

This document is downloaded from the VTT's Research Information Portal https://cris.vtt.fi

VTT Technical Research Centre of Finland

Safety Challenges of AI in Autonomous Systems: A Human Factors Perspective

Karvonen, Hannu; Heikkilä, Eetu; Wahlström, Mikael

Published: 26/11/2019

Document Version
Publisher's final version

Link to publication

Please cite the original version:

Karvonen, H., Heikkilä, E., & Wahlström, M. (2019). Safety Challenges of Al in Autonomous Systems: A Human Factors Perspective. Poster session presented at Al Day 2019, Espoo, Finland.



VTT http://www.vtt.fi P.O. box 1000FI-02044 VTT Finland By using VTT's Research Information Portal you are bound by the following Terms & Conditions.

I have read and I understand the following statement:

This document is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of this document is not permitted, except duplication for research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered for sale.



Safety Challenges of Al in Autonomous Systems: A Human Factors Perspective

Hannu Karvonen, Eetu Heikkilä, Mikael Wahlström

VTT Technical Research Centre of Finland Ltd

Al in autonomous systems

Autonomous systems (AS) are among the most potential application areas for AI technologies. AI, especially machine learning, is currently used in AS in limited areas such as object detection, but more advanced decision-making and adaptation in operations are also targeted.

Al safety challenges

Al-enabled AS hold big promises for increases in productivity and safety. However, the application of AI also introduces new safety and security risks. In the literature, several concrete safety challenges of AI applications and AS have been identified. For an overview, a list with numbers 1–5 in the left side of Figure 1 is provided. The safety challenges addressed here are adopted from the widely-cited work of Amodei et al. (2016). These issues are relevant for all AI systems, but become especially crucial with systems where AI is embedded as a part of a physical AS that interacts with humans.

Avoiding negative side effects

How to ensure that the system does not disturb its environment in a negative way while pursuing its goal?

Reward hacking

How to ensure that there are no loopholes in the system that allow unintended rewards?

Scalable oversight

How to ensure that the system performs reasonably with limited information?

Safe exploration

How to ensure that exploratory moves don't lead to bad repercussions?

Robustness to distributional shift

How to ensure that the system doesn't make bad decisions in a setting that is different than the one that was used for training?

Human factors views on Al safety

While the technical assurance of AI systems is crucial to create viable autonomous systems, also a wider systemic view on safety is needed. One of the key issues is how the users of autonomous systems interact with AI, and how the safety of these interactions is ensured. To answer the AI challenges presented above, some key human factors considerations are presented in the right side of Figure 1.

- Perform comprehensive task analyses (e.g., core-task or cognitive work analysis) to ensure that the Al's goals are not contradicting with the goals of other entities in the same environment
- Provide sufficient information of the decision-making rationale for the user and other persons as needed, and a possibility to safely intervene with the system when undesired behaviour is noticed
- Enforce AI transparency for the user to be able to see the rationale behind the decisions by the AI system
- Utilize simplification and visualisations to enhance this transparency and its understandability
- Require approval by a human user before rewarding the system, e.g., after new types of actions
- Allow the monitoring of the system by the user(s) from a supervisory perspective
- Create clear definitions for how the system should be supervised and define the scenarios where the human user should be asked to intervene and take over
- Ensure that the people who interact with the system have sufficient knowledge of the system capabilities and its limitations, and also about the extent the system is allowed to explore
- Simulate exploration scenarios and design means for human supervision
- Develop means for human-AI interaction and supervision to support the adaptation to new environment
- Ensure that people interacting with the system are aware of its limitations also when the environment changes

Figure 1. Al safety challenges and proposals for approaches to address them from the human factors perspective

Conclusions

- Autonomous systems are a major application area for AI technologies
- Although the level of autonomy increases, the systems will still be in interaction with human users
- Several concrete Al safety issues have been raised in literature
 - Technical assurance of AI systems is important, but also a broader systemic view is needed
 - Human factors is one part of this consideration
- The key safety issues of AI should be addressed from a human factors point of view to ensure the safety of users and other
 people interacting with autonomous systems that employ AI