MoSSAIC: AI Safety After Mechanism

Matt Farr   Aditya Arpitha Prasad   Chris Pang   Sahil K

1. Abstract

We identify a causal–mechanistic paradigm in AI safety, primarily through the example of mechanistic interpretability. Recent results suggest limits to this paradigm's utility in answering questions about the safety of neural networks, and we argue further that those results give a taste of what is to come, by proposing a sequence of scenarios in which safety affordances based upon the causal–mechanistic paradigm break down. This analysis conceptually connects current obfuscation results with some of MIRI's more pessimistic threat models (e.g., deep deceptiveness, robust agent-agnostic processes) and suggest how we might unify all under a common framework. The paper then introduces a supplementary framework, MoSSAIC (Management of Substrate-Sensitive AI Capabilities), which addresses some of the core assumptions that underlie the causal–mechanistic paradigm, and we sketch out the complementary research infrastructure we are currently designing to allow us to keep pace with evasive intelligence.

2. Introduction

Neural networks (NNs) are famously described as black boxes. Their inner workings resist reduction to human-understandable concepts. [4; 43] Their decision-making processes are therefore difficult to properly audit to ensure safety. [43]

Neural networks are also increasingly being deployed to make decisions on behalf of humans across high-stakes domains. Our lack of understanding of how trained NNs process information to arrive at decisions poses challenges to their safe deployment. [43]

The sub-field of AI safety known as interpretability seeks to produce human-understandable explanations of NN behaviors. [7] [31]

Bereska & Gavves (2024) classify interpretability approaches into four main paradigms, which we'll take a quick look at:

Behavioral Interpretability treats models as black boxes, analyzing input-output relations without examining internal processes. They are model-agnostic and practical for complex systems but lack real insight into internal processes. [7] For example, minimal pair testing compares model outputs on almost-identical inputs to test for specific linguistic capabilities (e.g., ”The cat sat on the mat” vs. ”The cats sat on the mat” to test pluralization), and perturbation analysis systematically alters inputs to see how the output changes (e.g., testing robustness to adversarial examples).

Attributional Interpretability examines how individual features of the input affect the output, using gradient-based methods. These approaches offer more transparency over black-box methods, but still do not provide any information on the internal structures of models. [7] The simplest version of this is vanilla gradients, which computes the gradient of the output with respect to a change in some input feature. Subsequent versions offer more refined techniques on this basic premise.

Concept-based Interpretability seeks high-level concepts governing network behavior. [7] For instance, it might classify model outputs into honest and dishonest categories, take averages over the intermediate activations for each class and work out the difference between these averages as a vector in latent space, representing the concept “honesty.” This paradigm allows for ”representation engineering”—manipulating these internal representations to upregulate desirable concepts. [43]

Mechanistic Interpretability starts from the bottom, identifying clusters of neurons (called ”circuits”) that together perform a function in the decision-making process, from there seeking to understand the relations between these circuits and how these give rise to system behavior. [43] [7] [28] This field treats human-understandable ”features” as the fundamental unit of analysis, trying to isolate these via a number of techniques. [7]

Here we're focusing on mechanistic interpretability. It has remained a popular subject in AI safety over the past few years, receiving considerable funding and attention. As a rough indication of current investment, we note that approximately a third of the topics listed in Open Philanthropy's recent request for technical AI safety proposals are on mechanistic interpretability or are closely related. [33]

3. Mechanistic Interpretability

The field of mechanistic interpretability (MI) is not a single, monolithic research program but rather a rapidly evolving collection of methods, tools, and research programs. These are united by the shared ambition of reverse-engineering neural computations and, though lacking a comprehensive uniform methodology, typically apply tools of causal analysis to understand a model from the bottom up.

MI research is built on a set of postulates. A central assumption is that NN representations can in principle be decomposed into interpretable ”features,” fundamental units that ”cannot be further decomposed into smaller, distinct factors”—and that these are often encoded linearly as directions in activation space. [43] [7] Further work has shown that NNs in fact compress features such that multiple features are encoded by the same neuron—a phenomenon called superposition. [18] [7] [19]

Some examples of mechanistic techniques include the following:

- **Linear Probes**: Simple models (usually linear classifiers) are trained to predict a specific property (e.g., the part-of-speech of a word) from a model's internal activations. The success or failure of a probe at a given layer is used to infer whether that information is explicitly represented there.
- **Logit Lens**: This technique applies the final decoding layer of the model to intermediate activations, to observe how its prediction evolves layer-by-layer. [7] [6]
- **Sparse Autoencoders**: These attempt to uncompress a NN's features such that they become monosemantic. This allows researchers to observe the monosemantic features used by a NN in its computation.[14; 7; 3]
- **Activation patching**: This technique attempts to isolate circuits of the network responsible for specific behaviours, by replacing a circuit active for a specific output with another, to test the counterfactual hypothesis.[7; 28; 19]

More recently, mechanistic interpretability has been developing from a pre-paradigmatic assortment of techniques into something more substantial. It has been the subject of a comprehensive review paper[7], has been given a theoretical grounding via causal abstractions[19], and has more recently been given a philosophical treatment via the philosophy of explanations.[4]

In particular, this philosophical treatment characterizes MI as the search for explanations with these four properties:

1. **Causal-Mechanistic** - providing step-by-step causal chains of how a computation happens. This contrasts with attribution methods like saliency maps, which are primarily correlational. A saliency map might show that pixels corresponding to a cat's whiskers are ”important” for a classification, but it doesn't explain the mechanism of how the model processes that whisker information through subsequent layers to arrive at its decision.
2. **Ontic** - MI researchers believe they are discovering real structures (the ontology) within the model. This differs from a purely epistemic approach, which might produce a useful analogy or simplified story that helps human understanding but doesn't claim to perfectly represent reality. The search for ”features” as fundamental, linearly encoded units in activation space is a core ontic commitment of the field.
3. **Falsifiable** - MI explanations are framed as testable hypotheses that can be empirically refuted. The claim that ”this specific set of neurons and attention heads forms a circuit for detecting syntax” is falsifiable. One can perform a causal intervention—such as activation patching or ablating the circuit and observe if the model's syntactic capabilities break in a predictable way. This contrasts with unfalsifiable, post-hoc stories that can't be rigorously tested.
4. **Model-level** - The focus of MI is on the neural network itself, its weights, activations, and architecture. This is distinct from system-level analysis, which might examine the behavior of an entire deployed product (e.g., a chatbot integrated with search tools and a chain-of-thought prompting wrapper). A system-level explanation might attribute a behavior to the prompt, whereas a model-level explanation seeks the mechanism within the neural network's computation graph.

Leaving aside questions regarding the aptness of this philosophical characterization, we focus here on the Causal-Mechanistic nature of the explanations that mechanistic interpretability produces. **We argue that MI exemplifies what we term the ”Causal-Mechanistic Paradigm” in AI safety, and we go on to present some scenarios in which this paradigm fails to deliver reliable safety guarantees.**

The *causal-mechanistic paradigm*: the belief that we can decompose AI systems into smaller causal components by analysing their computational structure, and from these components reconstruct precise explanations of the behaviour of AI systems. From Bereska, Greaves et al.: ”The theory of causality (Pearl, 2009) provides a mathematically precise language for mechanistic interpretability, forming the foundation for understanding high-level semantics in neural representations (Geiger et al., 2023a). Treating neural networks as causal models involves **considering the compute graph as the causal graph**, allowing for precise interventions and examining individual parameters' roles (McGrath et al., 2023)." [7] *[emphasis ours]* In other words, a direct relationship is drawn between the *inner structure* of an AI system (i.e., the compute graph) and the methods (i.e., interventions upon the corresponding causal graph) we can use to reliably predict all the *relevant outcomes*. That is, we can understand the relevant outcomes of an AI system if and only if we understand its inner structure.

However, recent work indicates a fundamental problem in how we interpret NNs. Bailey et al. (2024) applied a variety of latent space monitoring techniques to prevent networks from outputting harmful content. They then used a series of black-box and white-box jailbreak methods to coax a NN to evade these monitors by obfuscating its activations, whilst preserving the harmful output. [5]

The latent space monitoring methods they tested included sparse autoencoders, linear and MLP probes, and out-of-distribution detectors. All use the internal activations of the model to identify where harmful output is generated. They all embody the broader assumption that understanding internal structure allows for behavioral prediction and control.

Using obfuscation attacks—joint optimization procedures that simultaneously preserve target behaviors while evading latent monitors—Bailey et al. were able to make their network evade all tested latent space monitors.

Even more worryingly, they demonstrated just how robustly a NN can obfuscate its activations. They conducted red-team/blue-team games in which defenders iteratively retrained their detection methods using previously obfuscated activations, and attackers adapted their methods to evade each new iteration. The NN was able to successfully obfuscate its activations more than 70 times.

Whilst these result have arisen in a training environment explicitly designed to induce obfuscation, the main conceptual takeaway—and the hypothesis put forward by the authors—is this:

> *"Models can represent concepts and execute behaviors using a broad set of different activations. That is, neural representations are highly malleable: we can reshape activation patterns in a variety of ways, while preserving a network's behavior."*

Following similar trends in the philosophy of mind and science, we might call this the *multiple realizability of neural representations*. Rather than "harmfulness'" consisting of a single direction in latent space-or even a discrete set of identifiable circuits-Bailey et al.'s evidence suggests it can be represented through numerous distinct activation patterns, many of which can be found within the distribution of benign representations.

Such multiple realizability is deeply concerning. We submit that the results of Bailey et al. should be viewed not simply as a technical challenge to be overcome through better monitoring techniques, but as indicating a theoretical limit to the causal-mechanistic paradigm. We further believe that it forms part of a developing threat model: *substrate-flexible risk*, as described in the following section. As NNs become ever more capable and their latent spaces inevitably become larger, we anticipate substrate-flexible risks to become increasingly significant for the AI safety landscape.

4. Problematic scenarios for the causal-mechanistic paradigm

We first briefly overview our critique of the causal-mechanistic paradigm in AI interpretability.

4.1. Overview of Scenarios

We contend that the causal-mechanistic paradigm in AI safety research makes two implicit assertions:

1. **Fixity of structure**: That the structural properties [1] discovered in AI systems will be relatively stable as AI capabilities increase. [2]
2. **Reliability of extrapolation**: That the structural properties of neural networks can be reliably used to make safety assertions about AI systems.

If these assertions hold, we will be able to reliably uncover structural properties that lead to misaligned behaviour, and create either (i) new model architectures or training regimes that don't possess those properties, (ii) low-level interventions that address those properties in existing systems, or (iii) high-level interventions that take advantage of stable low-level properties.

We believe there are scenarios in the near or medium-term future that will challenge these assertions. We outline these scenarios below:

- **Scaffolding shift** - the core AI architecture (e.g., an LLM) does not change, but new tools are provided that amplify or unlock latent capabilities, for example

[1] Note that the term "structural properties" is ambiguous and important in these assertions. We will partially resolve this in the next section, though indeed much of the work involved in MoSSAIC is clarifying what these structural properties are.

[2] Correspondingly, the techniques with which researchers discover structural properties will also remain relevant as capabilities increase.

changes in decoding algorithm, meta-decoding algorithms or access to tool use within some agent scaffolding.

- **Human-initiated paradigm shift** - a new machine learning paradigm or architecture is discovered that is more efficient and capable but breaks from existing, legible paradigms.
- **AI-assisted paradigm shift** - Automated R&D is used to create paradigms humans have limited understanding and influence over.
- **Self-modifying AI systems** - AI systems with the high level (i.e. not backpropagation/SGD-based) of ability to modify their own model architecture or give themselves new tools.
- **Deep deceptiveness** - Models able to reconfigure their internal representations at a deep level to evade human scrutiny.
- **Robust Agent-Agnostic Processes** - Even if individual models are ”safe” in their operations, there may be an overall context in which models acting together produce unsafe outcomes

We see the above list as showing a rough scale from relatively limited to very radical modification of the architecture and structures behind AI systems, such that the AI system effectively evades any interventions humans have created based on mechanistic assumptions. From the point of view of MoSSAIC, we think that there is a significant theme underlying all of these, namely that of a shift in the substrate of a model.

4.2. Substrates

We provisionally define a substrate as the (programmable) environment or architecture in which a system is implemented. In other words, it is the essential context that enables an algorithm to be implemented beyond the whiteboard. As a useful reference point that is already established in the literature—and without committing ourselves to the strict level separation proposed—we cite Marr’s three levels of analysis. Marr defines three levels on which an information processing system can be analyzed. [27] These are explained below via his example of a cash register.

1. **Computational**: the actual process that is being performed. For the cash register, this is the details of addition as defined algebraically (associative, transitive, etc.)
2. **Algorithmic**: the particular method by which it is performed. A cash register uses a base 10 number system, though it could of course use binary.
3. **Implementation**: the physical system that realizes the above processes. This would be the specific mechanical gears of the register.

We position “substrate” as capturing both the algorithmic and implementational levels. As an example from the AI domain, an LLM performs the task of next token prediction at the computational level, this is implemented on the transformer architecture consisting of attention and MLP layers (algorithmic substrate), which are implemented in a physical substrate of GPUs.

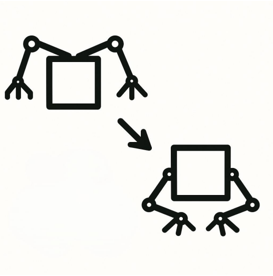
As another (non-AI) example, Conway’s Game of Life and von Neumann architectures can both be used to implement Turing machines. [34] [22] As such, both are in principle capable of executing any computer algorithm. However, even a deep understanding of Conway’s Game of Life would not help us debug or improve a complex application (e.g. a climate simulation, or a video game) designed to run on a conventional computer. In this case, the differences between the substrates render cross-domain knowledge transfer difficult.

It is also important to note that substrates are usually nested within other substrates. Functional and object-oriented programming languages, for example, are themselves substrates built upon lower-level substrates like x86 or ARM assembly. Both require different forms of interface and exhibit different domains of applicability.

Furthermore, substrates are not mere variations of a theoretical “perfect machine” – their differences have direct implications on their functionality.[21] For example, modern AI systems are built upon deep neural networks that exponentially scale in size and rely on matrix multiplication because one particular hardware-level substrate makes scaling matrix multiplications easier: GPUs with the CUDA toolkit. [24] [13] At a lower level, the divide between RISC and CISC instruction sets is a more realistic example of differences in substrate leading to significant differences in application programming. [9]

This dependence on substrate is not just a theoretical concern; it is visible even in the successes of mechanistic interpretability. Consider the well-studied task of modular addition, where small transformers are trained to compute (a + b) mod p. Researchers have successfully reverse-engineered these models and discovered that they learn a specific, intricate algorithm based on trigonometric identities and Fourier transforms, implemented via clock-like representations in the attention heads. The specific implementation is fundamentally tied to the architectural properties of the transformer substrate. A different architecture would almost certainly learn a different algorithm, rendering this detailed explanation obsolete. Thus, even our most complete mechanistic explanations are descriptions of a substrate–algorithm pairing, not of a universal computation. [42] [29]

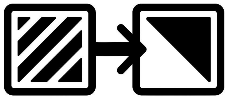
The foregoing definition of substrate is not fixed, and part of the work of MoSSAIC will be to develop this and other vectors of contingency rigorously. We invite the reader to hold this loose characterisation in mind as we present each scenario in more detail.



4.3. Scaffolding shift

Even if AI models remain unchanged from current day frontier systems, a large amount of work is conducted to ”unhobble” or otherwise enhance the abilities of existing models. This can be done by invoking models in a so-called ”agent framework” [37] with the aim of letting them achieve tasks independently [16], or offering models tools and function calls which allow them to access existing codebases [2]. In this case we can imagine the scaffolding as a substrate layer beneath that of the model itself, supplementing its operations and filling in deficiencies as an operating system does in a conventional computer. The tools provided might also directly circumvent core model deficiencies that were previously established by interpretability analysis, such as the failure of certain models to complete symbolic reasoning problems. [30]

To examine these changes through the lens of modern systems, these scaffolding implementations might plausibly elicit new capabilities previously hidden in models. For example, mechanistic interventions designed to prevent chatbot systems from creating harmful output may also fail when the LLM is removed from the guidance of a human conversation partner and told to iterate upon its own output in an agent loop.



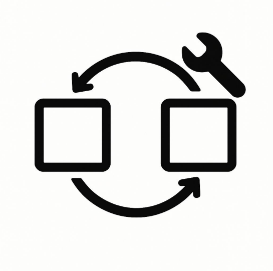
4.4. Human-initiated paradigm shifts

Most modern AI systems (i.e., before architectural variations) are underpinned by a top-level structure comprising layers of neurons connected via weighted edges with nonlinear activation functions (MLP), with ”learning” achieved via backpropagation and stochastic gradient descent.

The relative stability of this configuration has allowed mechanistic interpretability to develop as an instrumental science and to deliver findings which carry over from older to newer systems, such as circuit discovery. [12]

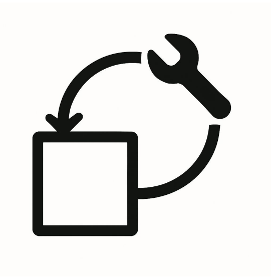
However, there is no guarantee this continuity will last: the transformer was an evolution in substrate that mixed conventional MLP layers with the attention mechanism[17], necessitating new mechanistic interpretability efforts to decode its inner workings and diminishing the value of previous work on networks such as CNNs and RNNs. Consider the difference between traditional MLP and Kolmogorov Arnold networks: the latter replace fixed activation functions on neurons with learnable functions for each edge in the network [26]. These paradigm or architecture shifts might set interpretability research back or—at worst—render it entirely obsolete, requiring new techniques to be developed from scratch.

As AI R&D accelerates, we expect new structures to emerge that are similarly incomprehensible from the perspective of someone studying a MLP network. Therefore, a deep understanding of today’s models—obtained via analysis of neural networks, transformer features, etc.—does not guarantee an understanding of future intelligences built upon different computational structures. This would be akin to trying to interpret a Python program as a pattern in Conway’s Game of Life.



4.5. AI-assisted paradigm shifts

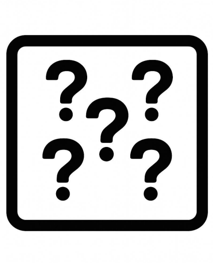
Another way we can progress from comparatively well-understood contemporary structures to less understandable ones is if we use AI to automate R&D—this is a core part of many projections for rapid scientific and technological development via advanced AI technology (e.g., PASTA[23]) [8] [40]. These changes can happen at various levels, from the hardware level (e.g. new neuromorphic chips) to the software level (e.g. new control architectures or software libraries). Furthermore, with R&D-focused problem-solving systems like o3[32], we may reach a scenario where humans are tasked with merely managing an increasingly automated and hard-to-comprehend codebase entirely produced by AI systems. Theoretical insights and efficiency improvements may be implemented entirely by AI, without regard for how easy the new architecture is for humans to interpret. This may leave interpretability researchers working with outdated models and outdated theories of how the models operate.



4.6. Self-modification

The natural extension of AI-assisted structural modification is self-modification. In contrast to the previous case, which implicitly includes a human-in-the-loop accepting any changes proposed (irrespective of whether they understand them or not), a self-modifying AI system is free to adapt itself, ostensibly in service of improving its performance in a given task or problem domain. We see a preliminary example of this behaviour with in context learning, where transformers learn the parameters of a task via one or a few iterations of the task within its context window. [1]

As AI capabilities continue to develop, an increasing pressure on AI systems to be adaptive and self-modifying seems a feasible scenario to anticipate, especially under race dynamics in which self-modifying AI systems outperform an AI-assisted human researcher/developer. The scope of self-modification could include changes on the level of architecture, training process, or even paradigm, most likely during training stages but feasibly during deployment if some form of test-time training process is adopted.

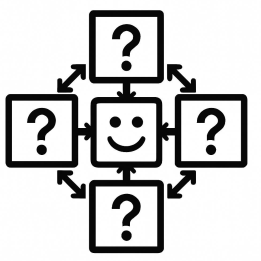


4.7. Deep Deceptiveness

It is also conceivable that mechanistic interventions—for instance, identifying neural correlates of deceptive thoughts and forcibly preventing them from developing—may be evaded by the model if the pressures to produce the original malign behaviour are not fully addressed.

This case was put forwards by Nate Soares in the article ”Deep Deceptiveness”[39]. In it, he presents a scenario in which an AI system has been trained against deceptive outputs, and any precursor to deception is shut down before it can develop. (Note that this is not explicitly a mechanistic intervention but a more general case of an advanced intelligence evading fixed targets via reconfiguration of internal processes.) In response to this restriction it develops several strategies, one of which involves translating the problem into an abstracted form such that it doesn't register as a deceptive thought process. This ability to translate the problem is not in itself deceptive and may indeed be one of the strengths of the model.

We can picture a similar thing happening with mechanistic deception monitoring and intervention techniques. Isolating the circuits that correlate to deception incentivizes the model to self-modify its sub-processes such that those deception monitors are not triggered, leaving the overall malign objective unaltered. In fact, this obfuscation is what the Bailey et al. (2024) experiments described in the introduction demonstrate, albeit in a contrived setting. Asides from pure obfuscation, we may encounter sub-processes that are individually innocuous (according to mechanistic observation methods) but that lead to unforeseen and unintended system behavior when combined together. These are examples of a deep misalignment in which the circuit-level monitoring and interventions become part of the network's loss landscape, leading to the model ”optimising away” their effects.



4.8. Aggregate Deception

The concept of aggregate deception is strongly related to prior work on robust agent-agnostic (multipolar) processes, and both can be viewed as box-inversions of the deep deceptiveness process outlined above.[3]

[3] Box-inversions show a correspondence between risk phenomena occurring inside a network (in the box) and those occurring

Aggregate deception takes 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, such that each component system contains only a benign-looking fraction of some overall deception/malicious behaviour. This massively increases the space of network components over which a search for deception must take place, further hampering mechanistic safety work.

4.9. Summary

Regardless of who implements changes in substrate, the current race dynamics strongly incentivise the development of more capable models over human-legible or human-understandable ones. This suggests that more capable models will consistently be selected for deployment over more interpretable ones, leaving AI developers who insist on producing human-legible models or retaining humans in the development cycle lagging behind in capabilities (sometimes described as paying an ”alignment tax”)[38] and at risk of being out-competed. Secrecy and competitive pressure in the development of frontier models may also incentivise AI developers to restrict access to—or even intentionally obscure—the architectures and paradigms they work with, via restrictive ”black box” APIs. In the absence of explicit regulations against this, conventional mechanistic interpretability work (and mechanistic safety work in general) will become more difficult.

5. MoSSAIC Overview

In this section we lay out a speculative proposal for how we might address the above problems with the reductionist/mechanistic paradigm.

5.1. Core Motivation

We can characterize the more general problem, inherent in the Causal–Mechanistic paradigm, in terms of substrate-dependent and substrate-independent approaches to alignment.

Many of the concerns we raise above center on the problem of moving and fixed targets: Mechanistic research is built on analysing fixed, legible substrate targets, whilst much of the problem posed by new models is that they present a moving target of novel substrate configurations. The results/insights obtained under the causal–mechanistic paradigm are tied too closely to a particular substrate instantiation of AI. Thus, they may fail to generalize to new substrates, and their safety assurances may be weakened.

The problem of generalizing beyond any one particular substrate—be that model, architecture, or paradigm—has already been noted. The main solution, typical of the early MIRI work, is the so-called agent foundational or theoretical “top-down” approach. This approach focuses on laying a theoretical foundation for how advanced AI systems will behave, using fields like decision theory[41], game theory, and singular learning theory[20]. The goal is often to capture the forms and behaviours an artificial manifestation of intelligence will exhibit under optimal conditions (sometimes called working on a problem in the worst case [11]) [10], with the assumption that AI systems which are suboptimal but superhuman will optimise themselves towards this end state. AIXI is a key example of such an approach.

These abstractions and idealised concepts of agency and decision-making should remain stable through successive generations of SOTA models and architectures. AIXI represents a mathematically perfect or optimal reinforcement learner, so any advance in reinforcement learning should resemble more and more AIXI’s core learning process, regardless of the substrate it’s realized in. These substrate-independent insights should therefore generalize. Unfortunately, these insights show limited applicability to the diversity of real-world AI development/deployment scenarios. It is naturally difficult to perform experiments to verify these hypotheses in existing systems, and harder still to empirically verify the theory behind them.

We view these two relationships to substrate as defining the two dominant research proxies [25] historically in AI safety research, corresponding loosely to the prosaic and agent-foundational camps. The prosaic camp uses insights typically derived via inductive methods to arrive at conclusions regarding current systems, thereby risking overfitting to those systems and the substrates that realize them. The agent-foundational camp proceeds more via deduction, arriving at mathematical proofs of idealized agency which apply generally but fail to apply to any existing instances of AI systems.[4]

That is, we must meet the challenge of substrate-dependance

across a wider infrastructure of connected AI systems (outside the box), arguing that the two are the instances of the same process.

[4] This is not to say that there is no work being performed between the two approaches. In brief, we view PIBBSS and the singular learning theory and causal incentives research agendas as failing to fall directly into one or the other approach.

by becoming more general without dichotomizing it with generalization that neglects substrate-specifics altogether. We contend that this might be made possible by investigating the power of generalization (in say, a "general result") and vastly re-imagining the possible solutions. We outline a specific proposal to this effect in the next section ("live theory"), as our tools and research infrastructures themselves are pervaded by moderate intelligence.

5.2. Live Theory

The research tools we propose are built upon an approach we call "live theory". Here we summarize the way in which live theory might help us meet the challenge of adequately characterizing intelligence in a generalizable way without trading off the specifics.

To generalize is to transfer some insight between contexts. For this, we typically extract that which is stable or invariant across multiple contexts and port that to the novel situation. The generalised product we produce is typically a theory; it captures the core insights that can be fitted to new contexts via abstraction and parametrization.[35] For example, Shannon formalized his intuitions regarding information into a set of equations that apply to multiple different domains. His equations provide an invariant structure that applies to handwriting, morse code, the Internet, telephone communications, biological brains, and numerous other substrates.

However, routing our insights through a single conceptual core as part of our theoretical practice requires us to overlook, or "abstract away from", local contextual features that are often significant, especially in contexts involving a lot of complexity. In practice, this re-insertion of context is done by intelligent humans. Shannon's insights must be married to an understanding of the specifics of neurons and their metabolic constraints if his insights are to be applied to information processing in biological brains.[15]

We believe that the causal–mechanistic paradigm we describe above is an example of a set of research assumptions and practices that depend heavily on local context. It is fine-grained, substrate-specific, and may fail to generalize as a single conceptual core. The typical "conceptual core" approach involves abstracted or high-level claims and can be overly coarse-grained, substrate-independent or model agnostic, and may fail to reflect the local evolving details, which is a vector of risks, as discussed in the first half of this paper.

We believe that the dichotomy between the two approaches is a byproduct of limitations in research infrastructure (and subsequent culture of methodology) rather than a necessary divide in the field.

We aim to look slightly ahead, where AI presents the opportunity to transcend this dichotomy at the level of outputs and artefacts. We expect moderately intelligent AI infrastructure to be able to support new kinds of conceptual transformations that do not rely on shared core formalisms in order to be robust across disciplines, mechanisms, substrates, paradigms etc.. This is a dynamic form of robustness, which aims to create dynamic artefacts that transmit novel insights reliably while remaining adaptive to local context.

This is possible in a time of moderate-intelligence, where AI technology is not too dangerous or agentic, but the cost and latency become extremely low. Humans in such a middle-period regime might produce partially complete mathematical theory prompts or contexts, rather than finished formalisms. The incompleteness is a feature rather than a bug, since it allows the specifics to weigh in on the mathematical structure itself, rather than as just parametric instantiations of a prefigured general mathematical structure.

5.3. AI assistance

As a contemporary example of what we mean, consider the adaptability that an LLM provides. You create a local context based on the task you want to achieve and the interface adapts its responses to suit the situation. For example, it can explain the same mathematical concepts at various intellectual levels, translate ideas between languages, and create examples to demonstrate theoretical intuitions. This interactivity goes beyond simply taking a formula and inputting context via parametrization.

We expect live theories to reflect this adaptability made possible by AI tools and interfaces, and we argue that substrate-flexible intelligence will require us to exploit this adaptivity to novel contexts when porting insights regarding "high-level" constructs such as deception, honesty, and so on. Therefore, we will be able to transform previous discoveries in mechanistic research (e.g. mechanistic interpretability techniques or intuitions) to fit new domains and substrates in a way that is more flexible than traditional mathematical formulae. Humans will supply partial formalisms that include a mix of formal and informal data, that undergo "smart substitution" using infrastructural AI, replacing the usual theoretical process of formal abstraction and substitution that is the basis of existing science. The details of this post-parametric generalization are beyond the scope of this paper, but articulated elsewhere.[36]

To be clear, live theory is still undergoing prototyping, but we view it as a plausible near-term outcome of progress currently being made, and we want to leverage these near-term tooling capabilities to counter the challenges of the medium-term developments we detail in this paper.

5.4. What MoSSAIC is not

We should stress a few things that we view live theory as separate from. Firstly, it is not about "autonomous AI scientists." The use of AI is backgrounded as a tool to assist with the communication of intuitions, ultimately based on intuitions and insights that come from humans. Secondly, we do not view our work as either high or low-level theorizing, but a way of generalizing theoretical insight that is transferable across different contexts, without finding a static "core thesis" that simplifies them. It is neither independent of substrate nor entirely substrate-dependent (and hence not model "agnostic"). We believe that the core steering role humans play makes this theory of change more subtle than the idea of simply "using AI to solve AI safety.". Instead, MoSSAIC looks to develop frameworks, tools, and intelligent infrastructure for porting human insights between different contexts, which we claim is the truer desideratum underlying "general" methods, or oversight that "composes" or "scales". We will still as researchers need to lay out seed ideas (such as 'deception' or 'power-seeking') and their guidance, without being restricted only to formal invariants.

5.5. Commitment

Rather than treating intelligent constructs such as deceptiveness or honesty as behaviours and meaningful phenomena in and of themselves, the mechanistic paradigm aims to explain them in terms of underlying neural mechanisms. As a result, it cannot respond to the inherent flexibility that intelligence has with respect to its substrates. As part of our work on MoSSAIC, we are attempting to formalize intuitions we feel to be relevant to deception in a flexible way. This work serves both as a demonstration of a live theoretic approach to tackling deception across substrates, and also an attempt to define a framework that might provide a cohesive articulation of the pessimistic, agent foundations problem cases mentioned earlier.

A range of terms are used in the literature to describe ideas associated with goal-directedness, each with its own connotations and implied structure. These terms include goal, objective, aim, intention, or drive. We choose to use the word "commitment" because we hope to capture something more general than the above terms. Commitments can vary from very basic commitments of self-preservation and physical integrity; to more complex ones such as parent–offspring relationships; to the intricate commitments social animals form in packs or groups. Commitments can include goals or objectives, but need not be actively optimized for or even consciously acknowledged in many cases. For instance, my emotional commitment to a sibling in another part of the world only becomes part of my conscious planning in some specific cases. Commitments can also include the more biological notions of intentions and drives, or the possibility of structured plans and in shared commitments between intelligences. Commitments as we define them need not be voluntary, the product of a deliberation process, or locally/globally optimal: a parent may sacrifice their life to save their child from a burning building without consciously weighing up their options. In short, we want to capture some notion of "for-the-sake-of" that guides action/behaviour and is robust across environmental changes and substrate alterations.

Crucially, we will aim to formalize "commitments" in a way that is substrate-sensitive rather than substrate-independent. This is because the nature of the substrate will dramatically affect the nature of the associated commitments—there is no static, universally applicable account of commitment.

6. Conclusion

We believe that the above lays out a picture of why we are interested in working on MoSSAIC, as well as some approaches we find promising. We have already begun preliminary work formalizing commitments through the lens of category theory, which provides a lightweight, domain-agnostic mathematical structure that will help us conceptualise networks of commitment as the dual of causal DAGs. We hope to continue this work both on the mathematical/formal and philosophical aspects and welcome engagement with the ideas we have raised in this preliminary outline.

7. Author Contributions

Sahil K generated the initial ideas and supervised the project. Matt Farr developed the arguments at MATS 6.0 and cowrote the initial and current drafts of the paper with Chris Pang and Aditya Arpitha Prasad, respectively. Chris developed many of the examples and connections to computer science, Aditya helped shape the ML aspects of this paper.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. URL https://arxiv.org/abs/2211.15661.

[2] Anthropic. Build with claude: Computer use (beta), 2024. URL https://docs.anthropic.com/en/docs/build-with-claude/computer-use. Accessed on 12/19/2024.

[3] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by

a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

[4] Kola Ayonrinde and Louis Jaburi. A mathematical philosophy of explanations in mechanistic interpretability – the strange science part i.i, 2025. URL https://arxiv.org/abs/2505.00808.

[5] Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses, 2025. URL https://arxiv.org/abs/2412.09565.

[6] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112.

[7] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. URL https://arxiv.org/abs/2404.14082.

[8] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014. ISBN 0199678111.

[9] Greg; Chen, Crystal; Novick and Kirk Shimano. Risc architecture, 2000. URL https://cs.stanford.edu/people/eroberts/courses/soco/projects/risc/risccisc/. Accessed on 12/10/2024.

[10] Ajeya; Christiano, Paul; Cotra and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you, 2021. URL https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit. Accessed on 12/16/2024.

[11] Paul Christiano. My research methodology, 2021. URL https://ai-alignment.com/my-research-methodology-b94f2751cb2c. Accessed on 12/20/2024.

[12] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

[13] NVIDIA Corporation. Uda toolkit, 2024. URL https://developer.nvidia.com/cuda-toolkit. Accessed on 12/10/2024.

[14] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

[15] Lori Dajose. Thinking slowly: The paradoxical slowness of human behavior, 2024. URL https://www.caltech.edu/about/news/thinking-slowly-the-paradoxical-slowness-of-hum Accessed on 12/4/2024.

[16] Google DeepMind. Project astra, 2024. URL https://deepmind.google/technologies/project-astra/. Accessed on 12/19/2024.

[17] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

[19] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2025. URL https://arxiv.org/abs/2301.04709.

[20] Jesse Hoogland. Neural networks generalize because of this one weird trick, 2023. URL https://www.lesswrong.com/posts/fovfuFdpuEwQzJu2w/neural-networks-generalize-because-of-this-one- Accessed on 12/23/2024.

[21] Sara Hooker. The hardware lottery, 2020. URL https://arxiv.org/abs/2009.06489.

[22] Intel Corporation. Intel® core™ processor family, 2024. URL https://www.intel.com/content/www/us/en/products/details/processors/core.html. Accessed on 12/12/2024.

[23] Holden Karnofsky. Forecasting transformative ai, part 1: What kind of ai?, 2021. URL https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/. Accessed on 12/13/2024.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[25] Jan Kulveit. Hierarchical agency: A missing piece in ai alignment, 2024. URL https://www.alignmentforum.org/posts/xud7Mti9jS4tbWqQE/hierarchical-agency-a-missing-piece-in-ai-alignment. Accessed on 1/7/2024.

[26] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks, 2024. URL https://arxiv.org/abs/2404.19756.

[27] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982. ISBN 0716715678.

[28] Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023. URL https://arxiv.org/abs/2307.15771.

[29] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

[30] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models, 2024. URL https://arxiv.org/abs/2406.02061.

[31] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits, 2020. URL https://distill.pub/2020/circuits/zoom-in/. Accessed on 25/6/2025.

[32] OpenAI. 12 days of openai, 2024. URL https://openai.com/12-days/?day=12. Accessed on 12/19/2024.

[33] Open Philanthropy. Tais rfp: Research areas, 2025. URL https://www.openphilanthropy.org/tais-rfp-research-areas/.

[34] Paul Rendell. This is a universal turing machine (utm) implemented in conway's game of life., 2010. URL http://rendell-attic.org/gol/utm/index.htm. Accessed on 12/10/2024.

[35] Sahil. The logistics of distribution of meaning: Against epistemic bureaucratization, 2024. URL https://www.lesswrong.com/posts/MhBRGfTRJKtjc44eJ/the-logistics-of-distribution-of-meaning-agains Accessed on 12/4/2024.

[36] Sahil. Sequence: Live theory, 2024. URL https://www.lesswrong.com/s/aMz2JMvgXrLBkq4h3. Accessed on 12/4/2024.

[37] Barry Schluntz, Erik; Zhang. Building effective agents, 2024. URL https://www.anthropic.com/research/building-effective-agents. Accessed on 12/19/2024.

[38] Rohin Shah. [an #95]: A framework for thinking about how to make ai go well, 2020. URL https://www.lesswrong.com/posts/9et86yPRk6RinJNt3/an-95-a-framework-for-thinking-about-how-to-mak Accessed on 12/16/2024.

[39] Nate Soares. Deep deceptiveness, 2023. URL https://www.alignmentforum.org/posts/XWwvwytieLtEWaFJX/deep-deceptiveness. Accessed on 12/19/2024.

[40] Roman V Yampolskiy. From seed ai to technological singularity via recursively self-improving software. *arXiv preprint arXiv:1502.06512*, 2015.

[41] Eliezer Yudkowsky and Nate Soares. Functional decision theory: A new theory of instrumental rationality, 2018. URL https://arxiv.org/abs/1710.05060.

[42] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks, 2023. URL https://arxiv.org/abs/2306.17844.

[43] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.