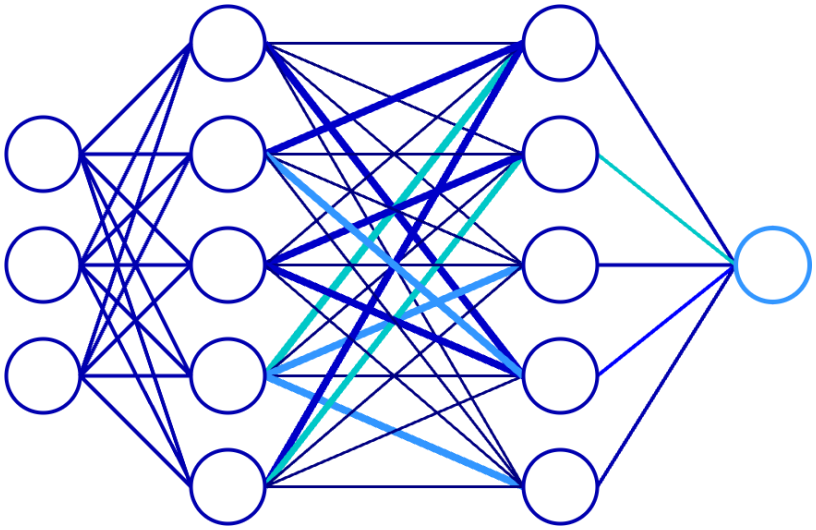

When Enough Is Enough: Assurance Cases for AI System Security Design

Susanna Cox | 26 November 2024 | susanna@anglesofattack.io

Safety, Security & Robustness

How do we know whether our security and robustness requirements for mission-critical AI systems are good enough? To answer this question, it makes sense to consider what makes an AI system's functions critical, and what that means for safety.

Proliferation of Mission-Critical AI

AI applications become mission- or even safety-critical when their failure can contribute to human suffering, injury, or death. An example of safety-critical AI applications is the use of models to enable autonomous vehicles to navigate through the physical world. It is easy to imagine how a critical failure of such a system might result in injury or even death.

The increase of AI-driven applications in everyday life has included an increase in the number and type of AI applications which may be considered safety- or mission-critical. Failure of finance AIML applications may not directly result in loss of life, but could certainly prove to have disruptive effects on larger economic systems. Conversely, failure of an AIML application in healthcare could have either result–economic disruption or loss of human life–or both.

Safety Standards & System Autonomy

To envision where AI safety standards must eventually progress, think of a use case when it would be especially difficult to argue safety standards had been met. One particularly challenging case is one where there is no human in the loop. It's generally easy to see why the more autonomous the system, the more difficult it becomes to argue its safety. Similarly, the more autonomous the system, the more stringent its engineering requirements must be.

It may be useful here to take a step back and reconsider the goals and applications of AI. The goal of most AIML systems is inference at scale. The first aspect of this goal, inference, is often applied as decisions at scale; it is arguable that all inference systems are in fact decision systems. Whether the decision is direct towards a user or subject–such as prison sentencing algorithms or a health application which aids medical professionals in certain diagnoses–or indirect, as in the case of a system which serves recommendations to users based on their in- or outside-of-platform behaviors, is inconsequential. A decision is being made.

Examining the second aspect of this goal, scale, gives a more full picture of AI's chief area of usefulness, as well as what might be its most critical weakness: the shift towards the scaling down of human oversight in the undertaking of some human affairs. In its best light, this is presented as reducing mundane human labor and thus improving overall human condition; however there can be no denying that a central purpose and means to this end is the reduction of human involvement, and thus, human oversight.

In this light, it's clear that AI systems must ultimately require the most stringent of safety arguments. To be practically actionable, these must include a means of assurance for both their sufficiency, and success of execution.

Approach: Assurance Cases

One approach to ensuring sufficient and successful safety controls across diverse applications is to adapt Assurance Cases (ACs), which are often used in challenging safety-critical engineering contexts.

"An AC is defined as a reasoned, auditable created artifact that supports the contention that its top-level claim (or set of claims) is satisfied, including systematic argumentation and the underlying evidence and explicit assumptions that support the claim(s)." [1]

One critical aspect of ACs is the defining of both what needs to be done, as well as the criteria for and/or metrics of successful fulfillment. In this respect, ACs can be thought of as a real-world development application of the four key questions for threat modeling:

- *What are we working on?*
- *What can go wrong?*
- *What are we going to do about it?*
- *Did we do a good enough job?*

*Source: **Threat Modeling Manifesto*** [2]

From this perspective, it makes sense that successfully engineering safety-critical software depends on successful threat modeling, both in theory and in practice.

In their 2021 **paper**, *"Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components"*, Kläs *et al* explore the underlying technical and philosophical questions involved in standardizing an understanding of when secure is secure enough–and also how ACs could be adapted for use in safety-critical AI systems.

A key quote:

"A risk acceptance criterion that seems reasonable to apply in the context of AI…states that the residual risk after the application of safety measures should be As Low As Reasonably Practicable (ALARP). The meaning of 'reasonably practicable' is not static but depends on the state of the technology and the intended application, including the underlying business case and related practical restrictions."

Once again, it becomes clear that successful application of ACs lies in proper understanding of the specific risks applicable to a given system. Critically, in the case of safety-critical AI components and systems, security risks may arise from either traditional/general cybersecurity vectors, or threats may be AI-specific, arising from the operation of any AIML system in production.

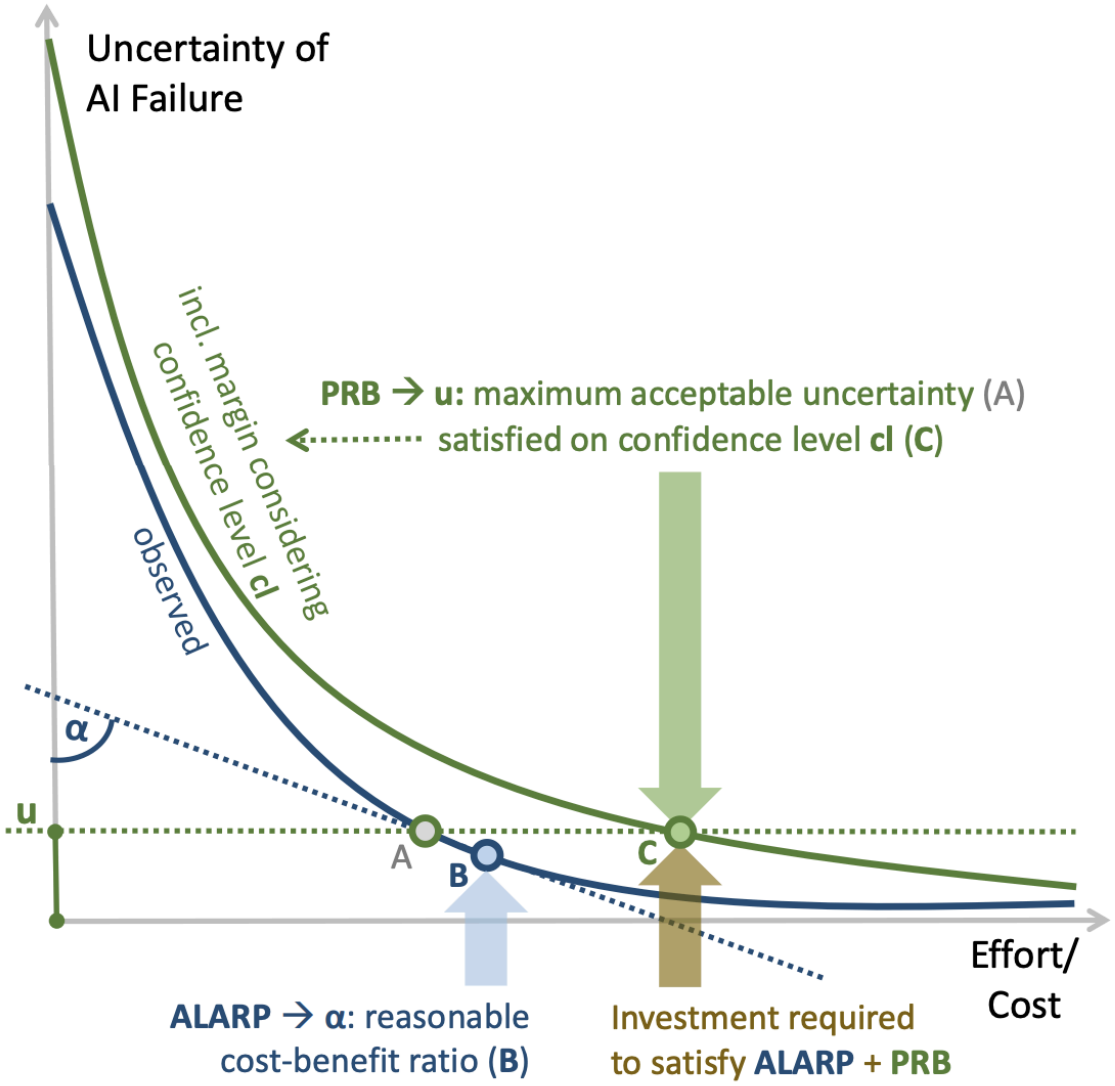

Image 1. [1]

Image 1, taken from the paper, describes some of the relationships among cost, effort, and risk associated with developing AI systems, with the goal of ALARP residual risk. This is an interesting conceptual understanding of the delicate balance among AI stakeholder interests in industry, governance, and legislation.

In any practical application, these considerations must include what is reasonable, practical, and possible within the field–or else run the risk of obsolescence before they have ever been enacted. Modeling threats to AI systems, both cybersecurity and AI-specific, remains a necessary first step to building and deploying them securely.

References

1. Kläs, Michael, Rasmus Adler, Lisa Jöckel, Janek Groß and Jan Reich. “Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components.” AISafety@IJCAI (2021).
2. “Threat Modeling Manifesto.” n.d. Threat Modeling Manifesto. Accessed November 27, 2024. https://www.threatmodelingmanifesto.org/.