Building Human Systems of Trust in an Accelerating Digital and AI-Driven World

Yoshija Walter

1 Department for Business, Institute for Management and Digitalization, Kalaidos University of Applied Sciences, Zurich, Switzerland, 2 Laboratory for Cognitive Neuroscience, Faculty of Mathematics and Natural Sciences, University of Fribourg, Fribourg, Switzerland, 3 Translational Research Center, University Hospital for Psychiatry, University of Bern, Bern, Switzerland

Keywords: trust, humans, computer, digital, digitalization, artificial intelligence, digital ethics, digital humanities

INTRODUCTION

We have become accustomed to navigating ourselves not only in the physical but also in the digital world. Both people in modern societies as well as AI-systems “learning” online make use of publicly available information online known as open source intelligence, or, OSINT (Glassman and Kang, 2012; Chauhan and Panda, 2015; Weir, 2016; Quick and Choo, 2018; González-Granadillo et al., 2021; Sebyan Black and Fennelly, 2021). One of the main challenges in this domain is that it has become difficult to discern fact from fabricated materials—sometimes even deliberately exploited through “fake news” and “disinformation campaigns” (Sood and Enbody, 2014; Martinez Monterrubio et al., 2021; Petratos, 2021; Beauvais, 2022; giahanou et al., 2022; Lin et al., 2022; Rai et al., 2022). Already with the standard algorithms employed today, we are continuously facing three looming problems:

- **Algorithmic manipulation**: how do I know that I am presented online with the full truth and that the algorithms don’t just show me a one-sided selection of information?
- **Deliberate disinformation campaigns**: how do I know that the information I see comes from an honest source and has not been produced by a party that deliberately tries to spread false information?
- **Veracity of the medium**: how can I know that the information (i.e., the message, report, picture, audio, or video) depicts real world facts and has not been fabricated by a cunning AI program?

As such, the question of how to deal with AI in respect to ethical norms and matters of trust is becoming a focal discussion point (Reynolds, 2017; Aoki, 2020; Chi et al., 2021; Shin, 2021; Lewis and Marsh, 2022). The following pages briefly outline two case reports, highlight some of the associated problems and propose how they could be addressed in the future through social endeavors. In principle, the epistemic standpoint of the present paper is neither political nor economic in nature. Rather it focuses on the problem that an AI’s instrumental goals are not automatically congruent with the terminal objectives of humans, especially in the domain of informational control, and can even be exploited deliberately by people with unethical intentions. Previous publications have taken stock of the theoretical, normative and social contributions made in the past years dealing with such issues of trust (Hohenstein and Jung, 2020; Tomsett et al., 2020; Godoy et al., 2021; Kerasidou, 2021; Sengupta and Chandrashekar, 2021) and have applied them to the problem of AI information processing (Kim et al., 2020; Mattioli et al., 2022; Wei et al., 2022; Zerilli et al., 2022). The present discussion builds upon them by acknowledging how increasingly fast the digital and AI developments are becoming —so under the consideration of two exemplary case reports, it aims to show which issues are still unresolved in terms of the inherent opacity in...
information transparency, and in which direction a societal discussion about these difficulties in AI alignment and safety engineering could go to potentially mitigate these problems.

**CASE REPORTS**

**Cambridge Analytica**

In 2014, the British data analysis company *Cambridge Analytica* was founded and shortly after has provoked a considerable scandal because they offered personality tests on Facebook, whereby the company not only collected data from the participants but also from their friends. This way, in a short amount of time they were able to collect around 50 million data sets of Facebook accounts for which they have invested around one million dollars. These data sets were the basis for manipulating the US elections, among others (Kaiser, 2019). In 2014, Cambridge Analytica was said to have been involved in 44 US-presidential candidates. The company boldly claimed that they were able to push Ted Cruz from being a “no name” to Donald Trump’s most notable contestant (Vogel, 2015). Using the psychometric data from millions of people, the goal was to deliberately target the voters with their fears and weaknesses in an automated fashion and to skew the outcome of the elections. Cambridge Analytica did not survive the scandal and declared insolvency in 2018. However, it seems like they are continuing their business model under a new company called *Emerdata* (Mijnssen, 2018; Murdock, 2018).

**OpenAI and Dall-E**

Artificial intelligence, or, machine learning, is a vibrant field of research that improved considerably in the past few years. Just recently, new models made headlines for opening new possibilities. The goal is to use artificial neural networks to find novel solutions to mathematical problems by letting the computer “learn” (in a figurative sense) from a large data set. Ever since the creation of GPT-3, a technology developed by OpenAI (a company that was co-founded by Elon Musk), the NLP capabilities have increased to another level (Zhang and Li, 2021). The application called *Dall-E 2* provides an interface between GPT-3 and computer vision, which allows users to put in commands in plain English and then creates images that can rarely be distinguished from real photos or human artwork (Ramesh et al., 2021, 2022). This yields new possibilities to further elude the boundaries between fact and fiction in the digital world for a mainstream audience. For ethical considerations, OpenAI has become hesitant to share this technology with the public (OpenAI, 2018; Schneider, 2022).

There are three common criticisms when building huge LLM (large language models) in machine learning: (i) it requires considerable computing power, which is environmentally demanding (Bender et al., 2021); (ii) the model “learns” from the biases on the internet and thereby becomes more prone, for example, to connect Islam with terrorism and to further discriminate minorities (O’Sullivan and Dickerson, 2020); and (iii) when a machine learns how to imitate human text processing, our academic and public institutions cannot check anymore if certain content is plagiarized or not (Rogerson and McCarthy, 2017; Mindzak and Eaton, 2021).

A fourth and less frequently discussed problem is the main objective of the present paper. It is linked to the previous criticisms, but it needs to be distinctly highlighted: it is the question of knowing how to trust the resulted output. This means either knowing if an information is veridical or knowing that one in fact is dealing with material stemming from an actual human if one believes this to be the case.

**ACKNOWLEDGING THE PROBLEMS**

These brief case reports illustrate that there are some challenges in digital, automated, and self-governed AI systems. The main ones are the following:

- **Reality-monitoring:** In our everyday physical interactions, it is often not difficult to verify a certain statement and “see it for oneself”. And if the context is more complex, one can ask an expert in the field. In the digital world, however, assertions can rarely be easily checked and it is also questionable if a comment indeed comes from a respected expert or if it is only fabricated by a third party.

- **Tailored information delivery:** In the digital world, information curation is often selected according to the trails we leave behind. In the case of Cambridge Analytica, this was done deliberately to manipulate voters, but in the general case of YouTube or Instagram, it is a generally accepted business model that they suggest material for us according to our previous online behavior. In a sense, there is no other choice because of the enormous amount of data online. Nevertheless, this poses the problem that one inevitably gets siloed into specific social and informational contexts—and often, users are not consciously aware of this fact.

- **Transparency:** For the most part, information on the web comes across as abstract and even anonymous information. There is no real way in which users can easily make sure to understand how certain information delivery is created and by which means it was delivered to us, let alone to know for sure who has created the data.

All of this creates a significant problem of trust in an increasingly digital and AI-driven world due to informational opacity (Lewis and Marsh, 2022; Zerilli et al., 2022). Thus, there is an increasing call for human control in the automated systems to warrant that everything is in order (Aoki, 2020, 2021).

**DEVELOPING SYSTEMS OF TRUST**

Based on the above comments, there are several recommendations that may be valuable for constructing human systems of trust in the digital world.

**Social Initiatives**

- **Leveraging the common problems:** We have already discussed the three problems of reality monitoring, tailored information and transparency concerning the current automation...
tendencies. The initiatives set up need to take these problems seriously and offer practical solutions to them.

- **Self-criticism:** Digital institutions and platforms (from search engines to news portals and chat bots) have to become highly self-critical and must be perceived as exceptionally honest. This means that they correct false information as soon as it is spotted and inform the users about their mistakes. Only if consumers establish a solid trust in the institution’s integrity can they also trust the data and information they distribute. Failing to be portrayed as responsibly self-critical would—or should—result in reputational consequences, letting them be perceived as unsuitable for reliable information delivery.

- **Institutions and networks:** Information should not be monopolized and should be networks and a market of institutions that are responsible for data curation. For example in the business world, rating agencies tell investors if their money is well spent with certain companies and it is crucial that there is no monopoly on this task so that they can criticize each other if one agency might be biased. This is important because these ratings have global consequences, and the same would be true for information processing on the internet.

- **Open data:** Data curators should include full transparency on how given information was created, who can warrant for its accuracy and how it is being distributed. The same is true for AI-systems, which usually learn from open source data banks and then fabricate a new answer or solution based on these inputs. These systems, too, need to tell us how the solutions came about and where the newly created information can be fact-checked (in other words: where it “learned” these things and how it can be verified). So far, conventional AI’s do not provide these kinds of information as they rather appear as black boxes, even to the ones who programmed them. To the public, this problem is exacerbated by the fact that the AI algorithms themselves are proprietary material and not generally available. However, the community of IT-developers has loudly demanded that there must be more transparency in this respect, which has had some effect. There are two noteworthy examples illustrating this: first, the Tesla founder Elon Musk has announced that he wanted to buy Twitter with the clear intent to enable free speech on the internet and to make its source code available (although whether the deal will be followed through remains yet to be seen; da Silva, 2022). Second, Facebook’s Meta has recently published their open pre-trained transformer language model (a new LLM) called OPT-175B with the distinct novelty that the details are now completely open to the public (Zhang et al., 2022).

- **Normative values and diversity of perspectives:** Currently, AI systems “learn” from the web as if it is a normative reference. Hence, the are prone to generate racist or sexist outputs. Computer scientists are working on integrating some pre-programmed normative values known as “process for adapting language models to society”, or, in short, PALMS (Solaiman and Dennison, 2021), but at the moment the systems are not very nuanced. It may be argued that the better such models can appropriate the real world, the more they might be able to handle complex social difficulties (such as dealing with racism or sexism). Problems of this sort can be mitigated by introducing a diversity of perspectives, so that the AI does not curate only the most likely output (after all, AI’s are based on statistical models) but provide us with a set of different perspectives that can be found (Johnson and Iziev, 2022).

### Digital Solutions

- **Building digital cultures and spaces of trust:** The social initiatives have to be embedded in organizational and digital environments. Hence, the information curators should consider themselves as not only curators of data but as curators of trustworthy content. This means that the reputation of projecting a culture of honesty and integrity is bound to become one of the most fundamental assets in the online world.

- **Brands and certificates:** Since a company wants to attract customers, it always acts as a brand. The more they can be identified with, the better they can attract customers. If an agency, for example, is perceived as a good rater for social justice and ethics, they can afford to hand out ratings and certificates. Like this, Max Havelaar has become one of the leading social justice stamps and if they approve a product, customers are usually confident that is unproblematic. The same can be the case for online brands that might hand out ratings and certificates for trustworthy data online. These certifications may even be embedded in up-to-date technology, such as blockchains and NFT’s (Adel et al., 2022). It is, of course, not easy to decide upon which agencies and certificates should count as “the” trustworthy ones. There are many questions associated with such an idea, like “Who can hand out such certificates?”, “Who decides which ones are good?”, or “What happens if such a brand misuses its position?”. This is where the word brand may be useful: just like in the areas of science and economy, people often worry about their reputation because if they get called out in a negative light, it has adversarial consequences to them. People would lose interest in what they have to offer and thus it could be imagined that there might be some market dynamics governing such certificates.

There are certain caveats that should be taken seriously when working on such endeavors: (i) online platforms should mitigate the risks for social and informational silos, which is currently a huge problem with the algorithms and AI-systems at play; (ii) information curators should form networks that hold each other accountable for malpractice; (iii) and as a society we need to work a large-scale digital literacy that would be enriched with strong critical thinking capabilities so that people know what they are dealing with online and can judge the content with due care. Promoting digital literacy is already one of the UN’s Sustainable Development Goals (SDGs) but there should be an added focus on developing critical thinking skills that can be applied to the interpretation of information in the digital world (UNESCO, 2018).
CONCLUSION

There is an ongoing technological revolution that comes along under the headings of digitalization and digital transformation. Human systems of trust are crucial to help us discern which outputs could be trusted and which ones may be questionable. They have to make sure that the systems are not used to create informational and social silos that eventually may become irreconcilable. We should work on a digital culture that entails proficiency in digital literacy, and one of its main interests should be the focus on large-scale critical thinking. Information curators have to make it their main priority that they are not hackable and that they act as brands in which the population can place its trust. Managing these challenges responsibly lies at the heart of a healthy development of our societies and personal wellbeing.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.
Reynolds, M. (2017). Peering inside an AI’s brain will help us trust it. *New Sci.*, 235, 10. doi: 10.1016/S0262-4079(17)31298-8

Rogerson, A. M., and McCarthy, G. (2017). Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *Int. J. Educ. Integr.*, 13, 1–15. doi: 10.1007/s40979-016-0013-y

Schneider, J. (2022). OpenAI’s New Tech Lets You Generate Any ‘Photo’ By Just Describing It. *PetaPixel*. Available online at: https://petapixel.com/2022/04/06/openais-new-tech-lets-you-generate-any-photo-by-just-describing-it/

Sebyan Black, I., and Fennelly, L. J. (2021). “Chapter 20—Investigations using open source intelligence (OSINT),” in: *Investigations and the Art of the Interview (Fourth Edition)*, I. Sebyan Black and L. J. Fennelly Eds. (Butterworth-Heinemann), 179–189

Sengupta, P. P., and Chandrashekhar, Y. S. (2021). Building trust in AI: opportunities and challenges for cardiac imaging. *JACC: Cardiovasc. Imag.* 14, 520–522. doi: 10.1016/j.jcmg.2021.01.002

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Human Comput. Stud.* 146, 102551. doi: 10.1016/j.ijhcs.2020.102551

Solaaiman, I., and Dennison, C. (2021). *Improving Language Model Behavior by Training on a Curated Dataset* [Research paper]. San Francisco, CA: OpenAI. Available online at: https://openai.com/blog/improving-language-model-behavior/

Sood, A. K., and Enbody, R. (2014). "Chapter 2—intelligence gathering." In *Targeted Cyber Attacks*, A. K. Sood and R. Enbody, eds. (Oxford, United Kingdom: Syngress) 11–21

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., et al. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, 1, 100049. doi: 10.1016/j.patter.2020.100049

UNESCO (2018). *Meet the SDG 4 Data: Indicator 4.4.1 on Skills for a Digital World* [UN Blog], Institute for Statistics. Available online on: http://uis.unesco.org/en/blog/meet-sdg-4-data-indicator-4-4-1-skills-digital-world

Vogel, K. P. (2015). Cruz partners with donor’s "psychographic" firm [News portal]. Virginia, US: POLITICO. Available online at: https://www.politico.com/story/2015/07/cruz-donor-for-data-119813

Wei, Y., Lu, W., Cheng, Q., Jiang, T., and Liu, S. (2022). How humans obtain information from AI: Categorizing user messages in human-AI collaborative conversations. *Inf. Process. Manage.* 59, 102838. doi: 10.1016/j.ipm.2021.10.2838

Weir, G. R. S. (2016). "Chapter 9—the limitations of automating OSINT: understanding the question, not the answer,” in: *Automating Open Source Intelligence*, R. Layton and P. A. Watters, eds. Oxford, United Kingdom: Syngress, (159–169)

Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*. 3, 1–10. doi: 10.1016/j.patter.2022.10.0455

Zhang, M., and Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.*, 1, 831–833. doi: 10.1016/j.fmre.2021.1.011

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., et al. (2022). OPT: Open Pre-trained Transformer Language Models (arXiv:2205.01068.), arXiv.

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

**Copyright © 2022 Walter. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.**