frontiers

# Building human systems of trust in an accelerating digital and AI-driven world

Yoshija Walter[1*]

[1]Université de Fribourg, Switzerland

*Submitted to Journal:*
Frontiers in Human Dynamics

*Specialty Section:*
Digital Impacts

*Article type:*
Opinion Article

*Manuscript ID:*
926281

*Received on:*
22 Apr 2022

*Journal website link:*
www.frontiersin.org

## Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

## Author contribution statement

YW is the sole author of this opinion or general commentary paper.


## Keywords

Trust, Humans, computer, Digital, digitalization, machine learning, AI, artificial intelligence, Digital ethics, Digital Humanities


## Contribution to the field

We are living in an increasingly digital world and the pace of AI development has accelerated. Only in the last few weeks, OpenAI (a company co-founded by Elon Musk) has provided access to its newest machine learning “playground” (as they call it) to the general public, based on Davinci-002, which is a model using GPT-3, the perhaps most potent large language model (LLM) to date. Its codex interface is not only good for generating regular text but also as a text-to-code algorithm that enables laypersons to create an app on demand with plain English – or any of the other 26 languages used in the system. Only a few days ago, OpenAI has now also issued its newest GPT-3 connection to computer vision called Dall-E 2 that is able to use English commands to create fabricated but photorealistic images that can barely be distinguished from real photos. This has serious consequences for how our societies have to deal with online information. There has been a call for human systems of trust. The present opinion paper highlights two brief case reports, pinpoints the social challenges and makes some recommendations in which direction the discussion should go.




## Funding statement

There is no external funding.

Opinion

# Building human systems of trust in an accelerating digital and AI-driven world

Yoshija Walter [1, 2, 3]

[1] Institute for Management and Digitalization, Department for Business, Kalaidos University of Applied Sciences

[2] Laboratory for Cognitive Neuroscience, Faculty of Mathematics and Natural Sciences, University of Fribourg

[3] Translational Research Center, University Hospital for Psychiatry, University of Bern

yoshija.walter@kalaidos-fh.ch

**Abstract**

The world is experiencing a strong trend towards digitalization, and machine learning is expected to play a major role in processing and distributing digital information. To both have a controlled development of these digital technologies and to make sure that society can manage to handle them responsibly, there is an increasing call for establishing human systems of trust. These systems should be set up to help us navigate in the digital sphere. There are three problems that can be seen today and with the advent of AI, they are likely to get more difficult to handle: (i) algorithmic manipulation, (ii) deliberate disinformation, and (iii) questions about the veracity of the media. These problems are discussed in light of two short case reports: (a) Cambridge Analytica, and (b) OpenAI with their new application called Dall-E. The present paper discusses these problems and makes the proposal that there is the need for creating novel and decisively human systems of trust by implementing social initiatives and digital solutions. There are three caveats that should be kept in mind when working on these systems, which is avoiding social silos, creating interdependent accountability and fostering a culture where digital literacy and critical thought are key.



**Keywords**

Trust, humans, computer, digital, digitalization, machine learning, AI, artificial intelligence, digital ethics, digital humanities

## 1. 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 (Chauhan & Panda, 2015; Glassman & Kang, 2012; González-Granadillo et al., 2021; Quick & Choo, 2018; Sebyan Black & Fennelly, 2021; Weir, 2016). 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" (Beauvais, 2022; Giachanou et al., 2022; Lin et al., 2022; Martinez Monterrubio et al., 2021; Petratos, 2021; Rai et al., 2022; Sood & Enbody, 2014). 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 (Aoki, 2020; Chi et al., 2021; Lewis & Marsh, 2022; Reynolds, 2017; Shin, 2021). 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.

## 2. 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 & 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 barely 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 & 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 (Mindzak & Eaton, 2021; Rogerson & McCarthy, 2017).

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.

**3. 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 (Lewis & 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).

**4. Novel systems of trust**

Based on the above comments, there are several recommendations that may be valuable for constructing novel 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.
- *Institutions and networks:* Information should not be monopolized and there 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.

- *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 & Dennison, 2021), but at the moment the systems are not very nuanced. 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 & 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 aways 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).

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 with strong critical thinking capabilities so that people know what they are dealing with online and can judge the content with due care.



**5. 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.

**References**


Adel, K., Elhakeem, A., & Marzouk, M. (2022). Decentralizing construction AI applications using blockchain technology. *Expert Systems with Applications*, *194*, 116548. https://doi.org/10.1016/j.eswa.2022.116548

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly*, *37*(4), 101490. https://doi.org/10.1016/j.giq.2020.101490

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior*, *114*, 106572. https://doi.org/10.1016/j.chb.2020.106572

Beauvais, C. (2022). Fake news: Why do we believe it? *Joint Bone Spine*, 105371. https://doi.org/10.1016/j.jbspin.2022.105371

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Chauhan, S., & Panda, N. K. (2015). Chapter 6—OSINT Tools and Techniques. In S. Chauhan & N. K. Panda (Eds.), *Hacking Web Intelligence* (pp. 101–131). Syngress. https://doi.org/10.1016/B978-0-12-801867-5.00006-9

Chi, O. H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, *118*, 106700. https://doi.org/10.1016/j.chb.2021.106700

Giachanou, A., Ghanem, B., Ríssola, E. A., Rosso, P., Crestani, F., & Oberski, D. (2022). The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers. *Data & Knowledge Engineering*, *138*, 101960. https://doi.org/10.1016/j.datak.2021.101960

Glassman, M., & Kang, M. J. (2012). Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). *Computers in Human Behavior*, *28*(2), 673–682. https://doi.org/10.1016/j.chb.2011.11.014

González-Granadillo, G., Faiella, M., Medeiros, I., Azevedo, R., & González-Zarzosa, S. (2021). ETIP: An Enriched Threat Intelligence Platform for improving OSINT correlation, analysis, visualization and sharing capabilities. *Journal of Information Security and Applications*, *58*, 102715. https://doi.org/10.1016/j.jisa.2020.102715

Johnson, S., & Iziev, N. (2022, April 15). A.I. Is Mastering Language. Should We Trust What It Says? *The New York Times*. https://www.nytimes.com/2022/04/15/magazine/ai-language.html

Kaiser, B. (2019). *Targeted: My Inside Story of Cambridge Analytica and How Trump, Brexit and Facebook Broke Democracy*. HarperCollins Publishers Ltd.

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33–49. https://doi.org/10.1016/j.cogsys.2021.11.001

Lin, T.-H., Chang, M.-C., Chang, C.-C., & Chou, Y.-H. (2022). Government-sponsored disinformation and the severity of respiratory infection epidemics including COVID-19: A global analysis, 2001–2020. *Social Science & Medicine*, *296*, 114744. https://doi.org/10.1016/j.socscimed.2022.114744

Martinez Monterrubio, S. M., Noain-Sánchez, A., Verdú Pérez, E., & González Crespo, R. (2021). Coronavirus fake news detection via MedOSINT check in health care official bulletins with CBR explanation: The way to find the real information source through OSINT, the verifier tool for official journals. *Information Sciences*, *574*, 210–237. https://doi.org/10.1016/j.ins.2021.05.074

Mijnssen, I. (2018, May 3). Cambridge Analytica: Nachfolger Emerdata gegründet. *Neue Zürcher Zeitung*. https://www.nzz.ch/international/cambridge-analytica-nachfolger-emerdata-gegruendet-ld.1382705

Mindzak, M., & Eaton, S. E. (2021, November 4). *Artificial intelligence is getting better at writing, and universities should worry about plagiarism* [Opinion Article]. The Conversation. http://theconversation.com/artificial-intelligence-is-getting-better-at-writing-and-universities-should-worry-about-plagiarism-160481

Murdock, J. (2018, March 5). What Is Emerdata? As Cambridge Analytica Shuts, Directors Surface in New Firm. *Newsweek*. https://www.newsweek.com/what-emerdata-scl-group-executives-flee-new-firm-and-its-registered-office-909334

OpenAI. (2018, April 9). *OpenAI Charter*. Official Company Website. https://openai.com/charter/

O'Sullivan, L., & Dickerson, J. (2020, August 7). *Here are a few ways GPT-3 can go wrong* [Opinion Article]. TechCrunch. https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/

Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. *Business Horizons*, *64*(6), 763–774. https://doi.org/10.1016/j.bushor.2021.07.012

Quick, D., & Choo, K.-K. R. (2018). Digital forensic intelligence: Data subsets and Open Source Intelligence (DFINT+OSINT): A timely and cohesive mix. *Future Generation Computer Systems*, *78*, 558–567. https://doi.org/10.1016/j.future.2016.12.032

Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, *3*, 98–105. https://doi.org/10.1016/j.ijcce.2022.03.003

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv:2204.06125 [Cs]*. http://arxiv.org/abs/2204.06125

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *ArXiv:2102.12092 [Cs]*. http://arxiv.org/abs/2102.12092

Reynolds, M. (2017). Peering inside an AI's brain will help us trust it. *New Scientist*, *235*(3133), 10. https://doi.org/10.1016/S0262-4079(17)31298-8

Rogerson, A. M., & McCarthy, G. (2017). Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *International Journal for Educational Integrity*, *13*(1), 1–15. https://doi.org/10.1007/s40979-016-0013-y

Schneider, J. (2022, April 6). *OpenAI's New Tech Lets You Generate Any 'Photo' By Just Describing It*. PetaPixel. https://petapixel.com/2022/04/06/openais-new-tech-lets-you-generate-any-photo-by-just-describing-it/

Sebyan Black, I., & Fennelly, L. J. (2021). Chapter 20—Investigations using open source intelligence (OSINT). In I. Sebyan Black & L. J. Fennelly (Eds.), *Investigations and the Art of the Interview (Fourth Edition)* (pp. 179–189). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-822192-1.00021-0

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Solaiman, I., & Dennison, C. (2021, June 10). *Improving Language Model Behavior by Training on a Curated Dataset* [Research paper]. OpenAI. https://openai.com/blog/improving-language-model-behavior/

Sood, A. K., & Enbody, R. (2014). Chapter 2—Intelligence Gathering. In A. K. Sood & R. Enbody (Eds.), *Targeted Cyber Attacks* (pp. 11–21). Syngress. https://doi.org/10.1016/B978-0-12-800604-7.00002-4

Vogel, K. P. (2015, July 7). *Cruz partners with donor's "psychographic" firm* [News portal]. POLITICO. https://www.politico.com/story/2015/07/ted-cruz-donor-for-data-119813

Weir, G. R. S. (2016). Chapter 9 - The Limitations of Automating OSINT: Understanding the Question, Not the Answer. In R. Layton & P. A. Watters (Eds.), *Automating Open Source Intelligence* (pp. 159–169). Syngress. https://doi.org/10.1016/B978-0-12-802916-9.00009-9

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4), 1–10. https://doi.org/10.1016/j.patter.2022.100455

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011