

**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



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

Hannu Karvonen, Eetu Heikkilä, Mikael Wahlström

VTT Technical Research Centre of Finland Ltd

## AI 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.

## AI safety challenges

AI-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.

## Human factors views on AI 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.

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

- Perform comprehensive task analyses (e.g., core-task or cognitive work analysis) to ensure that the AI'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

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

- 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

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

- 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

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

- 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

**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?

- 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. AI 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 AI 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



**Contact**
**Hannu Karvonen**
Tel. +358 40 021 6396
hannu.karvonen@vtt.fi