Addressing the soft impacts of weak al-technologies

Addressing the Soft Impacts of Weak AI-Technologies

Katleen Gabriels

Assistant Professor, Eindhoven University of Technology, The Netherlands
k.gabriels@tue.nl

Abstract

This paper argues that, in addition to assessing a design's *hard* impacts in terms of safety, environmental issues, and health, also *soft* impacts, such as ethical implications, need to be systematically and proactively addressed. This is particularly important for (weak) AI-technologies that are increasingly shaping today's society. When Microsoft released its AI chatbot Tay on Twitter in March 2016, Tay was supposed to learn to chat as an average American teenage girl. However, she quickly became sexist, racist, and anti-Semitic. Microsoft turned out to be overly naïve about the intentions of Twitter users who had to 'train' Tay and, in doing so, the developers did not properly acknowledge their ethical responsibility. Even though technology's non-neutrality has become generally accepted within the fields of ethics and philosophy of technology, other disciplines, such as engineering and computer science, often still adhere to the view that technology itself is neutral. Yet, technology mediates our actions and perceptions in numerous ways. Today, algorithms not only have an important share in how we see the world, they can also predict our future behaviour. Neither algorithms nor datasets are inherently neutral; on the contrary, users' and developers' biases can seep into them. Overall, this paper seeks to give attention to the non-neutrality of AI-technologies, the ethical responsibility of its developers, and the soft impacts of existing and emerging AI-technologies. In doing so, three ways are discussed in which the agenda of technology developers can be broadened with a more systematic focus on assessing soft impacts.

Introduction: Guiding Concerns, Objectives, and Structure

Nowadays, a lot of attention goes to potential detrimental consequences of AI robots. Some newspaper articles are rather amusing. In July 2015, *The Guardian* wrote that a robot *killed* a worker at a Volkswagen plant in Germany (Associated Press in Berlin, 2015). To hold someone morally and legally accountable for murder, this 'person' must, among other things, act freely and have consciousness (see van de Poel and Royakkers, 2011, p. 11). Following these conditions, present-day robots cannot be hold accountable. Last summer, *The Telegraph* wrote about a security robot that *drowned itself* (Titcomb, 2017). Even though it is tempting to project anthropomorphic characteristics on robots, they do not have consciousness and free will (i.e. 'strong' AI). Still, these reactions may not surprise: among other things, AI 'forces' us to question our place in the world and to question the boundary between man and machine.

Yet, exaggerated articles on robots that are killing employees should not distract us from more pressing issues today. We are already encountering numerous concerns and problems with so-called 'weak' AI. Existing and emerging AI-technologies raise several ethical questions (see e.g. Tegmark, 2017). This paper discusses problems that we are already facing today. It focuses on the non-neutrality of technology in general and of algorithms in particular, and on the ethical responsibility of technology actors, such as developers and computer scientists. It is my contention that we need to broaden our frameworks with, first, the soft impacts of technology (as opposed to hard impacts that generally receive much more attention in Technology Assessments) and, second, the acknowledgement of the non-neutrality of technology. My guiding concern is that all too often technology is looked upon as a neutral instrument whereas in fact it is value-laden.

The particular lens through which this paper should be read is ethics and philosophy of technology. This paper unfolds into four parts. In the first part, I address the intrinsic relation between technology and morality. Drawing upon mediation theory (see a.o. Verbeek, 2006), I illustrate how technology mediates, first, our perceptions of reality and, second, our actions. In the second part, ethical impacts of machine learning are discussed. In specific, I elaborate on Microsoft's AI chatbot Tay (March 2016) in order to show the importance of proactively focusing on soft impacts and, in doing so, of the ethical responsibilities of developers. In the third part, I present the distinction between hard and soft impacts of technology and I subsequently offer three recommendations in which soft impacts can be more systematically addressed: first, by including them by default in Technology Assessments; second, by incorporating compulsory courses on (computer) ethics in the training and education of developers and computer scientists; and finally, by designing ethical values *into* technology, such as by means of interdisciplinary approaches such as Value Sensitive Design. Some summarizing and concluding thoughts are formulated in the fourth and final part.

On the Intrinsic Relation Between Technology and Morality

In this first part, focus lies on the intrinsic relation between technology and values. Its main objective is to illustrate that technology is never merely a neutral instrument: it is value-

laden, mediates our perception of reality and our behaviour (cf. mediation theory), and influences our norms and values.

In its broadest sense, technology refers to any object that is made by humans (Murphie and Potts, 2003, p. 4). Yet, technology is more than a collection of artefacts, as it also refers to the practices and processes around it and its embeddedness in social structures (Murphie and Potts, 2003, p. 4; see also Bijker et al., 1987, p. 4). Design also conveys cultural and moral dimensions and biases (Murphie and Potts, 2003). 'Gendered technology' is an example of how technology can be loaded with cultural stereotypes; ladyshaves, for example, are often designed in pink, a design choice that conveys stereotypes about women.

The first iPhone entered the market in 2007. In just eleven years it shaped present-day society and has had a major impact on our lives. The smartphone shapes our behaviour, for example in self-presentation: everyday, millions of selfies are taken and shared. Furthermore, it shapes our norms, among other things, on immediacy, on being available, and on 'phubbing', i.e. "the act of snubbing someone in a social setting by concentrating on one's phone instead of talking to the person directly" (Chotpitayasunondh and Douglas, 2016, p. 9). There is research showing that phubbing is increasingly becoming the norm (see Chotpitayasunondh and Douglas, 2016).

Even though the non-neutrality of technology has become generally accepted within the fields of ethics and philosophy of technology, other disciplines, such as engineering and computer science, still often adhere to the view that technology itself is neutral. Mediation theory (see e.g. the works of philosophers of technology Don Ihde (1990) and Peter Paul Verbeek (2006; 2008)) analyses how technologies mediate the way we perceive reality and the way we act.

First, *mediation of perception* discloses how artifacts mediate the way we see the world and the way in which we interpret reality (Verbeek, 2006, p. 364). For instance, a microscope made bacteria visible, which we cannot see with the naked eye. This eventually led to changed views on hygiene. A telescope amplifies the universe, depicting how alone we actually are in the galaxy. Instead of disclosing reality and bringing it 'closer', technology can also increase distances, both in physical and moral terms (cf. 'moral distance'). Nowadays, ample academic attention goes to distancing technologies in warfare, such as drones. For instance, Coeckelbergh (2013) argues among other things that drone fighting, as an example of long-range fighting, creates increasing moral distance because 'screenfighting' makes the perception of the other disappear. When the enemy is depicted as a video image, this might lower the barriers to kill.

Second, *mediation of action* focuses on how technologies "mediate people's actions and the way they live their lives" (Verbeek, 2006, p. 366). A well-known example is the speed bump that urges drivers to slow down. In doing so, technology not only plays an active part in shaping our environment but also in shaping our behaviour. Design is often loaded with values and moral choices: for instance, the speed bump is designed to increase road safety. Related examples are alcohol locks in cars and the alarm sound that goes off as long as you do not put on your seatbelt in the car. In addition to moralizing people in order to obtain more desirable behaviour, moral choices and values can be designed and embedded *into* technology. An example of the so-called 'moralization of technology' (see Achterhuis, 1995) is the water-saving shower: in addition to telling people to spill less water, the design also offers a solution to the problem.

Technology can also mediate our behaviour in 'hidden' ways, without us even realizing (see Brey, 2004). A series of controlled experiments on manipulated rankings in a search engine, conducted by Epstein and Robertson (2015), revealed that top hits can influence people's voting preferences. In one of their experiments that focused on political elections, informants who did not have a preferential candidate *before* the experiment, turned out to significantly have a preferential candidate *after* the test: the candidate who always appeared first in the ranking. Up to 75% of the informants who were exposed to biased and manipulated results were not aware of the manipulation: they tended "to believe they have adopted their new thinking voluntarily" (Epstein and Robertson, 2015, web). Today, computer technology can increasingly affect our autonomy (see e.g. Kerr, 2010). Epstein and Robertson (2015) conceptualize their research outcome as the 'Search Engine Manipulation Effect' (SEME). Their findings are particularly worrying now algorithms increasingly make decisions for us, while their specific procedures generally remain secret (see also Pasquale, 2015). Epstein and Robertson's research results are also particularly interesting concerning present-day debates on the influence of data analytics to target voters, in particular, the role of Cambridge Analytica in the 2016 presidential race in the USA (see e.g. Hakim and Rosenberg, 2018).

Furthermore, in a massive-scale experiment in 2014, Facebook deliberately manipulated the News Feeds of 689,003 users, who did not know they were subjected to an experiment, in order to study emotional contagion (see Kramer, Guillory, and Hancock, 2014). Facebook and the involved researchers were widely criticized for this experiment, because they did not obtain informed consent and did not give informants the possibility to opt out. The research outcomes showed that Facebook users who are exposed to negative news on their News Feeds are more likely to share negative information themselves. The opposite holds true as well: being exposed to positive information increases the likelihood that you will share positive messages and updates.

Technology shapes the context in which we behave and make choices. The more data are being collected of people, the better companies can profile and target them. This creates a 'paradox of transparency': users and customers are increasingly becoming transparent for companies while the specific ways in which companies operate and in which their algorithms make decisions remain opaque.

Generally, engineering and designing are about shaping and, ideally, improving the world. Designers have a lot of power when they are making specific choices. For instance, there is a plethora of worrisome examples of design that can make people more addictive in order to spend more money, such as gambling machines (cf. 'addiction by design', see Schüll, 2014). In Verbeek's phrase (2006, p. 377): "The insight that technologies inevitably play a mediating role in the actions of users makes the work of designers an inherently moral activity". Especially because technology can also mediate our perceptions and behaviours into undesirable directions, it is important to give systematic attention to the

ethical responsibilities of developers in our algorithmic society. The focus of the next part lies on the non-neutrality of algorithms and machine learning, as a technique to develop AI.

On Learning Software and Developers' Responsibilities

Undoubtedly, algorithms are extremely helpful, especially in analysing complex data. In healthcare, for example, machine learning has already led to great promises. Esteva et al. (2017) trained a deep convolutional neural network (CNN) with a dataset of 129,450 clinical images of skin lesions. Their findings show that an AI is "capable of classifying skin cancer with a level of competence comparable to dermatologists" (Esteva et al., 2017, p. 115). Despite these promising outcomes, learning software can also lead to harmful and unfair results, especially when trained with biased datasets. In 2015, Google had to apologize because its algorithm wrongly labelled an image of African-Americans as 'gorillas' (see Grush, 2015). The Word2vec algorithm, which was used to find patterns in word embeddings in Google News articles, disclosed sexism: 'man is to computer programmer as woman is to homemaker' was one of the results. Bolukbasi et al. (2016) warn that blind application of machine learning might amplify biases already present in data.

Algorithms are not inherently or automatically unbiased (see e.g. O'Neil, 2017). Our biases can seep into the datasets, which can lead to harmful output, decided by algorithms. These unfair results can disadvantage underrepresented groups in society, such as black people. On a daily basis, millions of people inform themselves about the world through search engines, such as Google's recommender engine. Algorithms decide which information we will and will not see. A lot of factors influence the ranking you will be given, among other things your personal search history and Search Engine Optimization (SEO). Here as well, societal biases seep into the data: a well-known example, which led to public consternation in 2016, was the difference between searching for 'three white teenagers' (leading to top results of images of 'shiny happy' young white people) and searching for 'three black teenagers' (which showed mugshots in the top ranking) in Google Images (see Allen, 2016).

Humans shape the data that algorithms have to learn from and this can go wrong, as the case of Microsoft's AI chatbot Tay also illustrates. This case furthermore discloses the importance of ethical responsibility and of the awareness for 'soft' impacts in order to reduce potential harms by anticipating properly. When Microsoft released Tay on Twitter in March 2016, the chatbot was supposed to learn to chat as an average American teenage girl. She was trained by having actual conversations and interactions on Twitter. However, in doing so, she quickly became sexist, racist, and anti-Semitic. Some of her tweets included the following statements: 'bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got' and 'I fucking hate feminists and they should all die and burn in hell'. Microsoft turned out to be overly naïve about the intentions of users on Twitter who had to 'train' Tay and eventually had to shut down the system within 24 hours after its release.

Developers of computer technology have the responsibility not to cause harm, not only in legal but also in ethical terms. In several professional codes of conduct this is included as a moral imperative, such as the 'ACM Code of Ethics and Professional Conduct', "General Moral Imperative 1.2: Avoid harm to others"[i]. Microsoft did not properly assess and evaluate the potential impact of its system beforehand. As Wolf, Miller, and Grodzinsky (2017, p. 56) aptly remark: "Taking Tay offline when it became abusive was reactive, not proactive". Microsoft overlooked its ethical responsibility, especially because the learning software had to interact with an online audience. Tay was an experiment, conducted with the general public, and its harmful effects were immediately visible and directly impacted a lot of people.

On March 25, 2016, Peter Lee, Corporate Vice President Microsoft AI + Research, released an online statement on behalf of Microsoft, 'Learning from Tay's introduction' (see Lee, 2016). In the statement, he acknowledges that the research challenges in AI design "are just as much social as they are technical" (Lee, 2016, web). "We will remain steadfast in our efforts to learn from this and other experiences as we work toward contributing to an Internet that represents the best, not the worst, of humanity" (Lee, 2016, web). Machine learning is a dynamic process, of which the particular outcomes can indeed be hard to predict. Nevertheless, much more systematic attention must go to assessing soft impacts, especially because undesirable consequences can harm many people. Soft impacts are at the core of the next part.

Broadening the Framework with Soft Impacts

In order to assess risks involved with technological innovations, Risk and Technology Assessments were invented to evaluate undesirable impacts and subsequently trying to avert them by taking precautions (see Swierstra and te Molder, 2011, p. 1050). Yet, when assessing the risks of existing, new, and emerging technologies, attention generally goes to risks that qualify as *hard* impacts instead of on social and ethical impacts that qualify as *soft* impacts (see Swierstra and te Molder, 2011, p. 1050). Hard impacts can generally be characterized as 'objective', 'rational', 'public', 'factual', and 'neutral'. They focus on harms caused to health, the environment, and safety. Soft impacts are defined as 'subjective', 'personal', and 'value-laden', and are therefore more easily dismissible (Swierstra and te Molder, 2011, p. 1050). Soft impacts refer to the way in which technologies shape our behaviour, relationships, norms, values, expectations, et cetera (cf. supra, mediation theory).

The focus of hard impacts is narrower: hard impacts are clear and quantifiable. Furthermore, the causal link between the technology and its (hard) impact is direct and clear, which fits more easily into thinking favoured by policy makers. Soft impacts, on the other hand, are much more difficult to quantify, as the harms are less clear and more indirect.

In order to reduce undesirable consequences of weak AI-technologies, the agenda of developers (including research and education) should be broadened with a more systematic focus on soft impacts. In what follows, I discuss three ways ('recommendations') in which they can attain more organized attention: first, by including them by default in Technology

Assessments; second, by making (computer) ethics a standard course in the training and academic education of developers and computer scientists; and finally, by designing ethical values into technology by means of interdisciplinary approaches such as Value Sensitive Design (VSD).

First, soft impacts need to become a systematic part of Technology Assessments. Even though biased data might not be a direct threat to our health or environment, they can be harmful as previous examples illustrated. Also, current public debates on the role of Cambridge Analytica in political elections disclose big data's and algorithms' potential threats to democracy and political debate. From a broader historical perspective, concerns related to computer technology are not new. In 1950, the American mathematician Norbert Wiener already focused on the power of emerging cybertechnologies in *The Human Use of Human Beings: Cybernetics and Society*. Wiener underscored the importance of prioritizing humans. Another seminal book that highlighted the human consequences of computers is *Computer Power and Human Reason* by computer scientist Joseph Weizenbaum (1976). Throughout this book Weizenbaum warned for the risks of putting too much trust in machines without properly knowing how they work, which is echoed in present-day discussions on the opaque ways in which algorithms and neural networks operate. It is also quite ironic that academics have warned decades ago for several of the problems that we are currently facing. For instance, the American mathematician and computer ethicist James Moor already warned in 1985 for the negative consequences of computer technology that we are dealing with today, such as problems related to privacy and surveillance, because of computer technology's 'invisibility factor' (see Moor, 1985). Despite these warnings, a systematic approach to soft impacts is still easily dismissed. Therefore, in order to assess soft impacts of weak AI-technologies, interdisciplinary teams - consisting of developers, ethicists, cryptographers, legal counsellors, end-users, and so on - should be brought together to analyse an emerging technology proactively (instead of reactively). In so doing, taking a long-term perspective is important. Of course, we need to acknowledge that it is impossible to foresee every possible outcome. This problem has become known as the 'Collingridge dilemma' (see e.g. van de Poel and Royakkers, 2011), which refers to a methodological double-bind problem: before a technology is in use, there might be a lack of information to assess it properly, whereas, after a technology is widely dispersed there is a problem concerning power and control to solve potential undesirable impacts because of its scope. Nonetheless, even though it is impossible to predict 'every' impact, assessing soft impacts should become a systematic part of design and research. There are already some interesting initiatives that offer frameworks and guiding sets of questions to assess the social and ethical impacts of AI-technologies. An example is the 'Principles for accountable algorithms and a social impact statement for algorithms': its goal "is to help developers and product managers design and implement algorithmic systems in publicly accountable ways" (see FAT/ML, 2018, web). A group of academics and research scientists (working at Google Research and Microsoft Research) developed this list of principles and guiding questions in order to stimulate and increase awareness concerning developers' accountability.

Second, an ethical code for developers, computer scientists, and engineers no longer suffices: courses on (computer) ethics have to become a systematic part of their training and education, exactly because ethics and technology are not separate domains (cf. supra) and because of the far-reaching impacts of AI-technologies. Only very recently ethics courses for developers and engineers are receiving more public and academic attention. A recent article in *The New York Times* covers how universities like Harvard University, MIT, Stanford University, and the University of Texas at Austin are now introducing courses on ethics of computer science and AI (see Singer, 2018). Some universities have included compulsory courses on engineering ethics to technical students' curricula since several years (e.g. Eindhoven University of Technology), but this is not yet a widespread practice, especially not concerning computer ethics. In the same way as medical ethics has become an intrinsic part of the education of medical doctors, computer ethics should be incorporated in the curriculum of future developers, computer scientists, and engineers. By incorporating ethics courses that, among other things, address the increased accountability of designers, for instance concerning self-learning technologies, and that offer conceptual frameworks and theories to analyse specific cases and problems, students are trained to give more systematic attention to soft impacts.

Third, undesirable soft impacts can be reduced by seeking to design ethical values *into* computer technology. Responsible Research and Innovation (RRI) is receiving more attention, not only in academia but also by policy makers, for instance at the European Commission[ii]. A general idea underlying RRI is to design more socially responsible technology by incorporating ethical aspects in the design process. A specific example of RRI is Value-Sensitive Design, i.e. "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman, Kahn, and Borning, 2002, p. 1; see also Friedman, 1996; Flanagan, Howe, and Nissenbaum, 2008). A specific example of VSD is privacy enhancing technologies ('privacy by design'), for instance internet platforms that only allow strong passwords, in order to better protect users' privacy. VSD consists of an integrated methodology of three interrelated, iterative steps: conceptual, empirical, and technical. The first, conceptual stage consists of "philosophically informed analyses of the central constructs and issues relevant" to the technology under development (Friedman, Howe, and Felten, 2002, p. 2). Second, empirical social science studies (e.g. focus groups, questionnaires, …) are conducted in order to understand and gain increased insight into "the value-oriented perceptions and experiences of the direct and indirect stakeholders of a given system" (Friedman, Howe, and Felten, 2002, p. 3). This way, undesirable outcomes can already be reduced before the design process starts. And, finally, in the technical stage, focus lies among other things on how existing designs engender and protect ethical values, and on "how the identification of specific values can lead to new technical designs and mechanisms to support better those values" (Friedman, Howe, and Felten, 2002, p. 3).

Together, the three recommendations also offer a means to move away from the old paradigm in engineering and

computer science that focuses on the belief that engineering and designing technology is about functionality (i.e. a technology should work and function properly) and fulfilling legal requirements (see e.g. Verbeek, 2006). This paradigm ignores, amongst others, the non-neutrality of technology, its soft impacts, and the fact that designers have a lot of power in shaping people's behaviour when they are making specific decisions (cf. ethical responsibility and mediation theory). The ethical reflection should hence not come 'after' the design, but already before the start of the design process.

Concluding Thoughts

Throughout this paper I sought to show that technology's soft impacts, such as ethical and social implications, should be properly addressed beforehand. This is particularly relevant for (weak) AI-technologies that are increasingly shaping today's society. A number of cases, such as Microsoft's AI chatbot Tay, disclosed the importance of assessing soft impacts in advance, in order to minimize harm. I offered three ways in which soft impacts can be more systematically addressed: first, by including them by default in Technology Assessments; second, by making (computer) ethics a standard course in the training and academic education of developers, engineers, and computer scientists; and finally, by designing ethical values into technology by means of interdisciplinary approaches such as Value Sensitive Design. These three 'recommendations' are not exhaustive and are open to further discussion and debate.

In addition, this paper sought to illustrate that technology mediates our actions and perceptions in numerous ways, and that it also affects our norms and values. We often act less autonomously than we tend to think. Algorithms are increasingly permeating in our society and personal lives. No one before has lived in a society in which so much data is collected on a daily basis. Algorithms learn from all this data and make predictions and decisions: from self-driving cars, to healthcare, to the 'Internet of Toys', to drones in warfare, and so on. Neither algorithms nor datasets are inherently neutral. A particular problematic aspect of algorithms is that they can be manipulated (e.g. Facebook's study on emotional contagion). But even in cases without deception, the precise ways in which algorithms make decisions often remain opaque. AI-technologies subsequently raise compelling questions concerning core ethical values such as autonomy, justice, privacy, dignity, and so on (see also Royakkers et al., 2018). Because this paper focused on problematic cases, it might give the impression that my overall view is mainly negative; this is not true, many developers do take their responsibility.

To conclude, I fully realize that this paper did not specifically focus on artificial life developments. It, however, not only aimed to make a compelling case for addressing the soft impacts of AI-technologies, but also for more interdisciplinary collaborations between researchers, developers, scientists, and so on in order to create the best possible technologies, developments, models, and tools in the long run.

References

Achterhuis, H. (1995). De moralisering van apparaten. *Socialisme en democratie*, 52(1):3-12.

Allen, A. (2016). The 'three black teenagers' search shows it is society, not Google, that is racist. *The Guardian* online, June 10, 2016. https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet Accessed March 18, 2018.

Associated Press in Berlin (no author mentioned) (2015). Robot kills worker at Volkswagen plant in Germany. *The Guardian* online, July 2, 2015. https://www.theguardian.com/world/2015/jul/02/robot-kills-worker-at-volkswagen-plant-in-germany Accessed March 17, 2018.

Bijker, W. E., Hughes, T. P., and Pinch, T. J., editors (1987). *The Social Construction of Technological Systems. New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *arXiv*:1607.06520.

Brey, P. (2000/2004). Disclosive computer ethics. *Computers and Society*, 30(4):10-16.

Chotpitayasunondh, V., and Douglas, K. M. (2016). How "phubbing" becomes the norm: The antecedents and consequences of snubbing via smartphone. *Computers in Human Behavior*, 63:9-18. Doi: https://doi.org/10.1016/j.chb.2016.05.018

Coeckelbergh, C. (2013). Drones, information technology, and distance: Mapping the moral epistemology of remote fighting. *Ethics and Information Technology*, 15(2):87-98.

Epstein, E., and Robertson, R. E. (2015). The Search Engine Manipulation Effect (SEME) and its possible impact on the outcomes of elections. *PNAS*, 112(33):E4512-E4521. Doi: https://doi.org/10.1073/pnas.1419828112

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115-118. Doi: 10.1038/nature21056.

FAT/ML (2018). Principles for accountable algorithms and a social impact statement for algorithms. https://www.fatml.org/resources/principles-for-accountable-algorithms Accessed April 2, 2018.

Flanagan, M., Howe, D. C., and Nissenbaum, H. (2008/2010). Embodying values in technology: Theory and practice. In J. van den Hoven J., and Weckert, J., editors, *Information Technology and Moral Philosophy*, pages 322-353, Cambridge University Press, Cambridge.

Friedman, B. (1996). Value-sensitive design. *Interactions*:17-23.

Friedman, B., Kahn, P. H. Jr., and Borning, A. (2002). Value sensitive design: Theory and methods. *UW CSE Technical report*:1-8.

Grush, L. (2015). Google engineer apologizes after photos app tags two black people as gorillas. *The Verge*, July 1, 2015. https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas Accessed March 18, 2018.

Hakim, D., and Rosenberg, M. (2018). Data firm tied to Trump campaign talked business with Russians. *The New York Times* online, March 17, 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-russia.html Accessed March 18, 2018.

Ihde, D. (1990). *Technology and the Lifeworld*. Indiana University Press, Bloomington.

Kerr, I. (2010). Digital locks and the automation of virtue. In Geist M., editor, *"Radical Extremism" to "Balanced Copyrights":*

Canadian Copyright and the Digital Agenda, pages 247-303. Irwin Law, Toronto.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, 111(24):8788-8790. Doi: https://doi.org/10.1073/pnas.1320040111.

Lee, P. (2016). Learning from Tay's introduction. *Official Microsoft Blog*, March 25, 2016. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ Accessed March 18, 2018.

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, 16(4):266-275.

Murphie, A., and Potts, J. (2003). *Culture and Technology*. Palgrave MacMillan, New York.

O'Neil, C. (2016/2017). *Weapons of Mass Destruction. How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, New York.

Pasquale, F. (2015). *The Black Box Society. The Secret Algorithms that Control Money and Information*. Harvard University Press, Cambridge, MA.

Royakkers, L., Timmer, J., Kool, L., and van Est, R. (2018). Societal and ethical issues of digitization. *Ethics and Information Technology*. Doi: https://doi.org/10.1007/s10676-018-9452-x.

Schüll, N. D. (2012/2014). *Addiction by Design: Machine Gambling in Las Vegas*. Princeton University Press, Princeton, New Jersey.

Singer, N. (2018). Tech's ethical 'dark side': Harvard, Stanford and others want to address it. *The New York Times* online, February 12, 2018. https://www.nytimes.com/2018/02/12/business/computer-science-ethics-courses.html Accessed March 18, 2018.

Swierstra, T., and te Molder, H. (2011). Risk and soft impacts. In Roeser S., editor, *Handbook of Risk Theory*, pages 1050-1066. Springer, Dordrecht.

Tegmark, M. (2017). *Life 3.0. Being Human in the Age of Artificial Intelligence*. Knopf, New York.

Titcomb, J. (2017). Security robot 'drowns itself' in office fountain. *The Telegraph* online, July 18, 2017. https://www.telegraph.co.uk/technology/2017/07/18/security-robot-drowns-office-fountain/ Accessed March 17, 2018.

van de Poel, I., and Royakkers, L. (2011). *Ethics, Technology, and Engineering: An Introduction*. Wiley-Blackwell, UK.

Verbeek, P. P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology and Human Values*, 31(3):361-380.

Verbeek, P. P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In Vermaas P. E. et al., editors, *Philosophy and Design*, pages 91-103. Springer, Dordrecht.

Weizenbaum, J. (1976). *Computer Power and Human Reason. From Judgment to Calculation*. New York / San Francisco, W. H. Freeman and Company.

Wiener, N. (1950). *The Human Use of Human Beings* : *Cybernetics and Society*. Cambridge, The Riverside Press.

Wolf, M. J., Miller, K., and Grodzinsky, F. S. (2017). Why we should have seen that coming. Comments on Microsoft's Tay "experiment," and wider implications. *ACM Computers & Society*, 47(3):54-64.

---

[i] For the full text of the ACM Code of Ethics and Professional Conduct, see https://www.acm.org/about-acm/acm-code-of-ethics-and-professional-conduct Accessed March 18, 2018.

[ii] See e.g. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation Accessed April 2, 2018.