. 59–79). Oxford and New York: Routledge.

Van den Hoven, J., Lokhorst, G.-J., & Van de Poel, I. (2012). Engineering and the problem of moral overload. *Science and Engineering Ethics, 18*(1), 143–155. doi: 10.1007/s11948-011-9277-z.

Van den Hoven, J., Vermaas, P.E., & Van de Poel, I. (Eds.) (2015). *Handbook of ethics and values in technological design. Sources, theory, values and application domains*. Dordrecht: Springer.

Van Wynsberghe, A., & Robbins,S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*(3), 719–735. doi: 10.1007/s11948-018-0030-8.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago; London: The University of Chicago Press.

Wachter, S., Mittelstadt, B. D. M., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*(2), 841–887.

Winner, L. (1977). *Autonomous technology. Technics-out-of-control as a theme in political thought*. Cambridge (MA): MIT Press.

Winner, L. (1980). Do artifacts have politics? *Daedalus,* (109), 121–136.

Winner, L. (1986). *The whale and the reactor; A search for the limits in an age of high technology*. Chicago and London: The University of Chicago Press.

Ethics and Philosophy of Technology Section
Values, Technology & Innovation Department
School of Technology, Policy and Management
TU Delft
Jaffalaan 5
2628 BX Delft
The Netherlands
Email: i.r.vandepoel@tudelft.nl