Navigating Through Changes of a Digital World

Nathalie Hauk and Manfred Hauswirth

Abstract In this chapter, we address the question of how trust in technological development can be increased. The use of information technologies can potentially enable humanity, social justice, and the democratic process. At the same time, there are concerns that the deployment of certain technologies, e.g., AI technologies, can have unintended consequences or can even be used for malicious purposes. In this chapter, we discuss these conflicting positions.

Information technologies have become an integral part of work, health, entertainment, communication, and education. Yet, the great hope of this technological (r)evolution of opening up a world of possibilities—unlimited access to information, free expression for all, clean energy, sustainability, economic growth, and industrial innovations—has turned into a fear of living under a surveillance state with “transparent” citizens. Increasingly, society is divided into those who consider themselves as progressive, willing to jump on the bandwagon of technological innovation, and those for whom things are moving too fast, who feel powerless in defending their rights and safety. This debate within society has been called the “midlife crisis of the technological revolution” (Ars Electronica 2019), referring ironically to the search for orientation of people in their 40s. Consequently, this is the right moment to ask ourselves the fundamental questions of “why we develop technologies” and “what purpose they serve.” Historically, humans have always used tools to overcome their own (physical) limitations to ensure survival. Today, digital technologies carry the potential to not only overcome physical limitations but to promote and enable humanity, social justice, and the democratic process. For that, it is crucial to address

---

N. Hauk (✉)

Fraunhofer Institute for Open Communication Systems (FOKUS), Berlin, Germany

e-mail: nathalie.hauk@fokus.fraunhofer.de

M. Hauswirth

Technical University of Berlin, Berlin, Germany

Weizenbaum Institute, Berlin, Germany

e-mail: manfred.hauswirth@tu-berlin.de

the issues of morality, ethics, and legality in the development of technologies since the ultimate limit of technologies must be the ethical and moral limits. For more details on the topic, see chapter “Our Digital Mirror” in “Ethics and philosophy of technology.”

A prime example of the increasing reliance on technology in the modern society is Artificial Intelligence (AI). AI algorithms and technologies have already found their way into everyday life. Hence, the question of whether AI technologies should be employed no longer arises. The performance of routine tasks, such as using web search engines, opening a smartphone with face ID, or running automatic spell checks when writing an email, relies on AI, often unnoticed by the user. Nevertheless, as with any new technology, the use of AI brings both opportunities and risks. While AI can help with protecting citizens’ security, improving health care, promoting safer and cleaner transport systems, and enabling the execution of fundamental rights, there are also justified concerns that AI technologies can have unintended consequences or can even be used for malicious purposes.

Good exemplary demonstrations of these fundamental problems can be found in the area of Machine Learning (ML). ML is used to discover patterns in data, e.g., identifying objects in images for medical diagnoses. The big advantage is clear: An ML algorithm never gets tired and performs the tedious analysis task for enormous numbers of images, at high frequencies and speeds. With the invention of quantum computers, even larger amounts of data could/will be analyzed in real time, tackling problems that are out of reach until now. However, ML algorithms are often “black boxes”—capable of performing a learned or trained behavior without offering insight into how or why a decision is made (for a brief overview on explainable AI, see Xu et al. 2019). For the ML training process, the appropriate selection of training data sets is of crucial importance. The deployment of inappropriate or biased data sets often only becomes apparent after a training process has already been completed, as the following three examples illustrate:

1. Automated decision-making processes are deployed increasingly in recruiting and human resources management. In 2018, Amazon had to cease an AI recruiting tool after discovering that the underlying algorithms of their software discriminated against women. Presumably so because the initial training data set contained more male applicants than female applicants. Hence, the algorithm learned that the best job candidate was more likely to be a male.
2. Tay, a chatbot developed to research conversational understanding, released on Twitter by Microsoft in 2016, started using abusive language after receiving vast quantities of racist and sexist tweets from which Tay learned how to conduct a conversation.
3. An image recognition feature, developed by Google in 2015, miscategorized two black people as gorillas. The fact that the company failed to solve the problem (but rather blocked the categories “gorilla,” “chimp,” “chimpanzee,” and “monkey” in the image recognition entirely) demonstrates the extent to which ML technologies are still maturing.

A product must ensure the same standard of safety and respect for fundamental rights, regardless of whether its underlying decision-making processes are human-based or machine-based. Moreover, we create AI systems that are able to write texts and can communicate in natural language with us. Some of them do this so eloquently that we are no longer able to distinguish whether a real person communicates with us or a system. This fundamental problem has been shown by Joseph Weizenbaum with his simple natural language processing system ELIZA (Weizenbaum 1967). Since then, such systems, e.g., virtual assistants, have evolved significantly, while the problem still remains unsolved. The potential danger is that we do not know whether a given piece of information comes from a human or a machine. Thus, we cannot infer the reliability of a given information, or we may have to re-define the concept of reliability altogether. These issues become even more delicate and pressing when fundamental rights of citizens are directly affected, for example, by AI applications for law enforcement and jurisdiction. Traceability of how an AI-based decision is taken and, therefore, whether the relevant rules are respected is of utter most importance.

1 Trust as a Key Driver

Fears of surveillance and malicious use of technology potentially decelerate or even prevent technological and societal progress. So the fundamental question is:

How can we increase trust in technological development in order to generate value from the application of technologies?

Trust is a key antecedent of ensuring technology acceptance (Siau and Wang 2018) and, thus, a key requirement for continuing the progress of technological development. This is particularly important, when dealing with technologies that are not directly controlled by humans or if they make autonomous decisions. The importance of trustworthy AI has been identified and emphasized also at the political level, e.g., by the European Commission (European Commission 2020). It is important to distinguish between trustworthiness of a technology, e.g., AI, and trust in technologies. While trustworthy AI is comprised of normative ideas on the qualities and characteristics of a technology (that may or may not depend upon ethical considerations), trust in technologies is based on psychological processes through which trust is developed (Toreini et al. 2020). Yet, the concepts of trustworthiness of AI and trust in AI are intertwined. The concept of trustworthy AI is based on the idea that trust builds the foundation for sustainable technology development and that the full benefits of AI deployment can only be realized if trust can be established. At the same time, addressing the ethical considerations in the process of technology development or deployment influences the formation of trust, such as confidence that systems are designed to be beneficial, safe, and reliable.

Definitions of trust can have different emphases depending on the type of trust relationship, e.g., trust toward individuals, toward organizations, or toward machines

(Bannister and Connolly 2011). The comparability of the concept of trust in an interpersonal relationship and trust in machines is subject to ongoing scientific debate. Either way, the general concept of trust always includes a perception of risk, e.g., any type of negative consequence that might derive from using a technology. Trust can be defined as “the willingness [...] to be vulnerable to the actions of another party [...] irrespective of the ability to monitor or control that other party” (Mayer et al. 1995, p. 712). The formation of trust in technology specifically depends on the interplay of three characteristics: (1) human characteristics, such as personality and abilities; (2) environmental characteristics, such as morals and values of a given institution or culture; and (3) technology characteristics, such as the performance of the technology, its attributes, and its purpose (Schäfer et al. 2016; Siau and Wang 2018). Generally, the influence of human characteristics and environmental characteristics will be similar regardless of the type of trust relationship. Therefore, we continue with the specificities of technology characteristics in the human-technology trust relationship.

Human interaction with technology is increasingly moving away from the simple use of computers as tools toward building relationships with intelligent, autonomous entities that carry out actions independently (deVisser et al. 2018). As technological devices become ever-more sophisticated and personalized, the way humans bond with technology, e.g., by touching and talking to machines, intensifies as well. People have the tendency to anthropomorphize technology and, in the case of AI, to apply human morals to it (Ryan 2020). However, to apply human moral standards to machines is problematic since not even very complex machines like AI technologies possess consciousness, intentions, or attitudes at the moment (and possibly never will). Nevertheless, research suggests that the formation of trust in technologies depends on the level of perceived “humanness” of a technology—the perception of human-like traits, e.g., voice and animation features (Lankton et al. 2015). Furthermore, people develop trust in technologies in different ways, e.g., along more human-like criteria or more system-like criteria. According to the ABI+ model of trust (Mayer et al. 1995; Dietz and Den Hartog 2006), characteristics that enhance perceived trustworthiness in a person are ability, benevolence, integrity, and predictability. *Ability* refers to the skills and competencies that enable the trustee to have influence or deliver a desired outcome. Corresponding system-like criteria would be technical robustness and safety of a technology. *Benevolence* refers to the belief in the goodwill of the trustee. Applied to technologies, the perceived level of benevolence can be increased by the responsible deployment of a technology and its transparency. *Integrity* refers to a set of principles of the trustee that is perceived as respectable. To increase the level of perceived integrity in technologies, we must strengthen their reliability and accountability. Finally, *predictability* refers to the stability of perceived trustworthiness sustained over time.

For example, people are more likely to trust a new technology when that technology is provided by an institution with a high reputation—representing ability, benevolence, integrity, and predictability—in contrast to an institution without such a reputation (Siau and Wang 2018). Trust in chatbots, for example, depends partially on the perceived security and privacy of the service provider (Følstad et al. 2018).

Furthermore, if technologies are perceived as reliable, transparent, and secure, the trust in a technology increases (Hancock et al. 2020).

2 Conclusions

Society relies increasingly on technologies to stay competitive and to meet the growing complexities of life in a globalized world. AI algorithms and technologies have already made their way into everyday life, leading to improvement of human health, safety, and productivity. However, we need to balance these benefits with a careful deliberation of the unwanted side effects or even abuse of AI technologies. In compliance with ethical and moral principles, we need to ensure that AI systems benefit individuals and that AI's economic and social benefits are shared across society in a democratic and equal fashion.

If society approaches technological development primarily with fear and distrust, the technological progress will slow, and important steps toward ensuring the safety and reliability of AI technologies will be hindered. If society approaches technological innovation open-mindedly, technologies may have the potential to profoundly change society for the better. The basic building block to achieve this is trust. In order to foster and restore trust in technological advancement, we need to minimize risks, make systems verifiable, and build effective and accountable legislation, along with developing a new understanding of what trust in technologies can mean. The underlying psychological mechanisms of trust in human-technology relationships may extend the traditional trust concepts or reinvent the meaning of trust entirely. The process for this development has started already, but will require more practical experiences, experiments, and analyses in an open, discursive form with broad inclusion of societal stakeholders.

References

- Ars Electronica (2019) Out of the Box – the Midlife Crisis of the Digital Revolution. *Ars Electronica Festival*. Linz, Austria, September 5th-9th 2019.
- Bannister, F., & Connolly, R. (2011) Trust and transformational government: A proposed framework for research. *Government Information Quarterly*, 28, pp. 137-147. <https://doi.org/10.1016/j.giq.2010.06.010>
- De Visser, E.J., Pak, R., & Shaw, T.H. (2018) From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction. *Ergonomics*, 61(10), pp. 1409-1427. <https://doi.org/10.1080/00140139.2018.1457725>
- Dietz, G., & Den Hartog, N.D. (2006) Measuring trust inside organisations. *Personnel Review*, 35 (5), pp. 557-588. <https://doi.org/10.1108/00483480610682299>
- European Commission (2020) White Paper on Artificial Intelligence: a European approach to excellence and trust. Brussels, February 19th 2020. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed: March 3rd 2021)

- Følstad A., Nordheim C.B., Bjørkli C.A. (2018) What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In: Bodrunova S. (eds) *Internet Science*. INSCI 2018. Lecture Notes in Computer Science, 11193, pp. 194-208. Springer, Cham. https://doi.org/10.1007/978-3-030-01437-7_16
- Hancock, P.A., Kessler, T.T., Kaplan, A.D., Brill, J.C., & Szalma, J.L. (2020) Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors*, pp. 18720820922080-18720820922080. <https://doi.org/10.1177/0018720820922080>
- Lankton, N.K., McKnight, D.H., & Tripp, J. (2015) Technology, Humanness, and Trust: Rethinking Trust in Technology. *Journal of the Association for Information Systems*, 16(10), pp. 880-918. <https://doi.org/10.17705/1jais.00411>
- Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995) An Integrative Model Of Organizational Trust. *Academy of Management Review*, 20(3), pp. 709-734. <https://doi.org/10.5465/amr.1995.9508080335>
- Ryan, M. (2020) In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5), pp. 2749-2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Schäfer, K.E., Chen, J.Y., Szalma, J.L., & Hancock, P.A. (2016) A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, 58(3), pp. 377-400. <https://doi.org/10.1177/0018720816634228>
- Siau, K., & Wang, W. (2018) Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal*, 31(2), pp. 47-53.
- Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., & van Moorsel, A. (2020) The relationship between trust in AI and trustworthy machine learning technologies. In *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, January 27th –30th 2020, Barcelona, Spain. ACM, New York, NY, USA, pp. 272-283. <https://doi.org/10.1145/3351095.3372834>
- Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019) Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In: Tang, J., Kan, M.Y., Zhao, D., Li, S., & Zan, H. (eds) *Natural Language Processing and Chinese Computing*. NLPCC 2019. *Lecture Notes in Computer Science*, (11839), pp. 563-574. Springer, Cham. https://doi.org/10.1007/978-3-030-32236-6_51
- Weizenbaum, J. (1967). Contextual Understanding by Computers. *Communications of the ACM*, 10 (8), pp. 474-480. <https://doi.org/10.1145/363534.363545>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.