AI GUARDRAILS: CONCEPTS, MODELS AND METHODS

John McDermid, Kester Clegg, Yan Jia, Ibrahim Habli

Centre for Assuring Autonomy, University of York, UK

Centre *for* Assuring Autonomy

Abstract

It is now common to talk about "guardrails" as ways of controlling the risks associated with AI models, especially general-purpose AI (GPAI). This paper argues that this metaphor is misleading (over-simplistic) and that a more refined view of risk controls is needed, drawing on insights from safety engineering. This leads to the notions of a hierarchy of controls and layers of protection which can inform definition of a risk management strategy for a given AI application. The paper illustrates how concepts of layers of protection and hierarchies of controls, including precedence orders for controls, can be adapted from traditional safety engineering and used to guide the definition of risk management strategies for GPAI.

Introduction

When people refer to risks from artificial intelligence (AI), especially general-purpose AI (GPAI), they frequently also talk about introducing "guardrails". This is a compelling metaphor – it immediately conjures up images of steel or other strong materials to keep us on a safe path. But it is just a metaphor and, unfortunately, rather a misleading one. This position paper briefly explores the concept of guardrails, identifies some relevant concepts from safety engineering, e.g. layers of protection and hierarchies of risk controls, and outlines an approach to identify appropriate risk control strategies for a given problem.

The Guardrails Concept

Discussions of AI guardrails often include technical, policy and legal aspects. Our focus here is on the technical aspects[1] and the challenges that arise in developing effective guardrails.

First, the space of AI models is vast and very high-dimensional (perhaps billions of parameters). In 3D space physical guardrails plus the way we use them, e.g. holding on to them in crossing a bridge or descending stairs, guide us through a subset of the space. It is obvious from inspection how the guardrails work[2] – and that schemes such as those produced by M C Escher (see Fig. 1) wouldn't work. We have no way of visualising such "structures" in high-dimensional spaces and, even if we could, we would not be able to judge whether the guardrails have been placed effectively to keep the model behaviour "safe" (or the protections are illusory as in the Escher drawing) as there is no effective way of measuring or predicting their strength. For example, some of the common forms of guardrails are filters on the input and output, e.g. to remove offensive terms, or to prevent the model from providing advice on how to make a Molotov cocktail. However, it is often easy to bypass such filters – the term "jailbreaking" is typically used – so that harmful content is still produced.

[1] As they can give practical controls, rather than merely exhortations (policy) or redress (legal).
[2] Also, engineering analysis or standards mean that their strength can be assessed to show that they are good enough to offer the desired protection.

Second, the structure of the AI model is unknown and perhaps unknowable. The term "jagged frontier" [1] is used to mean a "boundary" in the model space which shows the limits of its capabilities. The idea has been used to distinguish those things that an AI model can do and those it can't – by comparison with a human notion of equally difficult tasks, see Figure 2 (taken from [1]). In our case, we are interested in the boundary of what the AI can do safely. If we use an AI model to do a task which happens to lie within the frontier and the AI model does it well, and safely, we may then use the model to do another task which is outside its capability, as we think it should do equally well on such a task but the results are inadequate (a degraded capability) or are even unsafe. Figure 2 again rather reinforces the metaphor – a simple 2D-rendering of the frontier which is smooth rather than jagged – but the challenge that the paper sets out is that we don't know where the frontier is, and therefore where to place guardrails to stop the model being employed on tasks it can't do safely. The challenge is also underscored by the ability to find single-pixel attacks on vision models that lead to misclassification [2]; this shows that the frontiers really are jagged in an important subclass of AI applications.

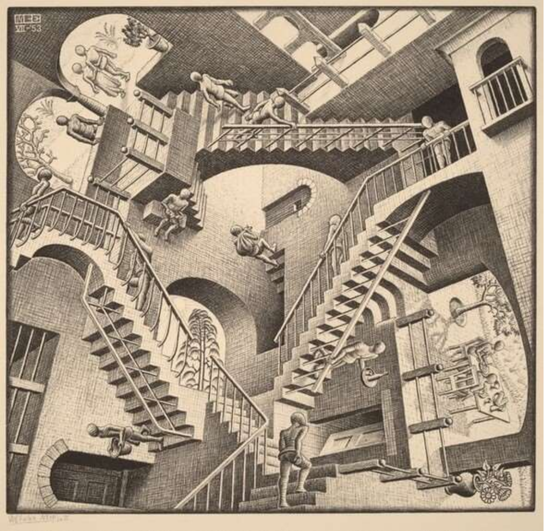

Figure 1: "Relativity", M C Escher

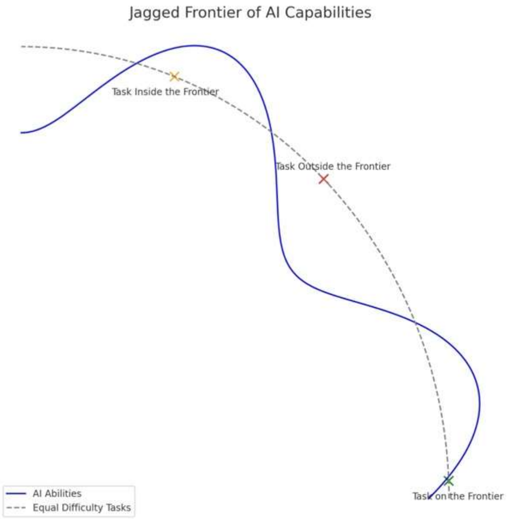

Figure 2: The Jagged Frontier [1]

Third, and perhaps most importantly, we must use guardrails properly – hold on to them, don't leap over them, etc. They do not keep us safe "despite ourselves". The metaphor of AI guardrails is often employed as if they will keep us safe regardless, but they can't, and jailbreaking of models is analogous to leaping over the guardrails. The idea of guardrails also rather implies a "one-size-fits-all" model of how to achieve safety. But it is both more subtle and complex than that.

Learning from Safety Engineering

Safety engineering includes analyses to identify failure modes of systems and their components and introduces controls to address the attendant risks. In defining controls, it is common to introduce layers of protection recognising that no one single control is likely to be completely effective. Further, there is a notion of a hierarchy of controls where those higher in the hierarchy are stronger and therefore preferable to those lower in the hierarchy. Using these concepts assists in defining safety architectures. In applying these concepts to AI models, it is helpful to distinguish between intrinsic and extrinsic properties of the AI models for example, weights in neural networks vs the performance of the network on some input data and this guides where to put risk controls, as illustrated in Figure 3. Further, it is important to understand when risk controls can be developed based on extrinsic properties or visible behaviour of the AI model.

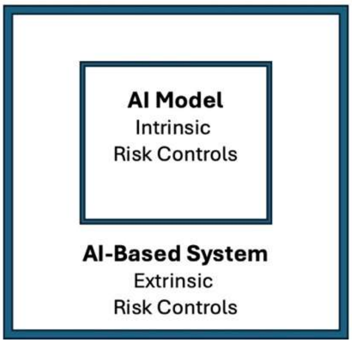

Figure 3: Intrinsic vs Extrinsic Risk Controls

We first give an example of an extrinsic risk control. Consider an unmanned aerial vehicle (UAV) with fixed wings, like a conventional aircraft, using AI for path planning and for flight control, e.g. operating the control surfaces using large language models (LLMs)[3]. Here, risk controls can be based on extrinsic properties. For example, stall depends on a combination of the speed of the UAV and the angle of attack, both of which can be measured and risk controlled by limiting commands to the actuators for the control surfaces. It would make sense to work in terms of a stall margin keeping the UAV flight parameters some way away from stall to allow for measurement uncertainty, local air turbulence, etc. More simply, perhaps, risk controls can be implemented on the planned flight paths, e.g. keeping the UAV out of exclusion zones (again with a margin to allow for inevitable uncertainties).

As well as identifying risk controls, safety engineering is concerned with assuring the system so it can be trusted. In this example the trust (and safety certification) can be focused on the risk controls not on the core control functions[4] and this can be done using a conventional certification process (also see the hierarchy of controls below). From a performance perspective we want some confidence in the flight control or path planning algorithms, as we would not want the safeguards to be routinely triggered (otherwise there is no benefit from the use of AI), but this is an operational/commercial consideration not a safety one.

In contrast, there are circumstances where the risk controls must be implemented using internal data and can be thought of as necessarily intrinsic. This occurs when

[3] For example, see: https://www.goodai.com/llm-agent-taught-to-control-drones/
[4] This philosophy is being adopted by some companies, e.g. SAIF Systems, see: https://saifsystems.ai.

there are no (immediate) external measurements that can be made to inform or direct risk control. An example of this would be recommending treatment for sepsis where the clinical impact of inappropriate recommendations might be serious but by the time they can be detected harm will have been done (of course, extreme erroneous doses, e.g. more than the patient's body weight, can be detected extrinsically). In this case, the risk controls must be in the AI model itself, e.g. shaped by an understanding of the relationship between patient characteristics and appropriate dosages of vasopressor (one of the drugs used to treat sepsis). This has been illustrated, taking a published model developed using reinforcement learning then showing its limitations (e.g. an unsafe rate of change of medication) and improving it through adaptations to the feature space and to the cost function [3]. Given the high criticality of the clinical scenario, the model needs to be adapted to include internal risk controls, even when there are external checks on the model outputs. From an assurance perspective, it is necessary to focus on the AI model itself, but there are emerging approaches for developing trust in the AI models, e.g. AMLAS[5]. But assurance is difficult with GPAI, e.g. LLMs, and the focus is necessarily more on extrinsic approaches.

For example, it is becoming increasingly popular to use retrieval augmented generation (RAG) approaches to focus LLMs on a particular company's data to get specific and relevant responses rather than generic ones reflecting the large quantities of data used to train the LLMs in the first place. However, as LLMs are prone to hallucination, or confabulation, the use of RAG methods also can be viewed as providing some form of risk control. In the authors' experience confabulation can arise from drawing on a general case in a specific situation to which it doesn't apply, i.e. giving the typical or modal answer in an atypical situation. This type of "extrinsic hallucination" is often easy to detect, as it will employ data that are clearly not relevant to the situation. Thus, the use of RAG frameworks for risk control can be thought of as something of a hybrid, but also a partial, approach. Hybrid, as using the local data on which the LLM is focused (extrinsic) helps to align the internal state of the AI model (intrinsic) with its operating context and makes erroneous outputs less likely. It is partial, as using RAG methods narrows the information used in making predictions to the relevant context but doesn't ensure that the outputs correctly reflect this context, i.e. that inferences are sound. For example, we have encountered cases of what we term "intrinsic hallucination" where the output is based on the local data and is perfectly plausible – but not correct [4] – and the checks in current RAG frameworks are insufficient to detect such cases.

Developers of GPAI, e.g. OpenAI, give guidance on ways of developing guardrails, e.g. on input filters (to detect potentially harmful prompts), and on outputs to detect and block inappropriate results, e.g. offensive language[6]. The approach proposed uses the same underlying GPAI model to "guard" the primary AI model – from a safety and

[5] Assurance of Machine Learning for use in Autonomous Systems, see: https://www.assuringautonomy.com/amlas
[6] https://cookbook.openai.com/examples/developing_hallucination_guardrails#2-constructing-our-hallucination-guardrail

assurance perspective there is a clear risk of a common mode failure, i.e. the primary model and the guardrail failing the same way (at the same time and/or on the same input data). Identifying and avoiding potential common mode failures is a well-established safety engineering principle.

Hierarchies of Risk Controls

In safety engineering it is common to identify a hierarchy of controls with the first choice being engineered mitigations that prevent hazards, and human-based mitigation of hazard consequences usually being a last resort. Here we sketch, as a tentative proposal, a hierarchy of risk controls for AI. This hierarchy would be used in choosing individual risk controls and in defining layers of protection for a particular AI-based system. The hierarchy and some illustrative examples of each are set out below:

1. The application allows an extrinsic, engineered control that can be assured using conventional means in the application domain, i.e. conformance with the standards.
    a. The above-mentioned UAV where directly programmed controls can be used for stall protection and preventing excursions into forbidden airspace.
    b. An on-line clinical triage chatbot which can use a lookup table to detect trigger words, e.g. "suicide", "self-harm" to stop the dialogue and connect the user to a human for immediate support.
2. The application allows an extrinsic, engineered control that uses AI, but which can be assured using emerging approaches such as AMLAS.
    a. A sliding-mode controller for a vehicle clutch is tuned and optimised using a radial basis function neural network (RBFNN) [5]; in this case safety and performance are aligned, hence control system optimisation is a guarantor of safety, and the RBFNN is simple (having few parameters), hence facilitating assurance.
3. The application allows an extrinsic, human-based, control where there is sufficient time for the human to make informed decisions and to exercise control[7].
    a. Many clinical diagnosis systems, e.g. used in pathology, use the AI system with the aim of reducing workload on pathologists, for example using AI + one pathologist, rather than having two pathologists read each image [6]; this gives performance/workload benefits but also retains the clinician in the role of decision-maker, with the intent that they can compensate for limitations in the AI models.
4. The application and AI development methods allow an intrinsic control, where the adaptations of the model can be justified using an appropriate analysis (safety, security, etc.) and the resulting model can be assured using an approach such as AMLAS.

[7] See: https://assuringautonomy.medium.com/human-control-of-ai-and-autonomy-the-art-of-the-possible-66779846e1d8 for a more extensive discussion of the circumstances under which human control is feasible, and the different types of control that can be exercised.

    a. Using safety analyses to inform the refinement of the feature space and loss function for a system recommending vasopressors for treatment of sepsis (as outlined above) [3].

5. Neither intrinsic nor extrinsic controls are possible, perhaps due to the choice of the AI model, such as LLMs, and a hybrid approach is the best option for reducing risk.
    a. Use of RAG-based approaches for generating safety analyses but iterating in the production of the analyses to allow an expert human to guide the AI model, e.g. by refining prompts, to produce a credible result [4].

Some observations are in order. First, options 1 and 2 are forms of fault-tolerant architecture, a topic which is well-understood for conventional computer-based systems, but little studied for AI-based systems (the discussion at [7] is a rare foray into this topic). Second, other precepts from reliability and safety engineering need to be applied, e.g. to avoid common mode or single points of failure; again, these ideas need to be adapted for systems employing AI, e.g. considering whether training data sets could give rise to common mode failure. Third, with option 3, care is also needed to avoid the human becoming a "liability sink" [8] and there is benefit in designing systems to support (augment) humans not to supplant them, an approach which is sometimes referred to as human-centred[8]. Fourth, the ranking of options 3 and 4 is debatable given the challenges for humans operating in a monitoring role; in general, analysis of the strengths and weaknesses of the options is needed in the application context, and it may be that option 4 is preferable to 3 in some situations. Thus, developing such analyses would be needed in a refinement of this proposed method.

Further, it should be noted that in two of the five examples (for cases 2 and 3) safety and performance are aligned, meaning that the overall system (purely technical in case 2, and socio-technical in case 3) are both efficient and can be assured for safety. Finally, there is a growing body of work on guardrails for LLMs and it is likely that a more refined approach to selecting guardrail models for LLM-based applications will emerge over time (as well as more guardrails). As a caveat, it should be noted that all these approaches reduce risk – they don't eliminate it – and work is still needed on more effective ways of assessing risk for GPAI.

As indicated above, we need to determine how to build AI-based systems with layers of protection to reduce risks to an acceptable level. This might involve, for example, an intrinsic AI-based control and an extrinsic human-based control. The hierarchy of controls informs the definition of layers – the stronger the mechanisms used the fewer layers are likely to be needed. But it's a little more subtle than that. AI and humans have different failure modes – for example a clinician reading pathology images might miss some artefacts, e.g. a lesion, that the AI can detect reliably, and vice versa. Thus, the definition of the layers needs to be informed by knowledge of the failure modes of

[8] A group in Stanford has pioneered this approach, see: https://hai.stanford.edu, although the term is now in more widespread use.

the humans in the task context and of the technology (AI model) being used. The layers should then provide at least one "barrier" to each failure mode with significant (unsafe) consequences. Attention also needs to be given to avoid common mode failures. In conventional safety architectures there are methods of analysis and information on failure modes that can inform such design. Although there is work published on AI risks, e.g. by NIST including refinements for GPAI[9], this tends to be quite generic and to realise effective layers of protection requires more specific failure mode analysis. At present, this idea remains aspirational except in some well understood contexts, e.g. in pathology.

Conclusions

Much has been written about the need for guardrails for AI, especially for advanced forms such as LLMs, which are perceived as presenting high levels of risk. We have argued that whilst this is a compelling metaphor it is quite misleading (over-simplistic) and that a more nuanced approach is needed to identify risk controls and layers of protection appropriate for a particular AI/ML model, in its context of use. Although the guardrails metaphor might mislead people to believing AI risks are simply controlled, the term has achieved wide enough acceptance that we don't think it will easily be changed. Instead, we advocate messages for different communities:

1. Safety Engineers – continue to promote classical approaches to safety engineering and fault-tolerant architectures but adapt them for AI and refer to them as guardrails when speaking to other communities to align with the accepted terminology.
2. Policy Makers and Regulators – ask for guardrails for model deployments (the effectiveness of the controls can only be properly assessed in the context of use) including why the risk controls were chosen, e.g. what failure modes they address and how the layers of protection manage the overall risk.
3. AI Developers and Deployers – recognise that solutions to AI "problems" are not only within the AI world (e.g. shaping loss functions or using LLMs to guard LLMs) but are to be found in other disciplines; develop more clarity on AI failure modes, seeking to move beyond the notions of "concrete problems in AI" to a more specific understanding of AI failure modes in their socio-technical context [9].

Perhaps most important is that these groups collaborate, with mutual respect.

It is acknowledged that the proposed approach to AI safety needs further exploration and validation/refinement, but we believe it is potentially instructive and that the notions of a hierarchy of controls and layers of protection can give critical insight. Finally, our hierarchy suggests that the use of frontier models, e.g. LLMs, in critical applications would be difficult to justify without further work to strengthen guardrails

[9] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

and an honest evaluation of how strong these guardrails are (what risk reduction they achieve), given the likely failure modes of the AI models and the impact of those failure modes in their context of use. In particular, this applies where LLMs are used to guard models produced by the same LLMs as it is hard to see how to avoid common mode failures making the guardrails illusory – and as useful as Escher's handrails.

Acknowledgements

M.C. Escher's "Relativity" © 2024 The M.C. Escher Company-The Netherlands. All rights reserved. www.mcescher.com

References

[1] Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayer L, Candelon F, Lakhani KR. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper. 2023 Sep 15(24-013).

[2] Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation. 2019 Jan 4;23(5):828-41.

[3] Jia Y, Burden J, Lawton T, Habli I. Safe reinforcement learning for sepsis treatment. In 2020 IEEE International conference on healthcare informatics (ICHI) 2020 Nov 30 (pp. 1-7). IEEE.

[4] Clegg KD, McDermid JA, Habli I. Using GPT-4 to Generate Failure Logic, in Proc SASSUR 2024.

[5] Zou J, Huang H and Gühmann C, "Improved Sliding-Mode On-line Adaptive Position Control for AMT Clutch Systems Based on Neural Networks," *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu, China, 2018, pp. 174-178, doi: 10.1109/IVS.2018.8500691.

[6] Ng AY, Glocker B, Oberije C, Fox G, Sharma N, James JJ, Ambrózay E, Nash J, Karpati E, Kerruish S, Kecskemethy P D, Artificial Intelligence as Supporting Reader in Breast Screening: A Novel Workflow to Preserve Quality and Reduce Workload, *Journal of Breast Imaging*, Volume 5, Issue 3, May/June 2023, Pages 267–276,

[7] Fenn J, Nicholson M, Pai G, Wilkinson M. Architecting safer autonomous aviation systems. arXiv preprint arXiv:2301.08138. 2023. (Also, in the proceedings of the 31st Safety Critical Systems Symposium, 2023, The Future of Safe Systems, M Parsons (Ed), ISBN-979-8363385520.)

[8] Ryan PM, Porter Z, Al-Qaddoumi J, McDermid JA, Habli I. What's my role? Modelling responsibility for AI-based safety-critical systems. arXiv preprint arXiv:2401.09459. 2023 Dec 30.

[9] Raji A D, Dobbe R, Concrete Problems in AI Revisited, https://arxiv.org/abs/2401.10899

Contact

☎ +44 (0)1904 325345

✉ assuring-autonomy@york.ac.uk

in linkedin.com/company/assuring-autonomy

📍 Institute for Safe Autonomy,
University of York, York, YO10 5FT