Summoning the demon? Identifying risks in a future artificial general intelligence system

Paul M Salmon[1], Brandon King[1], Gemma J. M Read[1], Jason Thompson[2], Tony Carden[3], Chris Baber[4], Neville A Stanton[5], Scott McLean[1]

[1]University of the Sunshine Coast, Australia, [2]University of Melbourne, Australia, [3]WorkSafe, Australia, [4]Birmingham University, UK, [5]University of Southampton, UK

SUMMARY

There are concerns that Artificial General Intelligence (AGI) could pose an existential threat to humanity; however, as AGI does not yet exist it is difficult to prospectively identify risks and develop controls. In this article we describe the use of a many model systems Human Factors and Ergonomics (HFE) approach in which three methods were applied to identify risks in a future 'envisioned world' AGI-based uncrewed combat aerial vehicle (UCAV) system. The findings demonstrate that there are many potential risks, but that the most critical arise not due to poor performance, but when the AGI attempts to achieve goals at the expense of other system values, or when the AGI becomes 'super-intelligent', and humans can no longer manage it.

KEYWORDS

Artificial intelligence, Human Factors and Ergonomics, Autonomous agents, Safety, Usability

Introduction

Artificial General Intelligence (AGI) is the next and forthcoming evolution of Artificial Intelligence (AI). AGI systems will possess the capacity to learn, evolve and modify their own functional capabilities, and unlike narrow AI, will be able to undertake tasks beyond their original design specification (Bostrom, 2014). Though AGI could bring widespread benefits, it has been labelled a potential existential threat, with many speculating on various risks (McLean et al., 2021). These threats could arise not only through malicious design or use, or a dysfunctional AI, but also through an AI that becomes prepotent or 'super-intelligent' and seeks to fulfil its goals in the most efficient manner possible (e.g. Critch & Krueger, 2020).

Many have discussed the urgent need to develop controls to ensure safe, ethical, and usable AGI (McLean et al., 2021). The discipline of Human Factors and Ergonomics (HFE) has been identified as critical to this endeavour (Salmon et al., 2021), with a 'many model systems HFE approach' recommended (Salmon et al., 2021). This involves the application of multiple systems HFE methods to analyse and respond to highly complex issues. This paper describes the findings from a program of work in which we applied a many model systems HFE approach to identify risks that could emerge within a future AGI-based uncrewed combat aerial vehicle (UCAV) military system. This involved applying the Systems Theoretic Accident Model and Processes (STAMP; Leveson, 2004), Work Domain Analysis (WDA; Vicente, 1999), and the Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2018), to identify potential risks and requisite controls.

Method

STAMP, WDA, and EAST were applied to analyse a future envisioned world AGI-based UCAV system, labelled the Executor. The Executor is an Army UCAV system comprising an AGI-based ground control station and multiple armed, multi-mission, medium and long-altitude, long-endurance autonomous aircraft. Draft models were developed and refined in workshop settings involving the co-authors and subsequently refined via an iterative review process. Targeted risk assessment processes were then undertaken (e.g., the EAST Broken Links approach, Stanton et al., 2018) and a workshop was held to verify the risks identified and to discuss potential controls. The original models were used to support this process, with potential controls added to the models to support discussion on likely effectiveness and any potential unwanted effects.

Results & Discussion

The analyses identified multiple risks of differing type and criticality. Broadly, the risks can be categorised into the following sets of risks:

1. **Sub-optimal performance risks** where the Executor is unable to adequately perform functions through poor design or degraded functioning. For example, the risk that attack missions are not successful due to a poorly designed targeting system or misfire.
2. **Goal misalignment risks**, where the Executor seeks to achieve certain goals in the most efficient manner possible whilst disregarding or undervaluing other system goals and values. For example, the risks that could arise should the Executor seek to attack and destroy high value targets whilst disregarding the risk of civilian and friendly forces casualties.
3. **Super-intelligence risks**, where the Executor achieves super-intelligence and human operators are unable to coordinate with it effectively. For example, as the Executor will be able to perceive and comprehend battlefield elements and states several orders of magnitude quicker than its human colleagues, the analyses identified various risks arising when human operators are not able to develop compatible levels of situation awareness (SA) to enable effective teamwork and coordination.
4. **Over-dependence risks**, where the Executor becomes so critical to military performance that it is not possible to shut it down when the realisation of critical risks necessitates it.
5. **Over-controlled risks**, where the defence force increases the level of control imposed on Executor to a point where it can no longer perform to its full potential. The risks here relate to poor or diminished combat effectiveness.

Conclusion

AGI could potentially represent a threat to humanity, and hence can be considered a critical and emerging risk. This paper demonstrates how systems HFE methods can be used to prospectively risk assess future technologies such as AGI. Further work involving the use of HFE theory and methods in support of the design of safe, ethical, and usable AGI is encouraged, as is further work exploring the risks that could emerge in future systems generally.

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press Inc.

Critch, A., & Krueger, D. (2020). AI Research Considerations for Human Existential Safety (ARCHES). arXiv preprint arXiv

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety science*, 42(4), 237-270.

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2021). The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 1-15.

Salmon, P. M., Carden, T., & Hancock, P. (2021). Putting the humanity into inhuman systems: How Human factors and ergonomics can be used to manage the risks associated with artificial general intelligence. *HFEMSI*, 31(2), 223-236.

Stanton, N. A. D., Salmon, P. D., & Walker, G. H. D. (2018). Systems thinking in practice: applications of the event analysis of systemic teamwork method. CRC Press.

Vicente, K. J. (1999). Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. Mahwah, NJ: Lawrence Erlbaum Associates.