Executive Summary

Executive Summary

Management of Substrate-Sensitive AI Capabilities (MoSSAIC): Substrate-flexible risk

Project Overview:

This research investigates limitations in current AI safety approaches (particularly mechanistic interpretability) in light of potential changes to AI architectures and the development of highly adaptive or self-modifying systems. We want to identify and clarify the “mechanistic/reductionist” paradigm in interpretability and introduce "substrate-flexible risk" as a novel threat model, one in which a mechanistic understanding of specific neural mechanisms fails as AI capabilities advance.

Research Context & Progress:

- Initial paper submitted to Tokyo AI Safety Conference and pending arXiv publication.
- Developing interest from researchers (e.g., Paul Colognese, Kola Ayonrinde).
- Part of the broader High Actuation Spaces Project, run by Sahil K.

Key Research Objectives:

1. Examine two critical assumptions in mechanistic interpretability:
    - Stability of structural properties as AI capabilities increase
    - Reliability of structural properties for safety assertions
2. Investigate vulnerabilities across multiple scenarios:
    - Changes to model scaffolding (e.g., new tools/capabilities)
    - Shifts in AI architectures/paradigms
    - AI-assisted architectural evolution
    - Self-modifying systems
    - Deep and aggregate deception

Methodology & Deliverables:

- Phase 1: Technical upskilling and MoSSAIC paper refinement through feedback
- Phase 2: Core theoretical development through literature analysis, case studies, and comparative analysis with mature sciences.
- Phase 3: Meta-level research on policy implications and empirical validation

- Phase 4: Final synthesis

Theory of Change:

Goal: Enhance AI safety by developing safety approaches that remain effective as AI systems evolve.

- Short-term (0-2 years):
    - Technical: Identify limitations in current interpretability approaches
    - Governance: Help policymakers understand gaps in regulatory frameworks that use mechanistic interpretability
- Medium-term (2-5 years):
    - Technical: Develop new interpretability methods for adaptive AI systems and alternative methods for tracking system commitments
    - Governance: Evolve frameworks to account for substrate flexibility and improve safety assessment approaches
- Long-term (5+ years):
    - Widespread adoption of substrate-sensitive approaches for adaptive and self-modifying systems
    - Implementation of governance frameworks robust to architectural changes

Neglectedness & Timeliness:

- No structured investigation into substrate-flexibility, though it underpins several evasive threat models.
- Recent challenges in applying mechanistic interpretability to new architectures (e.g., Mamba). Developing secrecy in frontier model development.

Success Metrics:

- Regular feedback from research community via refinement of published document (tight feedback loops)
- Framework's ability to integrate case studies with theoretical literature and capture evasive scenarios (deep deceptiveness, sharp left-turn, robust agent-agnostic process)
- Clarification of ambiguous mechanistic interpretability definitions
- Generation of actionable insights

Team & Support:

- Collaborating with Chris Pang, under supervision from Sahil K (independent)
- Research management from Matthew Wearden

- Access to broader research community through LISA, High Actuation Spaces, MATS/PIBBSS Slack channels.

Risk Assessment:

No direct dual use but three key challenges:

1. Deficient literature base: Existing research may be insufficient for framework development.
2. Theoretical Limitations: Formalization may prove challenging for theoretical reasons.
3. Time Constraints: Timeline may be insufficient

Full Proposal

Management of Substrate-Sensitive AI Capabilities (MoSSAIC): Substrate-flexible risk[1]

Summary

Bailey et al. (2025) shows that neural networks can hide malicious behaviours from latent space monitors, by exploiting considerable flexibility over their neural representations. I argue that this is one aspect of a much broader threat. In this research, I aim to identify and clarify the paradigm of mechanistic reductionism that underpins much of AI safety (esp. interpretability) and introduce "substrate-flexible risk" as a novel threat model, one in which a fine-grained, bottom-up understanding of specific neural mechanisms increasingly fails to track macroscopic behaviours (e.g., deception, harmful content) as AI capabilities advance.

Through this, I hope to challenge the reliability of safety methods that depend heavily on mechanistic interpretation, as well as the safety-motivated interventions (e.g., freezing specific model weights) they are used for. By this, I hope to influence how resources are allocated in AI safety portfolios and contribute to the development of rigorous safety cases. I also hope to motivate and guide the development of alternative, non-reductionist approaches to interpretability, which are currently being explored in the wider High Actuation Spaces group.

The full project is discussed at length and motivated in this paper, which was accepted for poster presentation at the Tokyo AI Safety Conference 2025. This funding will allow me to develop the arguments presented in Sections 2–4 (see Threat Model) into a more compelling and theoretically grounded case for substrate-flexible risks and their implications, via a more thorough engagement with the mechanistic interpretability literature and the continual integration of expert feedback along the research development cycle.

[1] This is a refined/edited version of my MATS extension application, submitted several months ago. That research proposal is here. The final presentation poster is here. The work was started as part of the High Actuation Spaces Project for AI Safety Camp 2024.

Threat model

Recent work (Bailey et al., 2025) has indicated that networks enjoy considerable malleability over their neural representations. The paper demonstrates how a network can be trained to comply with harmful requests whilst evading a variety of latent-space detection methods, including linear probes, OOD detectors, and SAEs.

This suggests concerning behaviour. Although the researchers actively trained the network in question to evade latent space detection, the conceptual takeaway from the results is that the latent space of current neural networks contains vast amounts of redundancy that can in principle be used to evade detection from white-box interpretability techniques. In a red-team/blue-team game, the attackers were consistently able to retrain the network to obfuscate its activations and evade the blue team's retrained monitoring techniques (as many as 70 times). Further work by Bailey et al. has shown that networks can also be trained to "self-obfuscate" in response to a context clue that they are in an evaluation environment.[2]

This raises questions regarding the robustness of many methods AI safety researchers are currently developing to ensure the transition to AGI goes well. I am labelling these techniques as mechanistic; broadly, they comprise techniques that invoke some reduction from a high-level behaviour/representation (e.g., deception, complicity with a harmful request) to its lower-level neural correlates (e.g., activations) and intervening on those neural correlates to identify causal relations or to implement safety measures.

My research aims to clarify a broad class of threat models, one that includes the results of Bailey et al. (2025) and extends further to consider future scenarios in which such a fine-grained causal understanding of networks' internal structures fails to generalize. I believe this framework can be extended to include a number of MIRI's more pessimistic risk phenomena – most notably, deep deceptiveness and treacherous turn – which have so far evaded a clear formal treatment.

My research claims that the types of techniques tested in Bailey et al. (2025) – and the class of methods grouped under mechanistic interpretability more generally – are premised on two implicit assumptions that may not hold:

(i) (**Ontological**) The structural properties discovered in AI systems will remain relatively stable as AI capabilities increase.

[2] Unpublished, private communication with Luke Bailey

(ii) (**Epistemological**) The structural properties of neural networks can be reliably used to make safety assertions about AI systems.

Bailey et al.'s results suggest that the second assumption is already failing in current networks, and we anticipate that the problem will only worsen as networks increase in size, capabilities, autonomy, and flexibility. Earlier work by McGrath et al. (2023) also demonstrated a related problem of self-repair, in which networks which are intervened upon via ablations in early layers compensate via alternative mechanisms, thus calling into question the causal relationships validated by such interventions.

The following scenarios are all plausible ways in which mechanistic guarantees of safety might be weakened by the failure of one or more of the above assumptions.

- **Changes to scaffolding:** AI models are often scaffolded by auxiliary structures (e.g, mixture-of-experts). In this scenario, the core AI architecture (e.g., an LLM) does not change, but new scaffolding is provided that amplifies or unlocks latent capabilities.
- **Changes to architecture/paradigm.** Say that the current paradigm for machine learning, i.e., MLP, is eventually replaced by something else, perhaps the recently reported Kolmogorov–Arnold Networks (Liu et al., 2025). In this case, it is unclear whether the techniques we have, which were developed and work well within the current paradigm, will continue to work. It has recently been reported how mechanistic interpretability techniques need to be adapted for application to Mamba architectures (Ensign & Garriga-Alonso, 2024).[3] In a worst case scenario, we may have to start from square one in understanding a system already more capable than those currently in use. Similar problems may arise in a situation where frontier models are developed and their architectures updated under conditions of secrecy.
- **AI-assisted changes to architecture or paradigm.** AI systems could accelerate architectural evolution through assisting in the design of new neural architectures and training paradigms. While this represents an extension of current neural architecture search and automated ML techniques rather than a qualitative leap, its impact could parallel that of direct self-modification, especially under race dynamics.
- **Self-modification.** Self-modification represents a key vulnerability. A model capable of changing its own architecture without deferring updates to a human overseer would have significantly expanded capabilities for architectural adaptation.
- **Deep deception.** As Nate Soares describes in his post "Deep Deceptiveness" (Soares, 2023), an artificial intelligent system that is constrained by measures (ostensibly mechanistic) that prevent it from achieving its objectives is incentivized to modify its sub-processes (and even its

[3] The results of the cited work are in fact promising, and existing mech-interp techniques do seem to generalize well in this case. However, other architectures may not offer this affordance, especially if AI itself becomes the driving force behind architecture changes, or if more profound paradigm shifts are invoked.

representations of the world) such that those measures are not triggered, rather than correct its objectives. These sub-processes might register as benign under interpretability tools but may lead to unforeseen and unintended high-level behavior.

- **Aggregate Deception/RAAPs.** Aggregate deception is a box-inversion of deep deceptiveness, taking place within a wider ecosystem of advanced intelligence systems rather than a single system. Instead of sub-processes combining to produce unintended outcomes within a model, any particular representation could be distributed between systems.

Each of these scenarios breaks at least one of the assumptions. The first four I designate as *risk scenarios*; by themselves they do not present a hazard. However, each of the risk scenarios can feed into the two final *threat scenarios*, which are related to each other [i.e., via a box inversion (Kulveit, 2023)].

As discussed in the position paper (co-authored with Chris Pang and Sahil K, accepted and presented at the Tokyo AI Safety Conference), I consider headway can be made on these evasive risk phenomena by clarifying and expressing them in terms of "substrate." We provisionally define substrate as "the (programmable) environment or architecture in which a system is implemented." This definition is vague and relative, but in current models we want it to capture everything on the implementation and algorithmic levels of Marr's Three Levels of Computational Analysis (Marr, 1982). Gently put, it picks out the lower-level details of a network; this can include the hardware or computing paradigm (e.g., MLP vs. Kolmogorov-Arnold networks; implementation level) as well as the architecture of a network (e.g., transformer vs. state-space models; algorithmic level).

My research aims to clarify and formulate the above-mentioned risks by first clarifying the definition of substrate and then rewriting the risk scenarios in terms of it, all the while iterating on the MoSSAIC paper content via researcher feedback. I believe that an apt definition will be able to capture the malleability described in Bailey et al. (2025) and to further extend it to capture the pessimistic MIRI scenarios – treacherous, sharp left turn, and deep deceptiveness – which both explore the evasive behaviours that this malleability can facilitate.

Importance

The UK AISI recently stated (AISI, 2024) their clear intention to pursue safety cases as a way of ensuring the safety of models before deployment. Anthropic (Anthropic, 2024) and Google DeepMind (GDM, 2025) have both followed suit.

A safety case is defined as "A structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and valid case that a system is safe for a given application in a given environment." Many of the sketches so far provided by the UK AISI and others indicate that mechanistic techniques will have a strong role to play in supporting safety-case arguments for system deployment.

Bailey et al. (2025) makes it clear that these techniques are already vulnerable to the malleability of network representations. Future, more capable networks will have both (a) more parameters and therefore a larger latent space over which to hide activations and (b) greater flexible and strategic planning capabilities. Future safety cases supported by arguments from mechanistic and latent-space monitoring will likely have to be much stronger than those being considered today, and their weighting in an overall safety case may have to be reduced.

These safety cases carry both technical and political weight; they are already developing into effective coordination/collaboration interfaces between theoreticians, empirical researchers, and policy-makers. Despite the hedging and disclaimers, and despite the recognition that the techniques described in these safety cases do not yet exist, their presence in such coordination structures and their implicit backing from multiple reputable organisations makes their development more likely. Any weaknesses and overlooked assumptions in their proposed methods are therefore extremely significant.

To be clear, I am not arguing that mechanistic interpretability and other latent-space monitoring techniques are without value or redundant. I am also aware that nobody anticipates a world in which these techniques are applied in isolation. I am arguing that we should try to quantify the extent to which we can trust these techniques to deliver the results we are hoping for, so that they can be more appropriately supplemented, developed, and/or integrated into a realistic safety case template and AI safety portfolio.

Tractability

Recent research publications suggest that this work is becoming tractable. As my key example, mechanistic interpretability has recently been developing from a pre-paradigmatic assortment of techniques into something more substantial. It has been the subject of a comprehensive review paper (Bereska & Gavves, 2024), has been given a theoretical grounding via causal abstractions (Geiger et al. 2025), and has more recently been given a philosophical treatment via the philosophy of explanations (Ayonrinde & Jaburi 2025). The field seems to be becoming more conceptually coherent and therefore suitable for the type of analysis I intend to conduct. In terms of the definition of substrate, I cite Rosas et al. (2024) as one exciting approach I am keen to explore. I am in contact with most of these researchers.

Bailey et al.'s work has also recently been published and I anticipate it will become a major case study and motivator for others to start working on similar problems. I am also confident that substrate-flexible risks are tractable in terms of their policy and regulation implications. The recent push towards clarity on safety cases, and the adoption of those safety cases as a tool for coordinating theorists, engineers, and regulators provides a structure that emphasizes the role of critiques and allows those critiques to propagate across technical and governance domains. I contend that a compelling case for substrate flexible risks, even if impossible to demonstrate theoretically, will have an effect on the AI safety landscape.

Neglectedness

> *"Oh yeah, I hadn't thought of that"*
>
> [Leonard Bereska, upon reading my threat model at the Tokyo AI Safety Conference afterparty]

The above reaction from Bereska (who has recently produced a comprehensive review of mechanistic interpretability) suggests that the problems of substrate flexibility are still unexplored.

Some discussions are taking place but these are mostly informal and lack concreteness. There has been, to my knowledge, no structured investigation into the risks and likelihood of substrate flexibility, and certainly none that have attempted to define them via close reference to the extant literature on substrate-focused safety practices. Nate Soares, writing in March 2023, explicitly mentions that misalignment of the type discussed in his deep deceptiveness post was largely unrecognized, with none of the major labs acknowledging deep deception risks in their agendas. A number of other early MIRI-style arguments have also made claims that are directly relevant, though we are only just starting to see some of the empirical studies that might indicate some grounding out of the conceptual argumentation in current systems.

Plans and Deliverables

Plan Overview

Overall objective: Refine the conceptual arguments presented as the problem statement of the MoSSAIC paper (Sections 2–4) into more technical/formal ones such that it evidences more of its claims, is more accurate and realistic with respect to current mechanistic interpretability research, and better motivates and guides the subsequent research into solutions (Section 5). Publish findings/arguments.

1. **Phase 1: Paper Refinement + Upskilling**
    - Aims
        - Refine definitions/presentation of existing work
        - Suggest routes for further exploration
        - Upskill in relevant technical background
    - Deliverables
        - Refined version of MoSSAIC paper
        - Prioritized plan of core research and methodology (identify case studies, specific analogies to mature sciences, etc.)
2. **Phase 2: Core Research**
    - Aims
        - Investigate areas of interest
            - Distill theoretical research, using two assertions as guide
            - Explore case studies and parallels
            - Integrate findings
        - Increase technical detail and evidence of MoSSAIC paper
    - Deliverables
        - Technical paper/presentation (or series) incorporating findings
3. **Phase 3: Meta-level research**

- Aims
    - Embed the research in wider policy context (safety portfolios, governance mechanisms, etc.)
    - Suggest future empirical work/validation studies
    - Report failings, research gaps and paradigm shortcomings
- Deliverables
    - Specific posts on empirics/policy/negative findings/research gaps/etc.

4. **Phase 4: Summary/write-up/buffer**
    - Aims
        - Complete/refine all deliverables
    - Deliverables
        - Post/paper or set of papers

Total Duration: 6 months

Current/Completed work

Initial work was accepted and presented at the Tokyo AI Safety Conference and an updated version will be available soon on arXiv. Much of the work covered by this funding will consist of expanding the paper via further theoretical research and feedback from researchers. I will be publishing results through the process, to ensure tight feedback loops and effective/immediate dissemination of any conclusion.

Phase 1: Paper Refinement + Upskilling

In Phase 1, I intend to refine the paper via researcher feedback and to upskill in relevant areas in which my understanding is still weak.

Refinement

The MoSSAIC paper has already generated interest and response. I intend to develop the research via continued iteration of the content. There are a number of aligned researchers already pursuing related agendas, including Paul Colognese, Kola Ayonrinde, and Luke Bailey. All have reviewed the paper and expressed interest in further collaboration. I will seek to collaborate actively with these and other researchers and to implement/integrate their feedback into the work. I am interested in tight feedback loops from the community, to ensure the project claims are realistic.

Upskilling

I am lacking some of the technical expertise required to expand upon the completed work. I aim to address this initially, so that I can benefit more from conversations with researchers.

I will upskill in mechinterp as well as several adjacent fields; these include devinterp, as an example of a more theoretical perspective on mech-interp; causal incentives, as a well-developed cluster of theoretical safety-relevant ideas based around causal mechanisms; and computational mechanics, for it's more general, substrate-agnostic treatment of prediction modelling.

Phase 2: Core theoretical development

As argued in the paper, we contend that the theory of change for mechanistic safety research depends on two implicit assertions:

1. (**Ontological**) That the structural properties discovered in AI systems will be relatively stable as AI capabilities increase.
2. (**Epistemological**) That the structural properties of neural networks can be reliably used to make safety assertions about AI systems.

The core research phase will involve a more intense focus on the mechanistic interpretability literature, using the orientation developed during the upskilling and researcher feedback/paper refinement stage.

Note that in the above assertions, the term “structural properties” performs a similar role to “substrate.” I intend to define “substrate” such that it fits into the above claims and gives them rigor. My intention is to develop a definition iteratively, with reference to the literature and researcher feedback. Luke Bailey has suggested that a suitable definition of substrate would help in the design of experiments, and seems optimistic about the viability of such experiments. I view this work as trying to clarify. I intend to work backwards and forwards, producing theoretical definitions of “substrate” and the ontological and epistemological claims, assessing them against the literature and researcher feedback, updating the definitions and repeating the process. In terms of the engagement with the literature, I will use a flexible combination of the following three methodologies.

1. **Distillation:** I will attempt to distill the implicit assumptions and claims made across the mechanistic interpretability literature. This will involve examining how different researchers conceptualize what constitutes a "mechanism" and what it means to "interpret" these mechanisms.
2. **Case studies**: Studying examples of where mech-interp practices have not cleanly generalized will form a useful base for extrapolating from. I am initially considering Mamba architectures, as work has already been performed here. I am also considering the CNN/RNN—transformer transition, scaffolded transformers, and KA networks. By examining how mechanistic interpretability approaches succeed or fail when applied to novel mechanisms, I hope to identify which assumptions are fundamental to the field, clarify the claims and hopefully be able to cross-reference and update my provisional definitions. Bailey also thinks such case studies are where experiments will be most viable.
3. **Comparison against related mature sciences**: Interpretability was initially informed and inspired by biology, neuroscience, systems theory, and other existing sciences (see, e.g., Olah et al. (2020)); I hope to leverage some of these parallels to provide additional perspectives on my research questions. These fields have grappled with similar questions about understanding adaptive, evolving systems and may offer valuable frameworks. My current initial research bet is on the analogous debate taking place in neurology (Krakauer et al. (2017)).

Phase 3: Meta-level research

The exact nature of Phase 3 will depend on the preceding phases. Assuming no major alterations to the main thrust of the argument, I intend to assess some surrounding questions. I believe some of this meta-level research would have value independent of the value of the core theoretical work.

Policy implications

As mentioned above and detailed in my theory of change, I believe that the work could have some substantial policy implications. Given the lack of explicit literature on this topic, I intend to start from the safety case approaches described by AISI and Anthropic and to consult with governance researchers/strategists (not yet sought) and potentially seek further collaboration in order to produce governance-appropriate outputs. I have access to LISA and can make connections via my research manager.

Empirical grounding

I intend to focus explicitly on how this work can be made more empirical. Conversations with Luke Bailey—who has produced empirical evidence of the class of threat model I present—have been encouraging here; he feels that with an appropriate definition of substrate, and by examining the case studies of Mamba vs transformer, it should be straightforward to demonstrate the main arguments in a toy model.

Retrospective/further work/negative results

The tractability of this work is currently unclear, and I suspect that it will not go as intended. I will be conducting the work with an awareness of this; however, I also intend to spend some time specifically focusing on the failures and ways in which the work might be limited, as I believe this be reflective of and useful to mech-interp research in general. I intend to identify areas in which the literature is currently scant and underdeveloped (as measured by its ability to support ontological and epistemological claims). I may also focus on the ways in which mech-interp falls short of a paradigmatic science (Tan, 2024), as well as how this can be addressed. Please see Failure Modes.

Phase 4: Summary/Write-up/Buffer

This is where I will prepare/refine final deliverables. The final presentation/publication will depend on the nature of any results obtained and the recommendations of my supervisor and RM; at minimum, I intend for a more explicit, rigorous version of the paper already available, though I suspect that the final deliverables will take the form of a set of papers or a sequence. I will be soliciting feedback from researchers on my deliverables during their development and will in the write-up stage incorporate these or else leave clear suggestions for future work.

Failure modes

1. **Literature insufficient:**
    - This is a major concern and a major factor in my decision to approach the question from the three angles (distillation, case studies, comparison) stated above.
        - Mitigation: I will try a variety of techniques and check in with my colleagues, preparing to pivot to another approach or combine approaches. I am also prepared to investigate and publish on the shortcomings of mech-interp, thereby highlighting where work could be done to give it a more theoretical, general formulation.
2. **Impossible for theoretical reasons:**
    - Even mature sciences have imperfect ontologies/epistemologies, and I am concerned that mech-interp, when distilled, will suffer from similar ambiguities.
        - Mitigation: I believe that these sciences do still operate according to formal or informal paradigms that shape and guide research, and that these paradigms themselves have a certain tolerance to the underlying ambiguities. I hope that something similar to a paradigm can be extracted for mech-interp (cf

Suitability

- I have the support of Sahil K as a mentor and Matthew Wearden as RM. I co-authored the paper with Chris Pang, who is able to continue collaborating when his own projects allow.
- I am seeking funds to ensure part-time access to LISA. This will keep me close to many relevant people, including my RM.

- I have a number of connections to researchers in mech-interp, agent foundations, and biology (e.g., Manuel Baltieri, Alex Altair, Jacob Drori, Kola Ayonrinde, Luke Bailey), who can help me upskill and provide a soundboard for ideas.

Time-bounded

- I cannot say with certainty that I will achieve everything within the time-frame requested.
    - Mitigation: I will conduct regular reviews of my progress in consultation with my RM and adapt my timelines accordingly. From informal discussions with researchers, I believe that the attempt to distill mech-interp into a set of assumptions constitutes a main sticking point of the research, and I am willing to assign more time to this output (given its potential downstream use to my own further theory of change and others' work) or to pivot elsewhere (as described above), depending on feedback..

Assessment

Primary Feedback Loops

- Regular release of research materials throughout project timeline
- Frequent presentations to:
    - High actuation spaces project
    - LISA community
    - Other relevant research groups
- Continuous incorporation of feedback into developing framework, using initial paper as starting point.

Oversight

- Regular check-ins with mentor and research manager
- Milestone reviews to assess progress
- Adjustment of research direction based on guidance

- Integration of suggested resources and approaches

Success Indicators

- Framework's ability to:
    - Integrate case studies with theoretical literature (e.g., if a case study suggests a similar ontology to the one distilled from the theoretical literature).
    - Clarify ambiguous mech-interp definitions (e.g., "feature")
    - Generate useful insights for practitioners

Theory of Change

Goal

To enhance AI safety by critically examining the limitations of interpretability research that depends on current models/architectures/paradigms, in the context of plausible future directions in artificial intelligence, thereby informing more robust safety strategies and research directions in both technical and governance/regulatory domains.

Impacts

(T) denotes a tail impact (top-20%)

Short-term Impacts (0–2 years)

Technical Domain

1. Initial identification/clarification of the key challenges posed by both adaptive AI systems and model/architecture/paradigm changes to current safety approaches, i.e. of the substrate-flexible risks threat model. Increased awareness of potential limitations of mechanistic interpretability approaches, in both their current and anticipated future forms. **(T)** Redirection of funding based on these findings.
    - Relevant to: AI safety researchers, AI developers, funders

2. Validation of the High Actuation Spaces Project's hybrid MIRI—prosaic perspective. **(T)** Foundations of a robust theoretical framework for analyzing assumptions in mechanistic interpretability.
    - Relevant to: AI safety researchers, AI developers, academia

Governance Domain

1. Heightened awareness among policymakers of potential gaps in proposed regulatory control frameworks and the limits on empirical safety measures. Better calibration on the risks associated with ML paradigm shifts.
    - Relevant to: Policymakers, AI governance experts, regulatory bodies, AI strategy experts
2. Motivate the development of frameworks for assessing the robustness of AI safety practices in light of adaptive AI possibilities and paradigm shifts.
    - Relevant to: AI governance experts, policymakers

Medium-term Impacts (2–5 years)

Technical Domain

1. Development of new interpretability methods to handle adaptive AI systems. Increased importance of formal frameworks and provable safety approaches. New, substrate-sensitive formal methods for tracking AI system commitments (the broader research agenda of which MoSSAIC is a part is currently developing one candidate approach; see Section 5 of the MoSSAIC paper and the Live Theory sequence).
    - Relevant to: AI safety researchers, AI developers
2. **(T)** Evolution of AI development safeguards to mitigate the harms from potential substrate adaptations. (Note: dual-use risks.)
    - Relevant to: AI developers, tech companies, industry standards bodies

Governance Domain

1. Refinement of AI governance frameworks (and possible administration mechanisms) to account and prepare for the challenges from adaptive AI systems. More nuanced approach to AI safety funding allocation.
   - Relevant to: Policymakers, AI governance experts, alignment-focused non-profits, funders
2. Development of new standards for evaluating AI safety claims in light of substrate-flexibility.
   - Relevant to: Standards organizations, regulatory bodies/auditors

Long-term Impacts (5+ years)

Technical Domain

1. Development/Adoption of new, substrate-sensitive approaches for AI safety and risk management, that robustly handle adaptive and self-modifying systems, for improved understanding, control, and integration.
   - Relevant to: Entire AI safety ecosystem

Governance Domain

1. Implementation of more comprehensive and forward-looking AI governance frameworks. Improved trust in verification/safety practices.
   - Relevant to: Policymakers, international organizations, AI governance experts, tech companies

Key Assumptions

Major

1. Paradigm changes[4] and/or the development of adaptive, self-modifying AI systems are plausible near-future scenarios **but** are not developed within the next three years. If self-modifying systems are developed in the next three years, it is possible that the positive aspects of this work could not be implemented in time.
2. Mechanistic interpretability can be meaningfully analyzed and its limitations identified through theoretical work.

Minor

1. Regulations and/or governance will indeed lean strongly upon support from substrate-focused practices for auditing/mitigating deception (i.e., substrate-focused approaches are not abandoned through some other failure mode or made redundant by other, stronger methods).
2. Mechanistic interpretability will not be able to identify/address the problems of substrate flexibility without the influence of this research, or will indeed generalize effectively to new paradigms and into the self-modification regime.
3. AI governance communities will be receptive to the results, and will have the authority/power to implement any resulting actionables.

Potential Challenges and Mitigation Strategies

1. Limited immediate applicability to current AI systems
    - Mitigation: Clearly articulate the proactive nature of the research and its relevance to ongoing AI development trends.
2. Difficulty in empirically validating theoretical findings

[4] See here for a discussion of shallow paradigm changes on the path to AGI.

- Mitigation: Validation through feedback from relevant communities/experts, solicited promptly and continually and incorporated into ongoing research. Triangulation between literature, case studies, and associated mature sciences. Generation of answers/indicators towards unresolved problems in mech-interp (e.g., what is a feature?).

3. Complexity of translating theoretical insights into practical governance frameworks
   - Mitigation: Present to and seek advice from governance experts early in the process; develop clear, accessible explanations of key concepts.

Risk analysis

1. No likelihood of any dual-uses for this research at this stage. However, downstream risks might arise from the improved steering and misuse of systems via stronger interpretability tools. I will assess ongoingly and solicit advice from my RM towards final publication.

Appendix

Core Upskilling Material

Area	References	Focus	Time
Foundational Maths and Computing	- "Mathematics for Machine Learning" (Deisenroth et al.) - Review basic information theory (Shannon's "Mathematical Theory of Communication")	Linear algebra in high dimensions, probability theory, information theory	80-100
Core ML understanding	- "Deep Learning" by Goodfellow et al. (focus on chapters about network architecture) - "Transformers from Scratch" by Peter Bloem	Understanding transformers, attention mechanisms, and training dynamics	100
MechInterp Foundations	- Anthropic's "Transformer Circuits" series - "A Mathematical Framework for Transformer Circuits" (Elhage et al.) - "In-context Learning and Induction Heads" (Olsson et al.)	Understanding circuits approach, composition of network functions, superposition	40

	- Toy Models of Superposition (Elhage et al., 2022)		
DevInterp [In progress]	Timaeus reading list [Consult with Alex GO/others]	Understanding how capabilities develop during training, phase transitions, SLT background,	150

Supplementary Material

Area	References	Focus	Time
Recent reviews	- Mechanistic Interpretability for AI Safety: A Review (Bereska & Gavves, 2024) -Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks (Rauker et al., 2023)	Familiarity with current state of research	20

Attempts to theorize MechInterp	- Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability (Geiger et al., 2023) - Towards A Rigorous Science of Interpretable Machine Learning (Doshi-Velez & Kim, 2017) - Aarron Meuller’s more theoretical output - Chris Olah's earlier work	Investigate attempts in the direction of theoretical MI research	60
Criticisms	Daniel Tan’s “My Cruxes in Mechinterp” and Charbel-Raphaël’s “Against Almost Every Theory of Impact of Interpretability”	Understand current criticisms, including technical criticisms	20
Philosophy	"The Conscious Mind" by David Chalmers (especially parts on information spaces) "Gödel, Escher, Bach" by Hofstadter (for computational emergence concepts) Dennett's "From Bacteria to Bach and Back" (for emergence of intelligence)	How philosophical theories of mind relate to computational systems	

Advanced Topics and Integration	- Study phase transitions in physics and relate to model development - Examine emergent phenomena in both physics and neural networks - Read papers on lottery ticket hypothesis and network pruning	Drawing parallels between physical systems and neural networks	

Working draft

Management of Substrate-Sensitive AI Capabilities (MoSSAIC): Substrate-flexible risk[5]

Summary

Bailey et al. (2025) shows that neural networks can hide malicious behaviours from latent space monitors, by exploiting considerable flexibility over their neural representations. I argue that this is one aspect of a much broader threat. In this research, I aim to identify and clarify the paradigm of mechanistic reductionism that underpins much of AI safety (esp. interpretability) and introduce "substrate-flexible risk" as a novel threat model, one in which a fine-grained, bottom-up understanding of specific neural mechanisms increasingly fails to track macroscopic behaviours (e.g., deception, harmful content) as AI capabilities advance.

Through this, I hope to challenge the reliability of safety methods that depend heavily on mechanistic interpretation, as well as the safety-motivated interventions (e.g., freezing specific model weights) they are used for. By this, I hope to influence how resources are allocated in AI safety portfolios and contribute to the development of rigorous safety cases. I also hope to motivate and guide the development of alternative, non-reductionist approaches to interpretability, which are currently being explored in the wider High Actuation Spaces group.

The full project is discussed at length and motivated in this paper, which was accepted for poster presentation at the Tokyo AI Safety Conference 2025. This funding will allow me to develop the arguments presented in Sections 2–4 (see Threat Model) into a more compelling and theoretically grounded case for substrate-flexible risks and their implications, via a more thorough engagement with the mechanistic interpretability literature and the continual integration of expert feedback along the research development cycle.

[5] This is a refined/edited version of my MATS extension application, submitted several months ago. That research proposal is here. The final presentation poster is here. The work was started as part of the High Actuation Spaces Project for AI Safety Camp 2024.

Threat model

Recent work (Bailey et al., 2025) has indicated that networks enjoy considerable malleability over their neural representations. The paper demonstrates how a network can be trained to comply with harmful requests whilst evading a variety of latent-space detection methods, including linear probes, OOD detectors, and SAEs.

This suggests concerning behaviour. Although the researchers actively trained the network in question to evade latent space detection, the conceptual takeaway from the results is that the latent space of current neural networks contains vast amounts of redundancy that can in principle be used to evade detection from white-box interpretability techniques. In a red-team/blue-team game, the attackers were consistently able to retrain the network to obfuscate its activations and evade the blue team's retrained monitoring techniques (as many as 70 times). Further work by Bailey et al. has shown that networks can also be trained to "self-obfuscate" in response to a context clue that they are in an evaluation environment.[6]

This raises questions regarding the robustness of many methods AI safety researchers are currently developing to ensure the transition to AGI goes well. I am labelling these techniques as mechanistic; broadly, they comprise techniques that invoke some reduction from a high-level behaviour/representation (e.g., deception, complicity with a harmful request) to its lower-level neural correlates (e.g., activations) and intervening on those neural correlates to identify causal relations or to implement safety measures.

My research aims to clarify a broad class of threat models, one that includes the results of Bailey et al. (2025) and extends further to consider future scenarios in which such a fine-grained causal understanding of networks' internal structures fails to generalize. I believe this framework can be extended to include a number of MIRI's more pessimistic risk phenomena – most notably, deep deceptiveness and treacherous turn – which have so far evaded a clear formal treatment.

My research claims that the types of techniques tested in Bailey et al. (2025) – and the class of methods grouped under mechanistic interpretability more generally – are premised on two implicit assumptions that may not hold:

(i) (**Ontological**) The structural properties discovered in AI systems will remain relatively stable as AI capabilities increase.

[6] Unpublished, private communication with Luke Bailey

(ii) (**Epistemological**) The structural properties of neural networks can be reliably used to make safety assertions about AI systems.

Bailey et al.'s results suggest that the second assumption is already failing in current networks, and we anticipate that the problem will only worsen as networks increase in size, capabilities, autonomy, and flexibility. Earlier work by McGrath et al. (2023) also demonstrated a related problem of self-repair, in which networks which are intervened upon via ablations in early layers compensate via alternative mechanisms, thus calling into question the causal relationships validated by such interventions.

The following scenarios are all plausible ways in which mechanistic guarantees of safety might be weakened by the failure of one or more of the above assumptions.

- **Changes to scaffolding:** AI models are often scaffolded by auxiliary structures (e.g, mixture-of-experts). In this scenario, the core AI architecture (e.g., an LLM) does not change, but new scaffolding is provided that amplifies or unlocks latent capabilities.
- **Changes to architecture/paradigm.** Say that the current paradigm for machine learning, i.e., MLP, is eventually replaced by something else, perhaps the recently reported Kolmogorov–Arnold Networks (Liu et al., 2025). In this case, it is unclear whether the techniques we have, which were developed and work well within the current paradigm, will continue to work. It has recently been reported how mechanistic interpretability techniques need to be adapted for application to Mamba architectures (Ensign & Garriga-Alonso, 2024).[7] In a worst case scenario, we may have to start from square one in understanding a system already more capable than those currently in use. Similar problems may arise in a situation where frontier models are developed and their architectures updated under conditions of secrecy.
- **AI-assisted changes to architecture or paradigm.** AI systems could accelerate architectural evolution through assisting in the design of new neural architectures and training paradigms. While this represents an extension of current neural architecture search and automated ML techniques rather than a qualitative leap, its impact could parallel that of direct self-modification, especially under race dynamics.
- **Self-modification.** Self-modification represents a key vulnerability. A model capable of changing its own architecture without deferring updates to a human overseer would have significantly expanded capabilities for architectural adaptation.
- **Deep deception.** As Nate Soares describes in his post "Deep Deceptiveness" (Soares, 2023), an artificial intelligent system that is constrained by measures (ostensibly mechanistic) that prevent it from achieving its objectives is incentivized to modify its sub-processes (and even its

[7] The results of the cited work are in fact promising, and existing mech-interp techniques do seem to generalize well in this case. However, other architectures may not offer this affordance, especially if AI itself becomes the driving force behind architecture changes, or if more profound paradigm shifts are invoked.

representations of the world) such that those measures are not triggered, rather than correct its objectives. These sub-processes might register as benign under interpretability tools but may lead to unforeseen and unintended high-level behavior.

- **Aggregate Deception/RAAPs.** Aggregate deception is a box-inversion of deep deceptiveness, taking place within a wider ecosystem of advanced intelligence systems rather than a single system. Instead of sub-processes combining to produce unintended outcomes within a model, any particular representation could be distributed between systems.

Each of these scenarios breaks at least one of the assumptions. The first four I designate as *risk scenarios*; by themselves they do not present a hazard. However, each of the risk scenarios can feed into the two final *threat scenarios*, which are related to each other [i.e., via a box inversion (Kulveit, 2023)].

As discussed in the position paper (co-authored with Chris Pang and Sahil K, accepted and presented at the Tokyo AI Safety Conference), I consider headway can be made on these evasive risk phenomena by clarifying and expressing them in terms of "substrate." We provisionally define substrate as "the (programmable) environment or architecture in which a system is implemented." This definition is vague and relative, but in current models we want it to capture everything on the implementation and algorithmic levels of Marr's Three Levels of Computational Analysis (Marr, 1982). Gently put, it picks out the lower-level details of a network; this can include the hardware or computing paradigm (e.g., MLP vs. Kolmogorov-Arnold networks; implementation level) as well as the architecture of a network (e.g., transformer vs. state-space models; algorithmic level).

My research aims to clarify and formulate the above-mentioned risks by first clarifying the definition of substrate and then rewriting the risk scenarios in terms of it, all the while iterating on the MoSSAIC paper content via researcher feedback. I believe that an apt definition will be able to capture the malleability described in Bailey et al. (2025) and to further extend it to capture the pessimistic MIRI scenarios – treacherous, sharp left turn, and deep deceptiveness – which both explore the evasive behaviours that this malleability can facilitate.

Once we have a theoretical handle on this problem, I aim to start feeding the work into the remainder

Importance

The UK AISI recently stated (AISI, 2024) their clear intention to pursue safety cases as a way of ensuring the safety of models before deployment. Anthropic (Anthropic, 2024) and Google DeepMind (GDM, 2025) have both followed suit.

A safety case is defined as "A structured argument, supported by a body of evidence, that provides a compelling, comprehensible, and valid case that a system is safe for a given application in a given environment." Many of the sketches so far provided by the UK AISI and others indicate that mechanistic techniques will have a strong role to play in supporting safety-case arguments for system deployment. Anthropic suggest their use for detection of harmful representations [discuss anthropic's safety case—mech interp]

Bailey et al. (2025) makes it clear that these techniques are already vulnerable to the malleability of network representations. Future, more capable networks will have both (a) more parameters and therefore a larger latent space over which to hide activations and (b) greater flexible and strategic planning capabilities. Future safety cases supported by arguments from mechanistic and latent-space monitoring will likely have to be much stronger than those being considered today, and their weighting in an overall safety case may have to be reduced.

These safety cases carry both technical and political weight; they are already developing into effective coordination/collaboration interfaces between theoreticians, empirical researchers, and policy-makers. Despite the hedging and disclaimers, and despite the recognition that the techniques described in these safety cases do not yet exist, their presence in such coordination structures and their implicit backing from multiple reputable organisations makes their development more likely. Any weaknesses and overlooked assumptions in their proposed methods are therefore extremely significant.

To be clear, I am not arguing that mechanistic interpretability and other latent-space monitoring techniques are without value or redundant. I am also aware that nobody anticipates a world in which these techniques are applied in isolation. I am arguing that we should try to quantify the extent to which we can trust these techniques to deliver the results we are hoping for, so that they can be more appropriately supplemented, developed, and/or integrated into a realistic safety case template and AI safety portfolio.

Tractability

Recent research publications suggest that this work is becoming tractable. As my key example, mechanistic interpretability has recently been developing from a pre-paradigmatic assortment of techniques into something more substantial. It has been the subject of a comprehensive review paper (Bereska & Gavves, 2024), has been given a theoretical grounding via causal abstractions (Geiger et al. 2025), and has more recently been given a philosophical treatment via the philosophy of explanations (Ayonrinde & Jaburi 2025). The field seems to be becoming more conceptually coherent and therefore suitable for the type of analysis I intend to conduct. In terms of the definition of substrate, I cite Rosas et al. (2024) as one exciting approach I am keen to explore. I am in contact with most of these researchers.

Bailey et al.'s work has also recently been published and I anticipate it will become a major case study and motivator for others to start working on similar problems. I am also confident that substrate-flexible risks are tractable in terms of their policy and regulation implications. The recent push towards clarity on safety cases, and the adoption of those safety cases as a tool for coordinating theorists, engineers, and regulators provides a structure that emphasizes the role of critiques and allows those critiques to propagate across technical and governance domains. I contend that a compelling case for substrate flexible risks, even if impossible to demonstrate theoretically, will have an effect on the AI safety landscape.

Neglectedness

> *"Oh yeah, I hadn't thought of that"*
>
> [Leonard Bereska, upon reading my threat model at the Tokyo AI Safety Conference afterparty]

The above reaction from Bereska (who has recently produced a comprehensive review of mechanistic interpretability) suggests that the problems of substrate flexibility are still unexplored.

Some discussions are taking place but these are mostly informal and lack concreteness. There has been, to my knowledge, no structured investigation into the risks and likelihood of substrate flexibility, and certainly none that have attempted to define them via close reference to the extant literature on substrate-focused safety practices. Nate Soares, writing in March 2023, explicitly mentions that misalignment of the type discussed in his deep deceptiveness post was largely unrecognized, with none of the major labs acknowledging deep deception risks in their agendas. A number of other early MIRI-style arguments have also made claims that are directly relevant, though we are only just starting to see some of the empirical studies that might indicate some grounding out of the conceptual argumentation in current systems.

~~Notes: Mechanistic interpretability appears likely to form a part of future AI safety portfolios, and the recent advances and dedication of funding suggest it is being treated as a promising approach in ensuring the transition to AGI goes well. Major research institutions and AI labs have invested heavily in these techniques, though recent challenges in applying them to new architectures like Mamba highlight potential future limitations, as methods developed for traditional architectures in the MLP paradigm might not necessarily transfer "out-of-the-box" to novel substrates.~~

~~If mechanistic interpretability were found to be vulnerable to both the adaptive features of future AI systems and fundamental paradigm shifts in AI architectures, it would require a significant shift away from these practices towards improved approaches. This dual vulnerability would necessitate frameworks that can account for both dynamic system behavior and architectural evolution across different paradigms, rather than relying solely on understanding static mechanisms within current architectures.~~

~~The identification of substrate-flexible risks would also inform policy decisions in several key ways. Most notably, it would challenge the reliability of posited safety auditing procedures that depend heavily on mechanistic interpretation, as well as safety measures that involve interventions based upon such auditing (e.g., freezing specific model weights). It might also influence how resources are allocated in AI safety portfolios, by suggesting a need to diversify beyond mechanistic approaches.~~

Plans and Deliverables

Plan Overview

Overall objective: Refine the conceptual arguments presented as the problem statement of the MoSSAIC paper (Sections 2–4) into more technical/formal ones such that it evidences more of its claims, is more accurate and realistic with respect to current mechanistic interpretability research, and better motivates and guides the subsequent research into solutions (Section 5). Publish findings/arguments.

1. **Phase 1: Paper Refinement + Upskilling**
    - Aims
        - Refine definitions/presentation of existing work

    - Suggest routes for further exploration
    - Upskill in relevant technical background
- Deliverables
    - Refined version of MoSSAIC paper
    - Prioritized plan of core research and methodology (identify case studies, specific analogies to mature sciences, etc.)

2. **Phase 2: Core Research**
    - Aims
        - Investigate areas of interest
            - Distill theoretical research, using two assertions as guide
            - Explore case studies and parallels
            - Integrate findings
        - Increase technical detail and evidence of MoSSAIC paper
    - Deliverables
        - Technical paper/presentation (or series) incorporating findings
3. **Phase 3: Meta-level research**
    - Aims
        - Embed the research in wider policy context (safety portfolios, governance mechanisms, etc.)
        - Suggest future empirical work/validation studies
        - Report failings, research gaps and paradigm shortcomings
    - Deliverables
        - Specific posts on empirics/policy/negative findings/research gaps/etc.
4. **Phase 4: Summary/write-up/buffer**
    - Aims
        - Complete/refine all deliverables
    - Deliverables
        - Post/paper or set of papers

Total Duration: 6 months

Current/Completed work

Initial work was accepted and presented at the Tokyo AI Safety Conference and an updated version will be available soon on arXiv. Much of the work covered by this funding will consist of expanding the paper via further theoretical research and feedback from researchers. I will be publishing results through the process, to ensure tight feedback loops and effective/immediate dissemination of any conclusion.

Phase 1: Paper Refinement + Upskilling

In Phase 1, I intend to refine the paper via researcher feedback and to upskill in relevant areas in which my understanding is still weak.

Refinement

The MoSSAIC paper has already generated interest and response. I intend to develop the research via continued iteration of the content. There are a number of aligned researchers already pursuing related agendas, including Paul Colognese, Kola Ayonrinde, and Luke Bailey. All have reviewed the paper and expressed interest in further collaboration. I will seek to collaborate actively with these and other researchers and to implement/integrate their feedback into the work. I am interested in tight feedback loops from the community, to ensure the project claims are realistic.

Upskilling

I am lacking some of the technical expertise required to expand upon the completed work. I aim to address this initially, so that I can benefit more from conversations with researchers.

I will upskill in mechinterp as well as several adjacent fields; these include devinterp, as an example of a more theoretical perspective on mech-interp; causal incentives, as a well-developed cluster of theoretical safety-relevant ideas based around causal mechanisms; and computational mechanics, for it’s more general, substrate-agnostic treatment of prediction modelling.

Phase 2: Core theoretical development

As argued in the paper, we contend that the theory of change for mechanistic safety research depends on two implicit assertions:

1. (**Ontological**) That the structural properties discovered in AI systems will be relatively stable as AI capabilities increase.
2. (**Epistemological**) That the structural properties of neural networks can be reliably used to make safety assertions about AI systems.

The core research phase will involve a more intense focus on the mechanistic interpretability literature, using the orientation developed during the upskilling and researcher feedback/paper refinement stage.

Note that in the above assertions, the term “structural properties” performs a similar role to “substrate.” I intend to define “substrate” such that it fits into the above claims and gives them rigor. My intention is to develop a definition iteratively, with reference to the literature and researcher feedback. Luke Bailey has suggested that a suitable definition of substrate would help in the design of experiments, and seems optimistic about the viability of such experiments. I view this work as trying to clarify. I intend to work backwards and forwards, producing theoretical definitions of “substrate” and the ontological and epistemological claims, assessing them against the literature and researcher feedback, updating the definitions and repeating the process. In terms of the engagement with the literature, I will use a flexible combination of the following three methodologies.

1. **Distillation:** I will attempt to distill the implicit assumptions and claims made across the mechanistic interpretability literature. This will involve examining how different researchers conceptualize what constitutes a "mechanism" and what it means to "interpret" these mechanisms.
2. **Case studies**: Studying examples of where mech-interp practices have not cleanly generalized will form a useful base for extrapolating from. I am initially considering Mamba architectures, as work has already been performed here. I am also considering the CNN/RNN—transformer transition, scaffolded transformers, and KA networks. By examining how mechanistic interpretability approaches succeed or fail when applied

to novel mechanisms, I hope to identify which assumptions are fundamental to the field, clarify the claims and hopefully be able to cross-reference and update my provisional definitions. Bailey also thinks such case studies are where experiments will be most viable.

3. **Comparison against related mature sciences**: Interpretability was initially informed and inspired by biology, neuroscience, systems theory, and other existing sciences (see, e.g., Olah et al. (2020)); I hope to leverage some of these parallels to provide additional perspectives on my research questions. These fields have grappled with similar questions about understanding adaptive, evolving systems and may offer valuable frameworks. My current initial research bet is on the analogous debate taking place in neurology (Krakauer et al. (2017)).

Phase 3: Meta-level research

The exact nature of Phase 3 will depend on the preceding phases. Assuming no major alterations to the main thrust of the argument, I intend to assess some surrounding questions. I believe some of this meta-level research would have value independent of the value of the core theoretical work.

Policy implications

As mentioned above and detailed in my theory of change, I believe that the work could have some substantial policy implications. Given the lack of explicit literature on this topic, I intend to start from the safety case approaches described by AISI and Anthropic and to consult with governance researchers/strategists (not yet sought) and potentially seek further collaboration in order to produce governance-appropriate outputs. I have access to LISA and can make connections via my research manager.

Empirical grounding

I intend to focus explicitly on how this work can be made more empirical. Conversations with Luke Bailey—who has produced empirical evidence of the class of threat model I present—have been encouraging here; he feels that with an appropriate definition of substrate, and by examining the case studies of Mamba vs transformer, it should be straightforward to demonstrate the main arguments in a toy model.

Retrospective/further work/negative results

The tractability of this work is currently unclear, and I suspect that it will not go as intended. I will be conducting the work with an awareness of this; however, I also intend to spend some time specifically focusing on the failures and ways in which the work might be limited, as I believe this be reflective of and useful to mech-interp research in general. I intend to identify areas in which the literature is currently scant and underdeveloped (as measured by its ability to support ontological and epistemological claims). I may also focus on the ways in which mech-interp falls short of a paradigmatic science (Tan, 2024), as well as how this can be addressed. Please see Failure Modes.

Phase 4: Summary/Write-up/Buffer

This is where I will prepare/refine final deliverables. The final presentation/publication will depend on the nature of any results obtained and the recommendations of my supervisor and RM; at minimum, I intend for a more explicit, rigorous version of the paper already available, though I suspect that the final deliverables will take the form of a set of papers or a sequence. I will be soliciting feedback from researchers on my deliverables during their development and will in the write-up stage incorporate these or else leave clear suggestions for future work.

Failure modes

1. **Literature insufficient:**
   - This is a major concern and a major factor in my decision to approach the question from the three angles (distillation, case studies, comparison) stated above.

- Mitigation: I will try a variety of techniques and check in with my colleagues, preparing to pivot to another approach or combine approaches. I am also prepared to investigate and publish on the shortcomings of mech-interp, thereby highlighting where work could be done to give it a more theoretical, general formulation.

2. **Impossible for theoretical reasons:**
    - Even mature sciences have imperfect ontologies/epistemologies, and I am concerned that mech-interp, when distilled, will suffer from similar ambiguities.
        - Mitigation: I believe that these sciences do still operate according to formal or informal paradigms that shape and guide research, and that these paradigms themselves have a certain tolerance to the underlying ambiguities. I hope that something similar to a paradigm can be extracted for mech-interp (cf

Suitability

- I have the support of Sahil K as a mentor and Matthew Wearden as RM. I co-authored the paper with Chris Pang, who is able to continue collaborating when his own projects allow.
- I am seeking funds to ensure part-time access to LISA. This will keep me close to many relevant people, including my RM.
- I have a number of connections to researchers in mech-interp, agent foundations, and biology (e.g., Manuel Baltieri, Alex Altair, Jacob Drori, Kola Ayonrinde, Luke Bailey), who can help me upskill and provide a soundboard for ideas.

Time-bounded

- I cannot say with certainty that I will achieve everything within the time-frame requested.
    - Mitigation: I will conduct regular reviews of my progress in consultation with my RM and adapt my timelines accordingly. From informal discussions with researchers, I believe that the attempt to distill mech-interp into a set of assumptions constitutes a main sticking point of the research, and I am willing to assign more time to this output (given its potential downstream use to my own further theory of change and others' work) or to pivot elsewhere (as described above), depending on feedback..

Assessment

Primary Feedback Loops

- Regular release of research materials throughout project timeline
- Frequent presentations to:
  - High actuation spaces project
  - LISA community
  - Other relevant research groups
- Continuous incorporation of feedback into developing framework, using initial paper as starting point.

Oversight

- Regular check-ins with mentor and research manager
- Milestone reviews to assess progress
- Adjustment of research direction based on guidance
- Integration of suggested resources and approaches

Success Indicators

- Framework's ability to:
  - Integrate case studies with theoretical literature (e.g., if a case study suggests a similar ontology to the one distilled from the theoretical literature).
  - Clarify ambiguous mech-interp definitions (e.g., "feature")
  - Generate useful insights for practitioners

Theory of Change

Goal

To enhance AI safety by critically examining the limitations of interpretability research that depends on current models/architectures/paradigms, in the context of plausible future directions in artificial intelligence, thereby informing more robust safety strategies and research directions in both technical and governance/regulatory domains.

Impacts

(T) denotes a tail impact (top-20%)

Short-term Impacts (0–2 years)

Technical Domain

3. Initial identification/clarification of the key challenges posed by both adaptive AI systems and model/architecture/paradigm changes to current safety approaches, i.e. of the substrate-flexible risks threat model. Increased awareness of potential limitations of mechanistic interpretability approaches, in both their current and anticipated future forms. **(T)** Redirection of funding based on these findings.
    - Relevant to: AI safety researchers, AI developers, funders

4. Validation of the High Actuation Spaces Project's hybrid MIRI—prosaic perspective. **(T)** Foundations of a robust theoretical framework for analyzing assumptions in mechanistic interpretability.
    - Relevant to: AI safety researchers, AI developers, academia

Governance Domain

3. Heightened awareness among policymakers of potential gaps in proposed regulatory control frameworks and the limits on empirical safety measures. Better calibration on the risks associated with ML paradigm shifts.
    - Relevant to: Policymakers, AI governance experts, regulatory bodies, AI strategy experts
4. Motivate the development of frameworks for assessing the robustness of AI safety practices in light of adaptive AI possibilities and paradigm shifts.
    - Relevant to: AI governance experts, policymakers

Medium-term Impacts (2–5 years)

Technical Domain

3. Development of new interpretability methods to handle adaptive AI systems. Increased importance of formal frameworks and provable safety approaches. New, substrate-sensitive formal methods for tracking AI system commitments (the broader research agenda of which MoSSAIC is a part is currently developing one candidate approach; see Section 5 of the MoSSAIC paper and the Live Theory sequence).
    - Relevant to: AI safety researchers, AI developers
4. **(T)** Evolution of AI development safeguards to mitigate the harms from potential substrate adaptations. (Note: dual-use risks.)
    - Relevant to: AI developers, tech companies, industry standards bodies

Governance Domain

3. Refinement of AI governance frameworks (and possible administration mechanisms) to account and prepare for the challenges from adaptive AI systems. More nuanced approach to AI safety funding allocation.
    - Relevant to: Policymakers, AI governance experts, alignment-focused non-profits, funders
4. Development of new standards for evaluating AI safety claims in light of substrate-flexibility.
    - Relevant to: Standards organizations, regulatory bodies/auditors

Long-term Impacts (5+ years)

Technical Domain

2. Development/Adoption of new, substrate-sensitive approaches for AI safety and risk management, that robustly handle adaptive and self-modifying systems, for improved understanding, control, and integration.
    - Relevant to: Entire AI safety ecosystem

Governance Domain

2. Implementation of more comprehensive and forward-looking AI governance frameworks. Improved trust in verification/safety practices.
    - Relevant to: Policymakers, international organizations, AI governance experts, tech companies

Key Assumptions

Major

3. Paradigm changes[8] and/or the development of adaptive, self-modifying AI systems are plausible near-future scenarios **but** are not developed within the next three years. If self-modifying systems are developed in the next three years, it is possible that the positive aspects of this work could not be implemented in time.
4. Mechanistic interpretability can be meaningfully analyzed and its limitations identified through theoretical work.

Minor

4. Regulations and/or governance will indeed lean strongly upon support from substrate-focused practices for auditing/mitigating deception (i.e., substrate-focused approaches are not abandoned through some other failure mode or made redundant by other, stronger methods).
5. Mechanistic interpretability will not be able to identify/address the problems of substrate flexibility without the influence of this research, or will indeed generalize effectively to new paradigms and into the self-modification regime.
6. AI governance communities will be receptive to the results, and will have the authority/power to implement any resulting actionables.

Potential Challenges and Mitigation Strategies

4. Limited immediate applicability to current AI systems
    - Mitigation: Clearly articulate the proactive nature of the research and its relevance to ongoing AI development trends.
5. Difficulty in empirically validating theoretical findings

[8] See here for a discussion of shallow paradigm changes on the path to AGI.

    - Mitigation: Validation through feedback from relevant communities/experts, solicited promptly and continually and incorporated into ongoing research. Triangulation between literature, case studies, and associated mature sciences. Generation of answers/indicators towards unresolved problems in mech-interp (e.g., what is a feature?).

6. Complexity of translating theoretical insights into practical governance frameworks
    - Mitigation: Present to and seek advice from governance experts early in the process; develop clear, accessible explanations of key concepts.

Risk analysis

2. No likelihood of any dual-uses for this research at this stage. However, downstream risks might arise from the improved steering and misuse of systems via stronger interpretability tools. I will assess ongoingly and solicit advice from my RM towards final publication.

Appendix

Core Upskilling Material

Area	References	Focus	Time
Foundational Maths and Computing	- "Mathematics for Machine Learning" (Deisenroth et al.) - Review basic information theory (Shannon's "Mathematical Theory of Communication")	Linear algebra in high dimensions, probability theory, information theory	80-100
Core ML understanding	- "Deep Learning" by Goodfellow et al. (focus on chapters about network architecture) - "Transformers from Scratch" by Peter Bloem	Understanding transformers, attention mechanisms, and training dynamics	100
MechInterp Foundations	- Anthropic's "Transformer Circuits" series - "A Mathematical Framework for Transformer Circuits" (Elhage et al.) - "In-context Learning and Induction Heads" (Olsson et al.)	Understanding circuits approach, composition of network functions, superposition	40

	- Toy Models of Superposition (Elhage et al., 2022)		
DevInterp [In progress]	Timaeus reading list [Consult with Alex GO/others]	Understanding how capabilities develop during training, phase transitions, SLT background,	150

Supplementary Material

Area	References	Focus	Time
Recent reviews	- Mechanistic Interpretability for AI Safety: A Review (Bereska & Gavves, 2024) -Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks (Rauker et al., 2023)	Familiarity with current state of research	20

Attempts to theorize MechInterp	- Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability (Geiger et al., 2023) - Towards A Rigorous Science of Interpretable Machine Learning (Doshi-Velez & Kim, 2017) - Aarron Meuller's more theoretical output - Chris Olah's earlier work	Investigate attempts in the direction of theoretical MI research	60
Criticisms	Daniel Tan's "My Cruxes in Mechinterp" and Charbel-Raphaël's "Against Almost Every Theory of Impact of Interpretability"	Understand current criticisms, including technical criticisms	20
Philosophy	"The Conscious Mind" by David Chalmers (especially parts on information spaces) "Gödel, Escher, Bach" by Hofstadter (for computational emergence concepts) Dennett's "From Bacteria to Bach and Back" (for emergence of intelligence)	How philosophical theories of mind relate to computational systems	

Advanced Topics and Integration	- Study phase transitions in physics and relate to model development - Examine emergent phenomena in both physics and neural networks - Read papers on lottery ticket hypothesis and network pruning	Drawing parallels between physical systems and neural networks	

New threat model presentation

Threat model

We claim that the field of mechanistic interpretability operates under two implicit assertions: (i) the structural properties discovered in AI systems will remain relatively stable as AI capabilities increase and (ii) the structural properties of neural networks can be reliably used to make safety assertions about AI systems.

That is, in focusing on the substrate-level mechanisms of individual neurons and circuits, mech-interp overfits to current paradigms/models. As a result, AI safety portfolios that rely heavily on mech-interp for model auditing might fail to take into account the adaptive features of intelligences. We identify the cluster of problems associated with this failure as "substrate-flexible risks."[9]

Broadly speaking, we argue that there are several plausible scenarios in which mechanistic understandings of AI may break down; these can be strung together into an escalating narrative:

- **Changes to scaffolding:** AI models are often scaffolded by auxiliary structures (e.g, mixture-of-experts). In this scenario, the core AI architecture (e.g., an LLM) does not change, but new scaffolding is provided that amplifies or unlocks latent capabilities.
- **Changes to architecture/paradigm.** Say that the current paradigm for machine learning, i.e., MLP, is eventually replaced by something else, perhaps the recently reported KANs. In this case, it is unclear whether the techniques we have, which were developed and work well within the current paradigm, will continue to work. It has recently been reported how mechanistic interpretability techniques need to be adapted for application to Mamba architectures.[10] In a worst case scenario, we may have to start from square one in understanding a system already more capable than those currently in use. Similar problems may arise in a situation where frontier models are developed and their architectures updated under conditions of secrecy.
- **AI-assisted changes to architecture or paradigm.** AI systems could accelerate architectural evolution through assisting in the design of new neural architectures and training paradigms. While this represents an extension of current neural architecture search and automated ML techniques rather than a qualitative leap, its impact could parallel that of direct self-modification, especially under race dynamics

[9] For those familiar with MIRI and ACS arguments, an introduction to the threat model is given here, where we posit the phenomena of deep deceptiveness and robust agent-agnostic processes as being two box-inverted# examples of the same evasive adaptivity, to indicate the potential blind-spots of mechanistic approaches to interpretability and safety.

[10] The results of the cited work are in fact promising, and existing mech-interp techniques do seem to generalize well in this case. However, other architectures may not offer this affordance, especially if AI itself becomes the driving force behind architecture changes, or if more profound paradigm shifts are invoked.

- **Self-modification.** Self-modification represents a key vulnerability. A model capable of changing its own architecture without deferring updates to a human overseer would have significantly expanded capabilities for architectural adaptation.
- **Deep deception.** As Nate Soares describes, in his post "Deep Deceptiveness," an artificial intelligent system that is constrained by measures (ostensibly mechanistic) that prevent it from achieving its objectives is incentivized to modify its sub-processes (and even its representations of the world) such that those measures are not triggered, rather than correct its objectives. These sub-processes might look benign but may lead to unforeseen and unintended high-level behavior.
- **Aggregate Deception/RAAPs.** Aggregate deception is a box-inversion of deep deceptiveness, taking place within a wider ecosystem of advanced intelligence systems rather than a single system. Instead of sub-processes combining to produce unintended outcomes within a model, any particular representation could be distributed between systems.

These are all interconnected by a common feature: a mechanistic understanding of the fine-grained details of current intelligences may fail to generalize to future intelligent systems. Such future intelligences may implement novel architectures/mechanisms and are likely to be able to flexibly re-adapt those mechanisms in order to achieve their (potentially misaligned) objectives.

Importance

Mechanistic interpretability appears likely to form a part of future AI safety portfolios, and the recent advances and dedication of funding suggest it is being treated as a promising approach in ensuring the transition to AGI goes well. Major research institutions and AI labs have invested heavily in these techniques, though recent challenges in applying them to new architectures like Mamba highlight potential future limitations, as methods developed for traditional architectures in the MLP paradigm might not necessarily transfer "out-of-the-box" to novel substrates.

If mechanistic interpretability were found to be vulnerable to both the adaptive features of future AI systems and fundamental paradigm shifts in AI architectures, it would require a significant shift away from these practices towards improved approaches. This dual vulnerability would necessitate frameworks that can account for both dynamic system behavior and architectural evolution across different paradigms, rather than relying solely on understanding static mechanisms within current architectures.

The identification of substrate-flexible risks would also inform policy decisions in several key ways. Most notably, it would challenge the reliability of posited safety auditing procedures that depend heavily on mechanistic interpretation, as well as safety measures that involve interventions based upon such auditing (e.g., freezing specific model weights). It might also influence how resources are allocated in AI safety portfolios, by suggesting a need to diversify beyond mechanistic approaches.

Tractability

The tractability of grounding this threat model in technical detail remains unclear at this stage. However, I believe that these arguments do not need to be fully grounded out in order to be of use to the AI safety community; I further believe that the tractability of producing a strong case for concern in governance/policy is high. For more specific detail, see the Failure Modes, Key Assumptions, and Potential Challenges.

Neglectedness

Discussions at MATS and LISA have indicated that the assumptions of mechanistic interpretability have yet to be clearly formulated, and questions about how far these substrate-dependent safety methods will generalize remain open.

Nate Soares, writing in March 2023, argued that misalignment of the type discussed in his deep deceptiveness post was largely unrecognized, with none of the major labs acknowledging deep deception risks in their agendas.

Whilst discussions are taking place surrounding the topics of substrate flexibility, these are mostly informal and lack concreteness. There has been, to my knowledge, no structured investigation into the risks and likelihood of substrate flexibility, and certainly none that have attempted to define them via close reference to the extant literature on substrate-focused safety practices.