designs

MDPI

Correction

Correction: Sharpening the Scythe of Technological Change: Socio-Technical Challenges of Autonomous and Adaptive Cyber-Physical Systems

Daniela Cancila [1,*], Jean-Louis Gerstenmayer [2,†], Huascar Espinoza [1] and Roberto Passerone [3]

1 CEA, LIST, CEA Saclay, PC172, 91191 Gif-sur-Yvette, France; huascar.espinoza@cea.fr
2 MEFi-DGE French Ministry of Economy, 94201 Ivry-sur-Seine, France; jean-louis.gerstenmayer@cea.fr or jean-louis.gerstenmayer@finances.gouv.fr
3 Dipartimento di Ingegneria e Scienza dell'Informazione, University of Trento, 38123 Trento, Italy; roberto.passerone@unitn.it
* Correspondence: daniela.cancila@cea.fr; Tel.: +33-(0)1-6908-0107
† Researcher CEA at present in position of project manager and policy advisor at the French Ministry of Economy. Ideas and opinion in this paper are these of the author, they are not representative of the French Ministry of Economy opinions.

Received: 30 January 2019; Accepted: 31 January 2019; Published: 11 February 2019

We, the authors, wish to make the following corrections to our paper [1].

Section 3.4 must be completely replaced by the following text, as a wrong version of the manuscript was used:

3.4. New Forms of Interaction between Humans and Autonomous Systems

There are unique challenges in moving from human-machine interaction in automation, where machines are essentially used as tools, to the complex interactions between humans and autonomous systems or agents. As autonomous systems progressively substitute cognitive human tasks, some kind of issues become more critical, such as the loss of situation awareness, or the overconfidence in highly-automated machines. The Tesla accident occurred in 2016 is a clear illustration of the loss of situation awareness in semi-autonomous vehicles, as stated by the National Transportation Safety Board (NTSB): "the cause of the crash was overreliance on automation, lack of engagement by the driver, and inattention to the roadway" [2]. One of the main causes of this problem is a reduced level of cognitive engagement when the human becomes a passive processor rather than an active processor of information [3].

In [4], the author addresses the "Ironies of Automation", especially for the control process in industries. The main paradox is that any automated process needs (human) supervisors. In the case of an accident or a problem in industry, the human supervisor may not be sufficiently prepared or reactive to solve it, because automation hides and directly manages ex-manual operations. Therefore, the supervisor could be less prepared in autonomous industrial control systems. Of course, this situation could appear in industry and not in traditional critical systems. In a nuclear plant, for example, engineers in the control rooms receive an expensive and strong (theoretical and experimental via simulation) training, required after the TMI-2 accident [5,6].

While the potential loss of situation awareness is particularly relevant in autonomous systems needing successful intervention of humans, there is a number of more general situations where risks in human-machine interaction must be better understood and mitigated:

- Collaborative missions that need unambiguous communication (including timescale for action) to manage self-initiative to start or transfer tasks.

- Safety-critical situations in which earning and maintaining trust is essential at operational phases (situations that cannot be validated in advance). If humans determine the system might be incapable of performing a dangerous job, they would take control of the system.
- Cooperative human-machine decision tasks where understanding machine decisions are crucial to validate autonomous actions. This kind of scenario implies providing autonomous agents with transparent and explainable cognitive capabilities.

3.4.1. Towards Trusted and Safe Human-Machine Relationships

Several authors [7–10] argue that interactions with autonomous agents must be considered as "human relationships", as we are delegating cognitive tasks to these entities. This perspective opens the door to the application of existing fundamental knowledge from the social sciences (psychology, cognitive modelling, neuropsychology, among others), to develop trustable and safe interactions with autonomous systems. For instance, the authors of [7] propose to encode the human ability of building and repairing trust into the algorithms of autonomous systems. They consider trust repair a form of social resilience where an intelligent agent recognises its own faults and establishes a regulation act or corrective action to avoid dangerous situations. Unlike other approaches where machine errors remain unacknowledged, this approach builds on creating stronger collaboration relationships between humans and machines to rapidly adjust any potential unintended situation.

In 2017, some influential leaders in AI established the Asilomar AI Principles aimed at ensuring that AI remains beneficial to humans [11]. One of these principles is value alignment, which demands that autonomous systems "should be designed so that their goals and behaviours can be assured to align with human values throughout their operation". In this context, some researchers such as Sarma et al. [9] argue that autonomous agents should infer human values by emulating our social behaviour, instead of embedding these values into their algorithms. This approach would also apply to the way human interact and it would be the basis to create learning mechanisms emphasizing trust and safe behaviours, including abilities such as intentionality, joint attention, and behaviour regulation. While these ideas look more appealing with the rise of AI and machine learning, the concept of learning "safe" behaviours in intelligent agents was already explored in 2002 by Bryson et al. [10]. Nevertheless, the problem of finding meaningful safety mechanisms for human-machine interaction inspired from human social skills remains largely open, because of the complex and intricate nature of human behaviour and the need of a provably-safe framework to understand and deploy artificial relationships.

References

1. Cancila, D.; Gerstenmayer, J.-L.; Espinoza, H.; Passerone, R. Sharpening the scythe of technological change: Socio-technical challenges of autonomous and adaptive cyber-physical systems. *Designs* **2018**, *2*, 52, doi:10.3390/designs2040052.
2. National Transportation Safety Board. *Collision between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailor Truck near Williston, Florida, 7 May 2016*; Accident Report NTSB/HAR-17/02-PB2017-102600; National Transportation Safety Board: Washington, DC, USA, 2017; p. 42.
3. Endsley, M.R. Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected. In *Advances in Intelligent Systems and Computing, Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Florence, Italy, 26–30 August 2018*; Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y., Eds.; Springer: Cham, Switzerland, 2018; Volume 824.
4. Bainbridge, L. Ironies of Automation. *Sci. Direct* **1983**, *19*, 775–779.
5. Clément, B.; Jacquemain, D. *Nuclear Power Reactor Core Melt Accidents*; Chapter Lessons Learned from the Three Mile Island and Chernobyl Accidents and from the Phebus FP Research Programme—Chapter 7; IRSN: Paris, France, 2015.
6. Walker, S. *Three Mile Island a Nuclear Crisis in Historical Perspective*; University of California Press: Berkeley, CA, USA, 2006.
7. de Visser, E.J.; Pak, R.; Shaw, T.H. From automation to autonomy: The importance of trust repair in human-machine interaction. *J. Ergon.* **2018**. [CrossRef] [PubMed]

8. Kohn, S.C.; Quinn, D.; Pak, R.; de Visser, E.J.; Shaw, T.H. *Trust Repair Strategies with Self-Driving Vehicles: An Exploratory Study*; SAGE Publications: Thousand Oaks, CA, USA, 2018; Volume 62, pp. 1108–1112.
9. Sarma, G.P.; Hay, N.J.; Safron, A. AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values. In *Computer Safety, Reliability, and Security*; Gallina, B., Skavhaug, A., Schoitsch, E., Bitsch, F., Eds.; Springer: Berlin, Germany, 2018; pp. 507–512.
10. Bryson, J.J.; Hauser, M.D. What Monkeys See and Don't Do: Agent Models of Safe Learning in Primates. In Proceedings of the AAAI Symposium on Safe Learning Agents, Palo Alto, CA, USA, 25–27 March 2002.
11. The Future of Life Institute. Asimolar AI Principles. Available online: https://futureoflife.org/ai-principles/ (accessed on 1 November 2018).