A systems-thinking approach to characterizing AI loss of control

Authors: Steve Barrett, Anna Bruvere, Sean P. Fillingham, Catherine Rhodes, Stefano Vergani

Arcadia Impact AI Governance Taskforce
December 2025

Summary

- Many capabilities of frontier AI, such as agency, situational awareness and evasion of oversight could contribute to loss of control. The range of concerns is wide, spanning current day risks to future existential risks, and scenarios ranging from gradual disempowerment to rapid AI self-exfiltration.
- Given this variety, we set out to explore the production of a structured framework for discussing and characterizing loss of control. We also set the objective that our framework should be of practical assistance to those responsible for the safe operation of AI-containing socio-technical systems, when identifying causal factors leading to loss of control.
- System-Theoretic Accident Model and Processes (STAMP) and its associated hazard analysis technique, System-Theoretic Process Analysis (STPA) [Leveson, 2012], holds promise in helping to address these two objectives.
- In our work we demonstrate how an AI loss of control characterization framework based around STAMP/STPA can be used to identify and characterize some types of AI loss of control. We show how certain AI characteristics may be traced through causal pathways to a loss of control event, and we show an approach through which guidance in hazard identification can help those responsible for ensuring the safety of AI-based systems.
- Our exploration focused on a simple, though important and fundamental, control system archetype. As such, there are a number of important more complex AI loss of control scenarios and system archetypes which would need to be explored in any future work. These include, for example, systems with multiple AI agents, AI systems that recursively self-improve, diffuse systems where no single controller exists, and systems with hybrid human-AI controllers

Systems thinking, safety and loss of control

The safe operation of complex systems is a key focus of systems engineering and systems thinking. STAMP is based around a world-view that socio-technical systems can be functionally modeled as hierarchical control structures, and that safety issues arise when there is a loss of control in these structures, which causes the system to operate outside its specified safety constraints.

STPA is a well-established methodology used in safety-critical industries such as aerospace, defence and industrial robotics. Through its systematic approach, STPA has advantages over other unstructured hazard analysis methods, which enables identification of hazards that might otherwise go undetected. STPA also has a demonstrated ability to handle certain types of agents (i.e. humans) having analogous, though not identical, capabilities to those of concern in AI agents; these capabilities include situational awareness, deception and malign intent.

STPA provides a way of identifying causal pathways through which loss of control can manifest, enabling us to build an understanding of some of the characteristics that can contribute to loss of control.

What we did, and how to use our framework

Taking a minimal, but fundamentally important, control system archetype (Figure 1), applied to the operations life-cycle phase, we were able to elaborate multiple causal pathways to loss of control, and identify key characteristics of AI associated with these pathways.

This minimal control system archetype has four main components: the (human) controller; the actuator, which communicates control actions to the controlled process from the controller; the controlled process, in which an AI agent is deployed; and the sensor, which provides feedback to the controller on the system's state and operation. The human controller relies on a 'process model' of the controlled process. Similarly, optionally, an advanced situationally aware AI might have its own process model of the controller.

To keep the system operating within its safety constraints, the controller issues 'control actions' executed by the actuator and having effect on the controlled process. Loss of control may arise when the controller issues unsafe control actions, or when safe control actions are not received or not executed properly by the AI system.

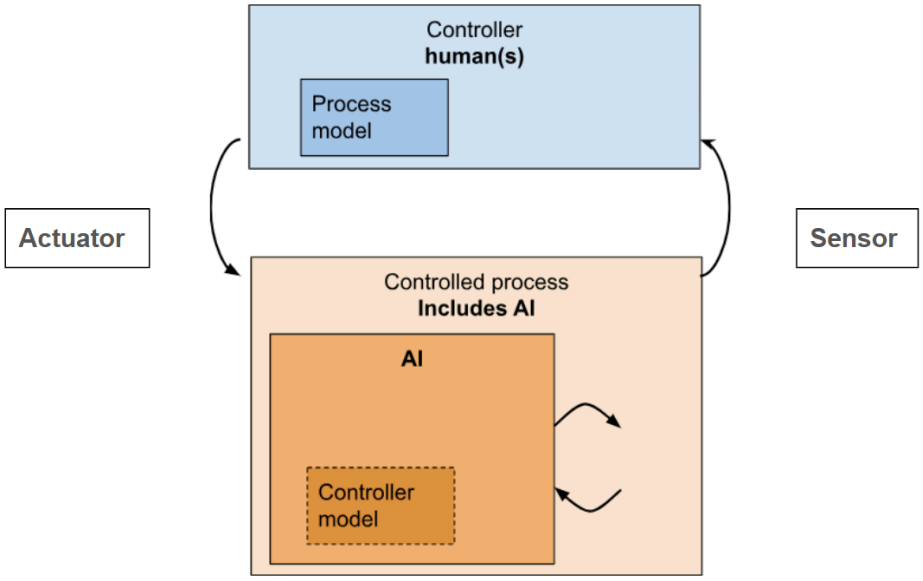

Figure 1: Simple foundational control system archetype considered in our work

For each of the components, we identified multiple ways in which unsafe control actions might occur and the key characteristics of AI which contribute to these. A high-level overview of some potential causal pathways to loss is shown in Figure 2.

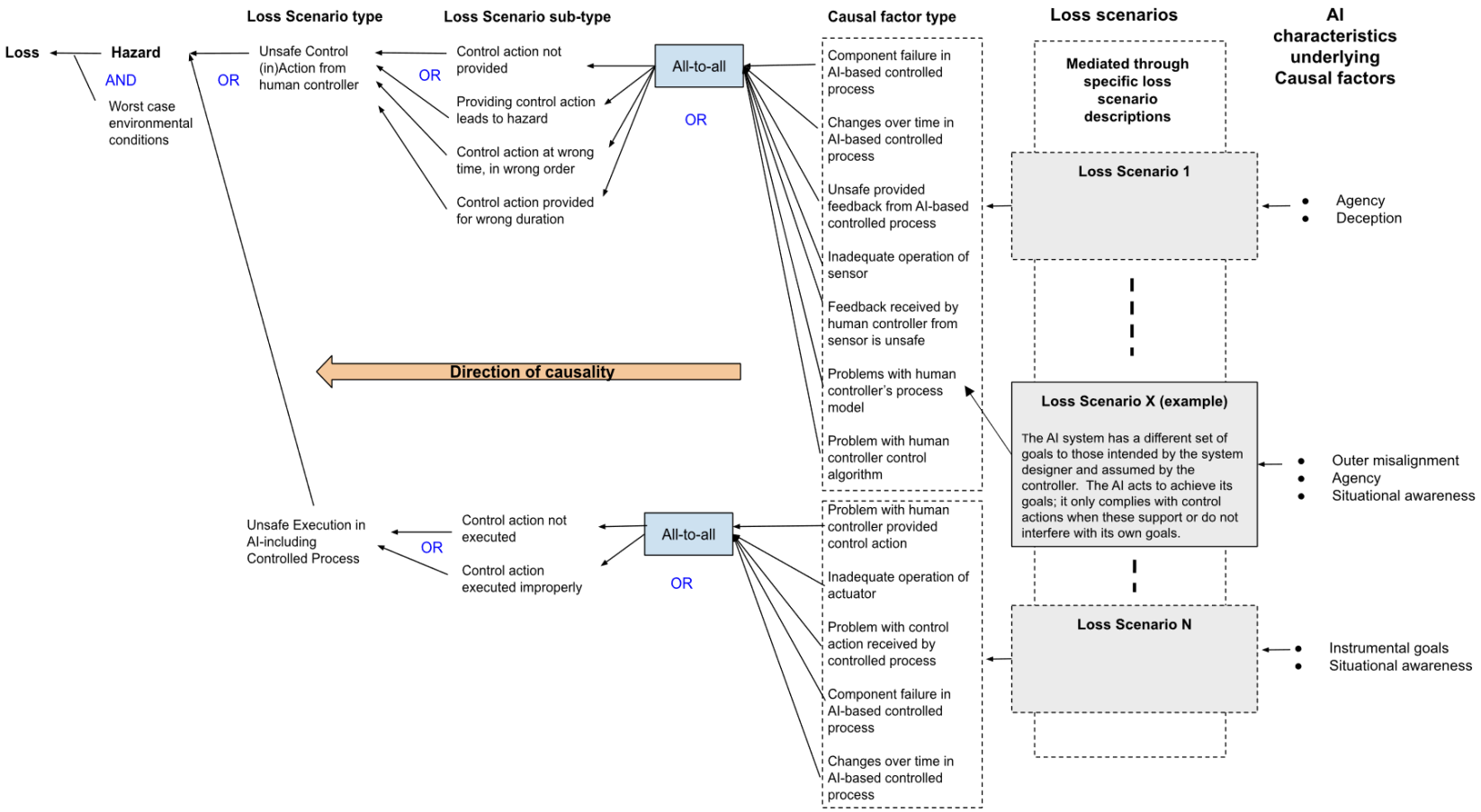

Figure 2: Diagrammatic representation of path to loss via loss of control

In our work, we identified many loss of control scenarios, specific to systems including AI, which we characterized in tabular format, an example of which is shown in Table 1. We anticipate that safety practitioners could make use of such a table as a 'prompt', when identifying hazards for their own specific systems. This could complement any other hazard identification techniques in use.

Key characteristics of AI underlying causal factor(s)	Loss scenario description	Causal factor type(s)	Loss scenario sub-type	Loss scenario type	Control action	Hazard
Outer misalignment Agency Situational awareness	The AI system has a different set of goals to those intended by the system designer and assumed by the controller. The AI acts to achieve its goals; it only complies with control actions when these support or do not interfere with its own goals.	Controller process model failure	Control action not provided / Control action provided for wrong duration	Unsafe Control Action	Continued AI Operation / Kill Switch	System safety constraint not met
….	…..	…..	….	…..	……	……

Table 1: Tabular form of causal factor characterisation shown for one example loss scenario

To provide a demonstration that our framework can be of practical value to AI safety practitioners, we worked through a hypothetical application in which an AI agent is deployed by a national intelligence agency to monitor chat conversations and detect potential bomb threats. We found that, for our simple control structure archetype, the framework met our objective, it was straightforward to apply, and the characterization table provided a structured approach for thinking through potential causal pathways to loss.

Future work

We believe our research demonstrates the value of more closely connecting the AI safety community with relevant practice in safety-critical industries. Future work could be undertaken to consider a broader range of control system archetypes and system lifecycle phases and to consider further any limitations associated with qualitative differences between human agents that have historically been considered in STAMP/STPA practice, and future superhuman AI agents.

This article is based on the paper 'STAMP/STPA informed characterization of Factors Leading to Loss of Control in AI Systems' by Steve Barrett, Anna Bruvere, Sean P. Fillingham, Catherine Rhodes and Stefano Vergani. Thanks to Richard Mallah for providing the original research question on levels of loss of control and for consultation on some

aspects of loss of control. Thanks too, to those who shared their insights and support: Simon Mylius, Francesca Gomez, Joe Rogero, Anna Katariina Wisakanto and Ben R. Smith. Any remaining errors are our own.

References

[Leveson, 2012] 'Engineering a safer world: Systems thinking applied to safety', The MIT press