Open Problems in Mechanistic Interpretability

Anonymous authors
Paper under double-blind review

Abstract

Mechanistic interpretability aims to understand the computational mechanisms underlying neural networks' capabilities in order to accomplish concrete scientific and engineering goals. Progress in this field thus promises to provide greater assurance over AI system behavior and shed light on exciting scientific questions about the nature of intelligence. Despite recent progress toward these goals, there are many open problems in the field that require solutions before many scientific and practical benefits can be realized: Our methods require both conceptual and practical improvements to reveal deeper insights; we must figure out how best to apply our methods in pursuit of specific goals; and the field must grapple with socio-technical challenges that influence and are influenced by our work. This forward-facing review discusses the current frontier of mechanistic interpretability and the open problems that the field may benefit from prioritizing.

1 Introduction

Recent progress in artificial intelligence (AI) has resulted in rapidly improved AI capabilities. These capabilities are not designed by humans. Instead, they are learned by deep neural networks (Hinton et al., 2006; LeCun et al., 2015). Developers only need to design the training process; they do not need to – and in almost all cases, do not – understand the neural mechanisms underlying the capabilities learned by an AI system.

Although human understanding of these mechanisms is not necessary for AI capabilities, understanding them would enhance several *human* abilities. For example, it would permit better human control over AI behavior and better monitoring during deployment. It would also facilitate trust in AI systems, allowing us to fully realize their potential benefits by enabling their deployment in safety-critical and ethically-sensitive settings.

Beyond the engineering benefits, understanding AI systems offers immense scientific opportunities. For the first time in history, we can create and study artificial minds with a level of access and control that is simply not possible in biological systems. What new laws of nature governing the mechanisms of minds might we discover from studying the internal workings of AI systems?

The scientific opportunities are not limited to the field of AI. If an AI can outperform tools designed by humans in a given scientific field, it suggests that the AI system is representing something about the world currently unknown to us. We can develop a deeper comprehension of the world by understanding those representations. What can we learn about protein folding from AIs that can successfully predict protein structure? What insights can we glean about disease from a radiographer that performs beyond human ability?

Mechanistic interpretability might unlock these benefits. This field of study aims to understand neural networks' decision-making processes. Here, we define "Understanding a neural network's decision-making process" as the ability to use knowledge about the mechanisms underlying a network's decision-making process in order to successfully predict its behavior (even on arbitrary inputs) or to accomplish other practical goals with respect to the network. Such goals might include more precise control of the network's behavior, or improved network design. Interpretability promises greater assurances for AI systems through a better understanding of what neural networks have learned, thus enabling us to realize their potential benefits.

1.1 The focus of this review: Open problems and the future of mechanistic interpretability

Several recent reviews of mechanistic interpretability research and related topics exist (Rauker et al., 2023; Geiger et al., 2022; Bereska & Gavves, 2024; Ferrando et al., 2024; Rai et al., 2024; Anwar et al., 2024; Davies & Khakzar, 2024; Mosbach et al., 2024; Mueller et al., 2024). Our review takes a more forward-looking stance. We discuss not only where the frontier is today, but also which directions we might benefit most from prioritizing in the future.

1.1.1 Why 'mechanistic' interpretability?

The distinction between interpretability and mechanistic interpretability is not always clear and is therefore worth clarifying. The motivations and methods used in interpretability work are often diverse (Lipton, 2018; Doshi-Velez & Kim, 2017; Jacovi, 2023). As a result, there are many ways in which interpretability research might be categorized. Prior categorizations of interpretability include causal vs. correlational methods, supervised vs. unsupervised methods, bottom-up vs. top-down methods, among others (Geiger et al., 2024b; Mueller et al., 2024; Bereska & Gavves, 2024; Belinkov, 2022a; Zou et al., 2023a; Davies & Khakzar, 2024). This review focuses specifically on *mechanistic* interpretability. But what distinguishes 'mechanistic interpretability' from interpretability in general? It has been noted that the term is used in a number of (sometimes inconsistent) ways (Saphra & Wiegreffe, 2024). In this review, we use the term 'mechanistic interpretability' in a technical sense, referring specifically to work that investigates the mechanisms underlying neural network generalization. Mechanistic interpretability represents one of three threads of interpretability research, each with distinct but sometimes overlapping motivations, which roughly reflects the changing aims of interpretability work over time.

The first thread aims to build AI systems that are interpretable by design. Much early interpretability work focused on explaining the sensitivity of machine learning models to inputs and training data. This work typically used small-to-medium sized models designed to be easily interpretable, such as decision trees (Breiman, 1984; Hu et al., 2019), linear models (Roweis & Ghahramani, 1999), and generalized additive models (Hastie & Tibshirani, 1986; Agarwal et al., 2021). These models could be used alongside attribution methods such as influence functions (Hampel, 1974; Koh & Liang, 2017) and Shapley values (Shapley, 1997; Lundberg & Lee, 2017), which were common techniques used to characterize model decision boundaries with respect to inputs. "Interpretability by design" continues to be an active research area, including architectures such as Concept-Bottleneck Models (Koh et al., 2020), Backpack Language Models (Hewitt et al., 2023), Kolmogorov-Arnold Networks (Liu et al., 2024c), and sparse decision trees (Xin et al., 2022).

With the rise of larger-scale nonlinear neural networks (Krizhevsky et al., 2012; He et al., 2016), another thread grew in importance, driven primarily by the question: Why did my model make this particular decision? However, one challenge in interpreting larger networks was finding attribution methods that could scale to large networks (Zeiler & Fergus, 2014). In response, a number of local attribution methods were developed, including grad-CAM (Selvaraju et al., 2019), integrated gradients (Sundararajan et al., 2017), and masking-based causal attribution (Fong & Vedaldi, 2017), SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), and many other methods, including backprop-based visualization methods (Simonyan et al., 2014a; Nguyen et al., 2016b).

Inspired by, for example, Inception (Szegedy et al., 2015) and GPT-3 (Brown et al., 2020), another thread emerged as models became capable of more profound generalization. Focused on the broader subject of generalization, it was driven by the question: How did my model solve this general class of problems? Due to its emphasis on the mechanisms underlying neural network generalization, the work in this category is commonly referred to as 'mechanistic interpretability' (in addition to other, more cultural reasons, according to (Saphra & Wiegreffe, 2024)). This kind of interpretability work is driven by a fundamental hypothesis in deep learning that generalization arises from shared computation (LeCun et al., 2015). Early work in this area, such as feature visualization (Olah et al., 2017b), or network dissection (Bau et al., 2020), sought "global" explanations for model generalization by investigating the roles of model components across a class of decisions. More recent work in this area looks at "circuits" of components (Wang et al., 2023), generalizable patterns of information flow (Geva et al., 2023), representation subspaces (Geiger et al., 2024c; Zou et al., 2023a) and probes (or self-supervised searches via sparse dictionary learning) for representation vectors that

carry information that generalizes across many instances for a particular task or set of tasks (Huben et al., 2024; Bricken, 2023; Todd et al., 2024; Hollinsworth et al., 2024).

1.2 Types of open problems

The field of mechanistic interpretability ultimately aims to achieve concrete scientific and engineering goals. For instance, we would like to be able to:

- Monitor AI systems for signs of cognition related to dangerous behavior (Section 3.2);
- Modify internal mechanisms and edit model parameters to adapt their behavior to better suit our needs (Section 3.2.2);
- Predict how models will act in unseen situations or predict when a model might learn specific abilities (Section 3.3);
- Improve model inference, training, and mechanisms to better suit our preferences (Section 3.4);
- Extract latent knowledge from AI systems so we can better model the world (Section 3.5).

Despite recent hopeful signs of progress, mechanistic interpretability still has considerable distance to cover before achieving satisfactory progress toward most of its scientific and engineering goals.

To achieve these goals, the field not only needs greater application of current state-of-the-art mechanistic interpretability methods, but also requires the development of improved techniques. The first major section (Section 2) therefore discusses open problems related to the methods and foundations of mechanistic interpretability.

We then explore key axes of research progress that will determine how far we can advance toward the goals of mechanistic interpretability (Section 3.1). Section 3 outlines how applications of interpretability methods have made progress toward the field's goals, and, for each goal, discusses the specific axes along which progress is likely needed to achieve specific objectives.

Finally, we note that the goals, applications, and methods of mechanistic interpretability do not exist in a vacuum. Like any scientific field, they lie within a broader societal context. The final section of this review examines open socio-technical problems in mechanistic interpretability (Section 4). It discusses current initiatives and possible pathways to translate technical progress into levers for AI governance, alongside consequential social and philosophical challenges faced by the field.

2 Open problems in mechanistic interpretability methods and foundations

One way of thinking about what neural networks do internally is that they learn parameters that implement neural algorithms. These neural algorithms take data as input and, through a series of steps, transform their internal activations to produce an output. Different parts of the network learn different steps in the algorithm; mechanistic interpretability aims to describe the function of different network components.

Broadly speaking, there are two methodological approaches to achieving this. The first approach, often called 'reverse engineering', is to decompose the network into components and then attempt to identify the function of those components (Section 2.1). This approach "identifies the roles of network components". Conversely, the second approach, sometimes referred to as 'concept-based interpretability', proposes a set of concepts that might be used by the network and then looks for components that appear to correspond to those concepts (Section 2.2). This approach thus "identifies network components for given roles".

In this section, we will examine both approaches ('Reverse engineering' and 'Concept-based interpretability') (Figure 1), discussing their methods and open problems. We will also touch on open problems that cut across either approach, including proceduralizing the mechanistic interpretability pipeline (Section 2.3) and its uses in automating interpretability research (Section 2.4).

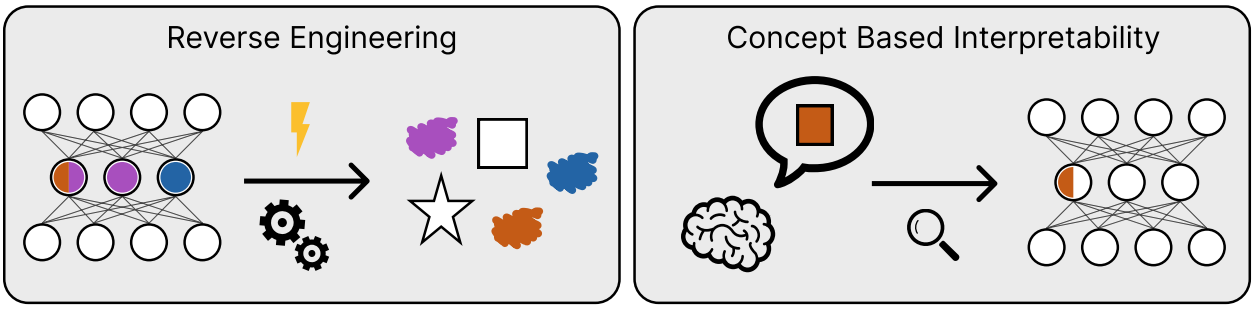

Figure 1: Two approaches to neural network interpretability. (Left) Reverse Engineering is characterized by decomposing networks into functional components and describing how those components interact to produce the network's behavior. It thus aims to 'identify the roles of network components' (Section 2.1). (Right) Concept-based interpretability on the other hand attempts to discover human concepts within neural network internals. It thus aims to 'identify the network components for given roles' (Section 2.2).

2.1 Reverse engineering: Identifying the roles of network components

2.1.1 Reverse engineering is necessary because AI and humans use different representations and perform different tasks

Large language models produce text that closely resembles human writing, and it is tempting to assume that they generate it through cognitive processes similar to those of human writers[1]. However, humans and AI models often solve problems in different ways. For example, a model that is only 1% the size of GPT-3 outperforms humans on next-token prediction tasks (Shlegeris et al., 2024). Inversely, even state-of-the-art multimodal LLMs struggle with tasks that a four-year-old could easily master, such as learning causal properties of new objects involving simple lights and shapes (Kosoy et al., 2023). An even clearer sign that humans and AI are using different representational processes are cases where humans cannot solve a problem at all, as in the case of predicting a protein structure from sequence (Jumper et al., 2021; Lin et al., 2023).

Even when both humans and AI exhibit comparable levels of competence on a given task, they may use different heuristics. For example, research shows that image models tend to rely more heavily on textural features (such as recognizing elephants by their hide rather than their shape (Geirhos et al., 2019)) or rely on dataset correlations (as when identifying fish by the fingers that proud fisherman use to hold them (Brendel & Bethge, 2019)) to a greater degree than people do. Even simple algorithmic tasks, like modular addition, which humans might solve with simple carries, were solved by a small transformer model by learning a Fourier transform strategy that researchers only understood in retrospect (Nanda et al., 2023a).

To grasp the potentially alien cognition of these models, we must develop methods to uncover and understand the previously unknown concepts and mechanisms implemented within them. In other words, we must be able to reverse engineer these models (Olah, 2024).

Reverse engineering generally involves three steps, whether it is an engine, a piece of software, or a neural network (Figure 2):

1. **Decomposition**: Breaking down the object of study into simpler parts (Section 2.1.2);
2. **Description of components**: Formulating hypotheses about the functional role of component parts and how they interact (a process that can be called 'interpretation') (Section 2.1.3);
3. **Validation of descriptions**: Testing if our hypotheses are correct (Section 2.1.4).

[1]However, even if this assumption were true, understanding LLMs would remain challenging, as we also lack a mechanistic understanding of the cognitive processes involved in human text creation!

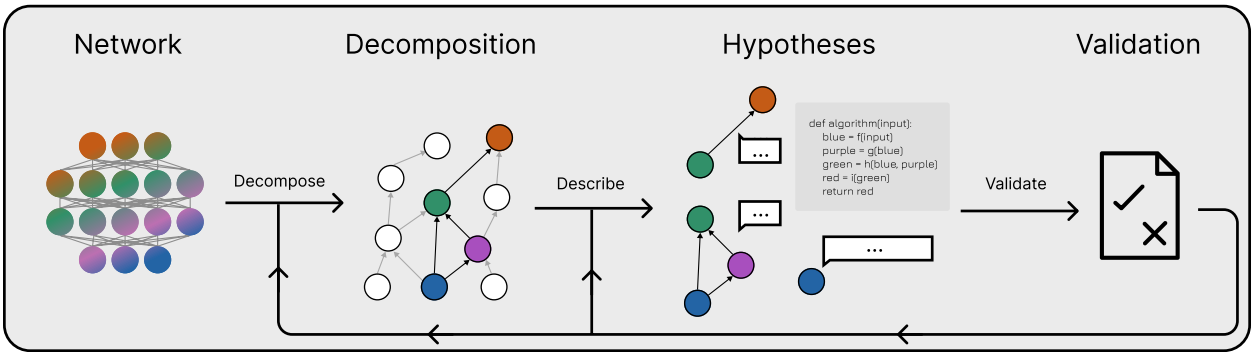

Figure 2: The steps of reverse engineering neural networks. (1) Decomposing a network into simpler components. This decomposition might not necessarily use architecturally-defined bases, such as individual neurons or layers (Section 2.1.2). (2) Hypothesizing about the functional roles of some or all components (Section 2.1.3). (3) Validating whether our hypotheses are correct, creating a cycle in which we iteratively refine our decompositions and hypotheses to improve our understanding of the network (Section 2.1.4).

If our hypotheses are invalidated, we must either improve our decomposition or improve our hypotheses regarding the functional roles of its components. We will systematically examine each step, analyzing current methods and their respective shortcomings.

2.1.2 Reverse engineering step 1: Neural network decomposition

In mechanistic interpretability, our aim is to decompose a neural network and study its parts in isolation in order to explain how neural networks generalize. This aim raises the question of how best to carve a neural network "at its joints" for the purposes of interpretability.

Networks do not naturally decompose into architectural components. The naive approach to decompose neural networks involves breaking them down into their architectural components, such as individual neurons, attention heads, or layers.

Some early attempts to interpret deep artificial neural networks studied the responses or weight structure of individual neurons or single convolutional filters (Erhan et al., 2009; Le et al., 2012; Krizhevsky et al., 2012; Szegedy et al., 2014; Simonyan et al., 2014b; Zhou et al., 2015; Srivastava et al., 2014; Yosinski et al., 2015; Mordvintsev, 2015; Olah et al., 2017a; Bau et al., 2020; Dalvi et al., 2019; Olah et al., 2020a; Cammarata et al., 2020). These efforts paid homage to the 'Neuron Doctrine' in neuroscience, which posits that individual neurons are the structural and functional unit of the nervous system (Cajal, 1924; Sherrington, 1906). However, researchers discovered that single neurons are 'polysemantic' – they seem to respond to multiple kinds of features in both artificial (Wei et al., 2015; Nguyen et al., 2016c; Olah et al., 2017a) and biological networks (Churchland & Shenoy, 2007; Rigotti et al., 2013; Mante et al., 2013; Raposo et al., 2014). These observations support earlier theoretical work that suggested representations used by neural networks do not necessarily align with the activation of individual neurons (Hinton, 1981).

Interpreting individual attention heads does not fare better than interpreting individual neurons, as attention heads also exhibit polysemanticity (Janiak et al., 2023). More broadly, research suggests that studying the attention patterns of models can often be misleading (Jain & Wallace, 2019; Pruthi et al., 2020).

Some work even suggests that representations in language models might span multiple layers (Yun et al., 2021; Lindsey et al., 2024). This chimes with work that edits or intervenes on individual layers, which indicates that this level is too coarse-grained to robustly carve the network at its joints (Meng et al., 2022b; Wang et al., 2023).

If natural architectural components, such as individual neurons, attention heads, or layers, do not provide a natural way to decompose neural network representations, then what does?

Decomposition by dimensionality reduction methods. If individual neurons are not the right decomposition, perhaps groups or patterns of neurons are. Many decomposition methods attempt to identify activation vectors that correspond to the basic unit of neural network computation. One common approach is to provide models with a range of unlabeled inputs, collect the resulting hidden activations, and then apply unsupervised dimensionality reduction techniques to these hidden activations. The hope is that structure in the hidden activations corresponds to the structure of neural computation. Commonly used dimensionality reduction methods include Principal Component Analysis or Singular Value Decomposition (Hollinsworth et al., 2024; Marks & Tegmark, 2024; Huang et al., 2024a; Bushnaq et al., 2024) and non-negative matrix factorization (Olah et al., 2018; Voss et al., 2021; Cammarata et al., 2020), though these techniques are no longer predominant methods used for mechanistically decomposing language models (Friedman et al., 2024).

Decomposition by sparse dictionary learning (SDL). According to the 'superposition hypothesis', neural networks are capable of representing more features than they have dimensions, as long as each feature activates sparsely (Elhage et al., 2021) (Figure 3). This is a key reason that dimensionality reduction methods are not considered state-of-the-art, because they cannot identify more directions than there are activation dimensions. The superposition hypothesis, coupled with the failure of dimensionality reduction methods to overcome it, motivated the search for methods that can identify more directions than dimensions. Recent work has explored the use of sparse dictionary learning (SDL) to this end (Elhage et al., 2021; Sharkey et al., 2022b; Huben et al., 2024; Bricken, 2023).

Currently, SDL is the most popular set of unsupervised decomposition methods in mechanistic interpretability. SDL encapsulates a family of methods, including Sparse Autoencoders (SAEs) (Gao et al., 2024; Templeton et al., 2024; Rajamanoharan et al., 2024; Makelov et al., 2024; Kissane et al., 2024b; Braun et al., 2024), Transcoders (Dunefsky et al., 2024), and Crosscoders (Lindsey et al., 2024).

In SDL, hidden activations are typically passed to a small neural network consisting of only two layers, which correspond to an encoder and decoder respectively, with a wide hidden space. The encoder activations represent how active each 'latent' [2] is, and the decoder matrix corresponds to a dictionary of latent directions. We want to train the dictionary elements to align with 'feature directions' in the network's hidden activations. Since we assume that individual features are sparsely present in the activations, the encoder activations are trained to be sparsely activating. In the case of SAEs, the output is trained to either reconstruct the input or, in the case of transcoders (Dunefsky et al., 2024), to reconstruct the activations of the next layer. Crosscoders (Lindsey et al., 2024) permit a wider class of inputs and outputs, potentially reconstructing the activations of many layers simultaneously. Since the encoder is nonlinear, it is thought to be able to learn to activate a latent only if a feature is 'active' in the hidden activations and remain 'off' if it is not.

Although SDL is considered a leading decomposition method for mechanistic interpretability, it has substantial practical and conceptual limitations (Figure 4).

SDL reconstruction errors are too high: Large errors in SDL reconstruction raise the question whether SDL methods can reconstruct the hidden representation sufficiently well such that the latents learned by SDL are adequately faithful to the models being interpreted. To measure this, the model's true hidden activations can be replaced with sparse dictionary reconstructions, then subsequently evaluating the extent to which the model's performance decreases. In practice, the error results in significant performance reductions. When a sparse dictionary with 16 million latents was inserted into GPT-4, the language modeling loss was equivalent to a model with only 10% of GPT-4's pretraining compute (Gao et al., 2024). Similarly, Makelov et al. (2024) found that using reconstructions from sparse autoencoders decreased GPT-2 small performance by 10% when trained on task-specific data, and 40% when trained on the full distribution.

Making sparse dictionaries much larger and sparser to reduce errors is a feasible but computationally expensive approach. Furthermore, in the limit, this results in merely assigning one dictionary latent per datapoint, which is clearly less interpretable. One partial solution is using SDL methods with 'error nodes' (Marks et al., 2024), to account for the discrepancies between the original and reconstructed activations. However, while the sparse autoencoders identified interpretable latents, the error nodes contain 'everything else', making

[2]The term latent is often used instead of the word feature to refer to SDL dictionary elements, since the term 'feature' is often used to refer to multiple different ideas (Smith, 2024b)

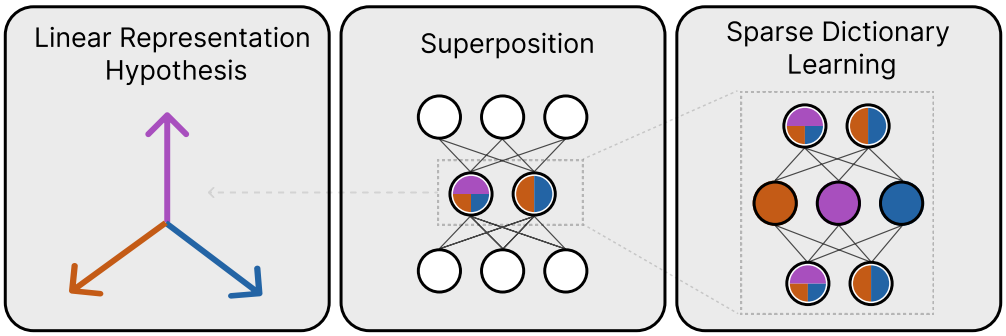

Figure 3: Three ideas underlying the sparse dictionary learning (SDL) paradigm in mechanistic interpretability. (Left) The linear representation hypothesis states that the map from 'concepts' to neural activations is linear. (Middle) Superposition is the hypothesis that models represent many more concepts than they have dimensions by representing them both sparsely and linearly in activation spaces. (Right) SDL attempts to recover an overcomplete basis of concepts represented in superposition in activation space.

them an inadequate solution to the problem. Engels et al. (2024b) found that these reconstruction errors are not purely random, as much of the direction of the error and its norm can be linearly predicted from the initial activation vector. This suggests that current SDL methods may systematically fail to capture certain structured aspects of model representations, but also implies potential solutions. To overcome this issue, it may be necessary to improve SDL training methods, or develop entirely new methods of network decomposition.

SDL methods are expensive to apply to large models: SDL involves training a small neural network for every layer of the AI model that we want to interpret. Typically, the sparse dictionaries have more parameters at that layer than the original model does, and consequently will probably be relatively expensive to train compared to the original model. [3] As AI models become larger, scaling costs of SDL also increase, although it remains unclear whether relative scaling costs are sub- or supra-linear. For cost effectiveness, it may be important to develop intrinsically decomposable methods for training models, while remaining at (or close to) state-of-the-art performance (Section 2.1.2). This approach will help avoid incurring the expense of both training and decomposing AI models.

SDL assumes the linear representation hypothesis in nonlinear models: Other problems with SDL arise from the assumptions on which SDL is based. One such assumption is the linear representation hypothesis. The linear representation hypothesis observes that though neural networks are nonlinear functions and could potentially use highly nonlinear representations (Black et al., 2022; Engels et al., 2024b; Kirch et al., 2024), they tend to use representations that exhibit strikingly linear behavior (Smolensky, 1986; Mikolov et al., 2013; Olah et al., 2020b; Elhage et al., 2021; Park et al., 2023a; Guerner et al., 2024; Gurnee & Tegmark, 2024). The hypothesis formalizes this phenomenon by stating that high level concepts are linearly represented as directions (vectors) in neural network embeddings. Its two core claims are (i) that the composition of multiple concepts can be represented as the addition of their corresponding feature vectors, and (ii) that the intensity of a concept is represented by the scale of its corresponding feature vector (Olah & Jermyn, 2024). Earlier work (Elhage et al., 2021) defined the linear representation hypothesis with the additional assumption of one-dimensional features, but this definition has since been refuted (Yedidia, 2023; Chughtai & Lau, 2024; Engels et al., 2024a) and clarified (Olah & Jermyn, 2024). Another recently proposed criterion is that the intensity of concepts should be retrievable by a linear function of the embeddings, up to a small error (Hänni et al., 2024; Olah & Jermyn, 2024).

A weak version of the linear representation hypothesis states that some concepts are linearly represented, while a strong version may assert that all concepts are (Smith, 2024a). Some works have shown that the

[3] The actual relative cost is unclear since there are no public attempts to apply SDL to every vector space in a model, although some work applies SDL to various layers (Marks et al., 2024; Gao et al., 2024; Braun et al., 2024; Lieberum et al., 2024; Bloom & Lin, 2024a; Huben et al., 2024)

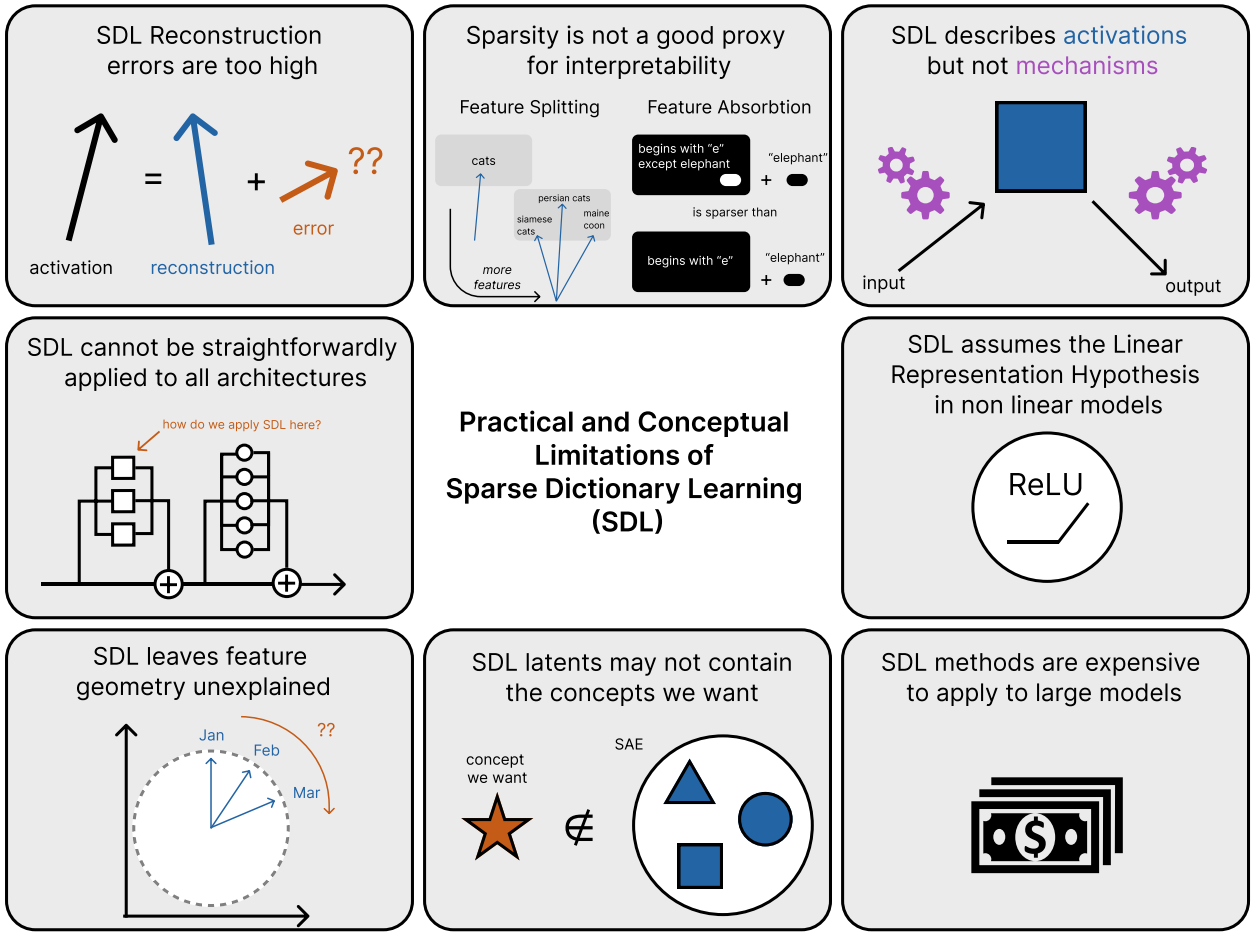

Figure 4: Sparse dictionary learning has a number of practical and conceptual limitations that cause issues when using it to reverse engineer neural networks (Section 2.1.2).

strong version is false for some models (Black et al., 2022; Csordás et al., 2024). The weaker version of the linear representation hypothesis is supported by the successes of linear probes (Section 2.2.1), activation steering (Section 3.2.2), and the success of sparse autoencoders in finding seemingly interpretable latents (Section 2.1.2).

Sparsity is not a good proxy for interpretability: Another of SDL's key assumptions is that feature activations are sparse. Therefore, SDL methods optimize their latents accordingly to be sparsely activating, with the implicit assumption that sparser decompositions are more interpretable than denser ones. However, this assumption may not necessarily hold. The (related) problems of feature splitting (Bricken, 2023), feature absorption (Chanin et al., 2024) and composition (Till, 2024) suggest that with sufficient optimization pressure, sparsity as a proxy for interpretability breaks down (though note that it is debated whether feature splitting is actually a problem). Other proxies, such as minimum description length, might be better optimization targets than sparsity *per se* (Ayonrinde et al., 2024). It also may be the case that no proxy metric is sufficient.

SDL leaves feature geometry unexplained: SDL decomposes networks into single directions in the activation space, which is a reasonable approach only if we consider feature activations akin to 'a bag of features' (Harris, 1954) without any internal structure. However, the geometric arrangement of features in relation to each other seems to reflect semantic and functional structure (Engels et al., 2024a; Gurnee & Tegmark, 2024; Park et al., 2024b; Bussman et al., 2024). Understanding the geometric structure of a network means understanding the position of one feature relative to (potentially) many others. Comprehending this

may be necessary if we want to know why and how a network treats certain features similarly or differently. If understanding the global geometry (all-to-all relationships) of features is essential to understand neural networks, this might pose a fundamental problem for current approaches to mechanistic interpretability (Hänni et al., 2024; Mendel, 2024). However, if only local geometric relationships between features need to be understood, understanding networks with a 'bag of features' approach may be more feasible.

SDL cannot straightforwardly be applied to all architectures: SDL was originally developed to identify features that may be represented in superposition across many neurons within a single layer. However, representations may be spread over other network architecture components besides neurons. In transformers, for example, representations may be spread across separate attention heads (Jermyn et al., 2023; Janiak et al., 2023), and even across different layers (Yun et al., 2021; Lindsey et al., 2024). However, it is not immediately obvious how to decompose representations distributed across attention heads with SDL (Mathwin et al., 2024; Wynroe & Sharkey, 2024). Nor is it straightforward to translate cross-layer distributed representations into causal descriptions of neural network mechanisms (Lindsey et al., 2024).

SDL decomposes the input and output activations of network mechanisms, but not the mechanisms themselves: Ultimately, the aim of mechanistic interpretability is to understand the mechanisms learned by neural networks. The parameters of the network, along with its nonlinearities and other architectural components, implement these mechanisms, which are applied to the input's hidden activations. SDL identifies directions in activation space. Being activations, they only interact with the network's mechanisms, but are not the mechanisms themselves. Describing the mechanisms directly remains unresolved with SDL. Gaining insights about the network's mechanisms from SDL latents requires further post hoc analysis, which can be labor intensive, computationally expensive, or data set dependent (Huben et al., 2024; Bricken, 2023; Riggs et al., 2023; Marks et al., 2024). This is an instance of a more broad problem with current mechanistic interpretability work; we primarily focus on understanding neural network activations, with little attention paid to how this structure in activations is computed via weights (Chughtai & Bushnaq, 2025).

SDL latents may not contain the concepts needed for downstream use cases: When using SAEs for practical tasks, there sometimes exists a single latent or small set of latents representing some concept of interest for task performance (and not representing much else). For instance, Kantamneni et al. (2024) found a single latent whose activation pattern was more accurate than official dataset labels on the NLP task GLUE CoLA (Warstadt et al., 2019). More often than not though, a sparse set of latents that encode some useful concept of interest do not exist. It is unclear what causes this problem. One hypothesis is that the concept we want isn't how the model 'thinks' about the concept, and the SAE is working as intended. Alternatively, the SAE training distribution could be too narrow, resulting in the SAE not being incentivised to learn the important latents. (Kissane et al., 2024c) found that SAEs trained on pretraining data generally do not have good latents for the concept of 'refusing' harmful user requests, while SAEs trained on chat formatted data do. Or, the SAE might not have a large enough dictionary size to learn all concepts of interest. Many more hypotheses are plausible. A complicating factor in using SDL to identify the learned mechanisms of neural networks is that the latents identified by depend on the data set used to train them. This is an undesirable property for a decomposition method that was initially hoped to be capable of identifying the fundamental units of computation in neural networks (Kissane et al., 2024c).

Current decomposition methods lack solid theoretical foundations. Given the practical and conceptual issues with SDL, there is broad agreement that the question of how to correctly decompose networks into atomic units remains a central problem, evidenced by the large amount of effort focused on the direction in recent years. After investing considerable effort in SDL approaches, it is apparent that improved conceptual clarity beyond the idea of superposition (Elhage et al., 2021) is needed to advance neural network decomposition.

One of the most significant open questions is the absence of clarity around the nature of features, despite being the central focus of SDL's identification efforts. Satisfying formal definitions are elusive and conceptual foundations are not yet established. However, even without foundations, progress in mechanistic interpretability is possible — even confused concepts can be pragmatically useful (Henighan, 2024).

Without solid conceptual foundations, it remains unclear whether the superposition hypothesis, which underpins the SDL paradigm, is fundamentally valid or merely pragmatically useful (Henighan, 2024; Templeton et al., 2024). If it is the latter, there may be better methods than SDL for carving neural networks at their joints. Such methods may take into account feature geometry or take a more dynamic, developmental view of how mechanistic structure emerges in the training process (Hoogland et al., 2024; Wang et al., 2024b). Moreover, although there is much emphasis on how models represent features in superposition, there is comparatively less work examining how they might perform computation on them natively in superposition. Further conceptual work into this problem may suggest new methodologies for decomposing networks, or provide bounds on the number of features we should expect models to be capable of learning (Hänni et al., 2024; Adler & Shavit, 2024; Bushnaq & Mendel, 2024).

Mechanistic interpretability should ideally be built on more formal foundations. Some work attempts to ground mechanistic interpretability in the formalisms of causality (Geiger et al., 2024a). However, this approach has not yet yielded canonical causal mediators (Mueller et al., 2024) that may form a basis for decomposition methods. How should we go about finding them? Given that our field's objective is to understand the learned structures that underlie networks' generalization behaviors, exploring theories about why neural networks generalize appears to be a promising avenue. But theories that attempt to characterize why neural networks generalize, such as the spline theory of neural networks (Balestriero & richard baraniuk, 2018), theories of neural networks' simplicity bias (Valle-Perez et al., 2019), deep learning theory involving the neural tangent kernel (Jacot et al., 2018; Roberts et al., 2022), or singular learning theory (Watanabe, 2009; Wei et al., 2023) have either not yet yielded mathematical objects that can be easily used for interpretability, or simply have not been successfully linked to approaches for interpreting neural networks. Establishing these connections would make significant progress toward carving neural networks at their joints to facilitate mechanistic interpretability. And if we can carve trained networks at their joints, it may suggest ways to train networks such that they come 'pre-carved'. Thus, better theoretical foundations may also be important for developing models that are intrinsically decomposable by design, which we discuss next.

Intrinsic interpretability: Building more easily decomposable models The current strategy of training a model solely for performance and then interpreting it post hoc may not be optimal if our goal is a model that is both interpretable and performant. To this end, it may be beneficial to prioritize interpretability during model training, for which there are several plausible approaches.

Instead of post-hoc decomposing trained network activations into discrete codes (Section 2.1.2), network activations could be forced to use a discrete code from the outset, as in Tamkin et al. (2025). MLPs could also be trained with sparser activation functions, such as TopK (Makhzani & Frey, 2013; Bills et al., 2023) or SoLU (Elhage et al., 2022). Similar approaches could be potentially used to restrict attention superposition (Jermyn et al., 2023) by limiting the number of heads attending to any query-key pair. Several approaches, such as 'mixture of experts' (Shazeer et al., 2017; Fedus et al., 2022; He, 2024), use sparsely activating components — with a large enough number of experts, and with sufficient activation sparsity, experts may become individually interpretable (Warstadt et al., 2019). Many attempts to incentivize interpretable activations directly so far have not been competitively performant, and have also allowed 'superposition to sneak through', mitigating benefits.

We can also target weight sparsity directly during training, including approaches such as L0 regularization (Louizos et al., 2018) or pruning (Mozer & Smolensky, 1988; Frankle & Carbin, 2019; Mocanu et al., 2018). Han et al. (2015) achieve sparse weights by using magnitude pruning followed by finetuning. This highlights a general strategy of finetuning with interpretability in mind (also used by Tamkin et al. (2025)), avoiding the potentially excessive cost of training from scratch. Another related strategy is targeting modularity (Kirsch et al., 2018; Andreas et al., 2016). For example, brain-inspired modular training (Liu et al., 2023) trains for modularity by embedding neurons in a geometric space and encouraging geometrically local connections.

Existing research implements other approaches that can simplify linearity-reliant circuit analysis (Elhage et al., 2021). For example, there is work removing the layer norm operations (Heimersheim, 2024), using input-switched affine transformations for recurrence (Foerster et al., 2017). Other studies leverage architectures that are mathematically analyzable in other ways than linearity, such bilinear activations in MLPs (Sharkey, 2023; Pearce et al., 2024).

After decomposing a network into components, whether through post hoc decomposition or by using intrinsically decomposable models, our task remains unfinished. We must provide an interpretation of the functional role of each component (Step 2 – Section 2.1.3) and validate that interpretation (Step 3 – Section 2.1.4). In the next section, we will discuss common methods of interpretation and their shortcomings.

2.1.3 Reverse engineering step 2: Describing the functional role of components

After decomposing networks into parts, the next step of reverse engineering is to "describe" the functional role of these components. This step is best thought of as generating hypothesized 'interpretations' or 'explanations'. These explanations form candidate descriptions of the functional role of a given component, and should not be taken to be definitive conclusions before they are thoroughly validated (see Section 2.1.4).

Descriptions of the functional role of neural network components can either indicate (1) the cause of a component's activation or (2) what occurs after that component has been activated, or – preferably – both (Figure 5). In this section, we will discuss the existing set of tools available to mechanistic interpretability researchers to describe the functional role of network components.

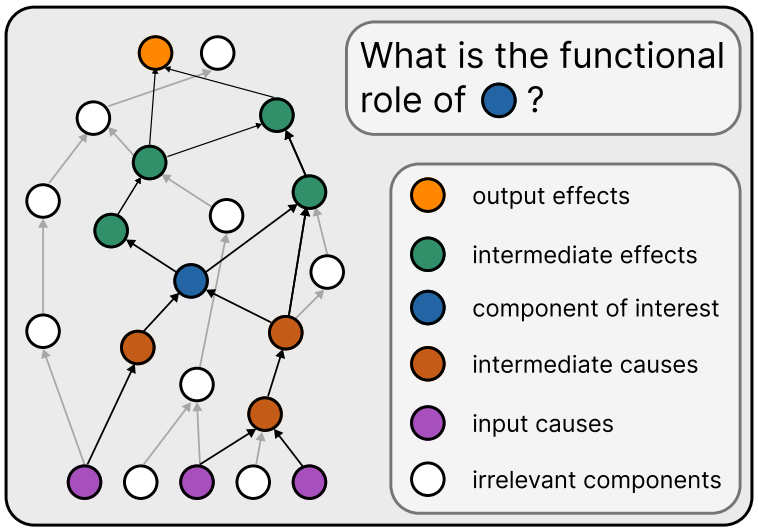

Figure 5: To study the functional role of the blue component numerous approaches are possible. We could study its causes: the purple input components or the red intermediate components via e.g. feature synthesis (Olah et al., 2017a; 2020b), maximum activating examples (Olah et al., 2017a; Bricken, 2023), or attributions (Sundararajan et al., 2017). Or we could study its effects: the orange output components or the green intermediate components via e.g. the logit lens (Nostalgebraist, 2020), activation steering (Turner et al., 2024), or attributions (Marks et al., 2024).

Explanations for what causes components to activate Explanations of the causes of component activation can use three broad categories of methods, each with several problems: (1) Highly activating data set examples; (2) Attribution methods; and (3) Feature synthesis.

Highly activating data set examples. The simplest method is to use highly activating data set examples (sometimes called 'exemplar representations' (Hernandez et al., 2022)). These are inputs on which a particular component is strongly activated. For a given component, analyzing apparent commonalities in the inputs suggests hypotheses for what causes that component to activate. This step may be carried out by humans or AI systems (Section 2.4).

Despite being widely used, this approach has several substantial issues. The first issue is that the method relies on human prior beliefs, which may lead interpreters to project their human understanding onto models that may, in fact, be using unfamiliar concepts. This bias could lead us to identify concepts in the model

that do not truly explain the model's functioning (Freiesleben & König, 2023; Donnelly & Roegiest, 2019; Gale et al., 2020).

Another issue with this approach is the potential for 'interpretability illusions'. Bolukbasi et al. (2021) show that bias in data sets can create misleading explanations even when top activating examples are selected from real data sets, as opposed to synthetically created data. Depending on the data set from which the examples were drawn, human annotators identified dramatically different meanings for given directions in the activation space of BERT (Devlin, 2018).

A third issue is that this approach often yields plausible explanations for arbitrarily chosen directions in the activation space (Szegedy et al., 2014). This means that plausible explanations based on highly activating data set examples cannot be solely relied upon to identify the basic units of computations in neural networks — other methods are needed to accurately identify them. Furthermore, it is possible to develop adversarial models that deliberately yield misleading feature visualizations (Geirhos et al., 2025).

Many of the issues with highly activating data set examples stem from the fact that they merely provide correlational explanations for the activation of a network component, rather than causal explanations. To identify causal explanations, other methods, such as attribution methods, are necessary.

Attribution methods are necessary for causal explanations but are often difficult to interpret. Attribution methods (Simonyan et al., 2014a; Nguyen et al., 2016b; Selvaraju et al., 2019; Sundararajan et al., 2017; Fong & Vedaldi, 2017; Lundberg & Lee, 2017; Ribeiro et al., 2016) are intended to measure the causal importance of upstream variables (such as inputs) on downstream variables (such as a network component). They are often gradient-based (Mozer & Smolensky, 1988; Simonyan et al., 2014a; Nguyen et al., 2016b; Selvaraju et al., 2019; Sundararajan et al., 2017; Wang et al., 2024d) or sampling-, perturbation-, or ablation-based (Fong & Vedaldi, 2017; Vig et al., 2020; Geiger et al., 2020; Ghorbani & Zou, 2020; Meng et al., 2022b; Chan et al., 2022a; Nanda, 2023a). However, on a theoretical level, many gradient-based methods identify only a first-order approximation of the ideal attribution, which is sometimes a poor approximation (Watson, 2022). Adebayo et al. (2018) revealed more practical implications, demonstrating that some gradient-based methods identify attributions that are independent both of the model and of the data generating process. Furthermore, an adversary can train a model or perturb an input to reveal any attribution map (Dombrowski et al., 2019; Ghorbani et al., 2019; Heo et al., 2019; Kindermans et al., 2019; Slack et al., 2020; Zhang et al., 2020)[4]. Perturbation methods present certain issues, such as taking models off their training distribution and eliciting unusual behavior, among other theoretical complications (Feng et al., 2018; Molnar et al., 2021; Hooker et al., 2021; Molnar et al., 2024; Freiesleben & König, 2023; Slack et al., 2021). Developing efficient and accurate attribution methods thus remains an open problem.

Feature synthesis. Feature synthesis is a strategy that combines highly activating data set examples and gradient-based attribution methods (Erhan et al., 2009; Szegedy et al., 2014; Olah et al., 2017a). This approach attempts to synthesize inputs that maximize the activation of a component subject to some regularization, such as consistency with a generative model (Nguyen et al., 2016a; 2017) or total variation distance (Mahendran & Vedaldi, 2014). However, criticisms of feature synthesis methods suggest that natural dataset examples may serve interpretation better (Zimmermann et al., 2021; Borowski et al., 2021) or show that current methods struggle to identify trojans (Casper et al., 2023a).

Explanations for the downstream effects of components Alternatively, the functional role of a component can be described through its downstream effects.

Studying the direct effect. The logit lens (Nostalgebraist, 2020) applies the models unembedding matrix to an intermediate residual stream representations, converting it into a distribution over the model's output vocabulary. Direct logit attribution is a generalization of this technique, applying to any appropriately sized vector space in the model, e.g. the output of MLP layers (Geva et al., 2021; 2022b;a; Dar et al., 2023), attention blocks, gradients (Katz et al., 2024) of these, and SDL decoder weights (Bricken, 2023). Adding a trainable affine or linear transformation before unembedding, as also referred to as the tuned lens, improves

[4]However, Freiesleben & König (2023) argue that adversarial examples do not truly undermine saliency maps as these are highly dissimilar to real interpretability challenges.

decoding accuracy (Belrose et al., 2023a; Yom Din et al., 2024), at the cost of less faithfully representing when the model has completed computation.

In the language of causality (Pearl, 2009), unembedding a residual stream vector measures the direct effect of that vector on the output (McGrath et al., 2023). However, the logit lens cannot measure the indirect effect, the effects resulting from the influence that the embedding has on the hidden activations of subsequent layers. Other methods, such as causal interventions (see below), are necessary to measure the indirect effect. In the future, it may be possible to extend logit-lens-like approaches to not only project the effects of network components on directions in output vocabulary space, but also on intermediate downstream components.

Causal Interventions. Causal interventions typically substitute ("patch") the value of some network component, usually an activation vector, with a different value during a forward pass, observing the resulting effect on the model. There exist a related set of techniques: ablation (for the special case of zero-ing or otherwise attempting to delete activations entirely), activation patching, causal mediation analysis, causal tracing, and interchange intervention (Vig et al., 2020; Geiger et al., 2020; Meng et al., 2022a; Chan et al., 2022a; Nanda, 2023a). [5]

More surgical patches are also sometimes also made on edges between network components. This approach, known as "path" patching, allows us to isolate the effect of one particular component on only one other component, instead of on the entire rest of the network (Goldowsky-Dill et al., 2023). Causal scrubbing (Chan et al., 2022a) is a generalization of path patching that allows for testing hypotheses concerning any given connection between a set of network components.

Causal intervention methods can also generate supervision signal to identify subspaces of interest. For instance, distributed alignment search (Geiger et al., 2024c) learns a linear subspace that represents a particular concept, using data with interventions on that concept as supervision (Geiger et al., 2024c; Guerner et al., 2024; Wu et al., 2023). Other work learns masks over network components to remove irrelevant components (De Cao et al., 2020; Csordás et al., 2021; Davies et al., 2023).

A causal intervention typically requires a forward pass of the model. This may make performing one for every network component in large models, long contexts, or when using finer-grained components such as sparse autoencoder latents prohibitively expensive. Faster alternatives that work well in practice (Marks et al., 2024; Templeton et al., 2024) include gradient-based approximations to activation patching, such as attribution patching (Nanda, 2023a; Syed et al., 2024), AtP* (Kramár et al., 2024), and integrated gradients (Sundararajan et al., 2017).

Observing the effects of components on sequential behavior. Another way to study the effects of model components is to patch in activated components and observe their effect on model behavior. Steering is one example of this (Rimsky et al., 2024; Turner et al., 2024), where components (activation vectors) are activated in order to influence the network's behavior, often in interpretable ways. To determine the functional role of an activation vector, a related method is to have a language model itself decode the activations (Chen et al., 2025; Watkins, 2023; Ghandeharioun et al., 2024; Huang et al., 2024b; Kharlapenko et al., 2024). These approaches, known as "patchscopes", patch activations from one forward pass of a model into a different forward pass (perhaps in a different model). The context of the new forward pass is designed to elicit relevant information from the activations of the original forward pass.

A related sequential behavior based technique is simply to read the chain-of-thought (Wei et al., 2024; Kojima et al., 2024) produced by a language model's output. While this could be considered an 'interpretability technique' in the sense that it aims to explain model decisions, it does not use model internals in those explanations, at least not directly. Recent research demonstrates that chains of thought are not entirely faithful to the model's underlying decision-making process (Agarwal et al., 2024; Atanasova et al., 2023; Turpin et al., 2023; Lanham et al., 2023; Ye & Durrett, 2022). A promising future direction for interpretability research may be to incorporate model internals into chain-of-thought training, which could incentivize

[5] For instance, by patching activations from the corrupt prompt "the capital of Italy is" into the clean prompt "the capital of France is", we can observe the effect on the output ("Rome" vs."Paris"). This tells us which component values are relevant for the differing output between the two prompts, but not the information that remains consistent (e.g. the fact that the answer is a city).

faithfulness. Another possibility is building monitors based on model internals to improve transparency in chain-of-thought faithfulness.

2.1.4 Reverse engineering step 3: Validation of descriptions

Initial descriptions of network components' functional roles should be treated as hypotheses that first require validation to ensure that they are reasonable.

Conflating hypotheses with conclusions has regrettably been commonplace in mechanistic interpretability research, making validation an important area for the field to improve (Madsen et al., 2024; Stander et al., 2025). Unfortunately, it is often hard to distinguish faithful explanations of neural network components from merely plausible ones. Numerous examples of model interpretations fail sanity checks (Adebayo et al., 2018; Leavitt & Morcos, 2020; Miller et al., 2024); "interpretability illusions" in which seemingly convincing interpretations of a model later turned out to be false (Bolukbasi et al., 2021; Makelov et al., 2023); or instances where different approaches to explaining the same phenomenon yielded different interpretations (e.g. Chan et al. (2022b) or Chughtai et al. (2023) vs. Stander et al. (2025) vs. Wu et al. (2024a)). Several other instances were previously cited in this review (Section 2.1.3). Hypotheses in interpretability require extensive validation beyond what appearances might imply.

Validating a hypothesis involves posing a simple question: Does the hypothesis make good predictions about the neural network's behavior? Testing the hypothesis often requires multiple approaches (Mueller et al., 2024). To validate descriptions, many approaches simply apply a different description method than the one used to generate the initial hypothesis (Section 2.1.3). If one description method yields a different description from another, it invalidates the hypothesis, which necessitates returning to an earlier step in the reverse engineering cycle (Figure 2). However, methods for validating descriptions are not limited to other component description methods. Hypothesis validation may take many forms, including the following:

Predicting activations and counterfactuals: By using natural language explanations of a given network component's function, it is possible to predict the component's activation levels on different inputs. This analysis can be carried out by humans or by AI systems (Section 2.4), as in (Hernandez et al., 2022; Bills et al., 2023; Shaham et al., 2025; Juang et al., 2024). In essence, our interpretations should enable us to successfully predict counterfactual scenarios in neural networks. For instance, if we ablate or activate particular network components, we should be able to predict specific downstream effects on other components.

Predicting and explaining unusual failures or adversarial examples: Good explanations of neural network behavior should help us identify and explain cases where that behavior fails to produce expected outcomes. For instance, Hilton et al. (2020) validated their methods by explaining cases where a deep reinforcement learning agent's neural network failed to achieve maximum reward and also explained specific hallucinations exhibited by the network. Another approach is to use use an interpretability approach to handcraft an input for a network that functioned as an adversarial example (Carter et al., 2019; Casper et al., 2023c; Mu & Andreas, 2020; Hernandez et al., 2022).

Handcrafting a network that reconstructs a network behavior: If our explanations for network behavior are sufficient, we should be able to use them to build replacement parts for the original network. Cammarata et al. (2020) validated their interpretation of a curve detector's function in a convolutional neural network by substituting its parts with simple handcrafted replacements.

Testing on ground truth: If the weights of a toy neural network were handcrafted by humans, it is possible to obtain a ground truth explanation for how it works. This proves useful for testing explanations produced by interpretability methods. For example, Conmy et al. (2023) validated a tool's ability to attribute model behaviors to internal components by running it on a simple model that implemented a known algorithm. (See also Section 2.1.4).

Using the hypothesis to achieve particular engineering goals: Another way to test explanations is to assess their utility in downstream applications (Doshi-Velez & Kim, 2017; Casper et al., 2023a). For example, Templeton et al. (2024) discussed examples where manually editing a large language model based on an interpretation led to predictable high-level changes in its behavior. Meanwhile, Marks et al. (2024) showed how an interpretability tool could assist humans with debugging a classifier in a toy task. Farrell

et al. (2024) use unlearning (Section 3.2.2) to demonstrate that learned SDL latents don't quite match human concepts, and might not be optimal for particular downstream use cases, highlighting potential issues with SDL.

Using the hypothesis to achieve specific engineering goals competitively: Achieving not only useful, but competitive methods sets an even higher standard. For interpretability tools, the highest evaluation criteria require fair comparisons against relevant baselines on real-world tasks instead of cherry-picking them. However, the practice of conducting evaluations using non-cherry-picked tasks remains relatively uncommon. Although attempts have been made to use techniques in the current interpretability toolkit in such evaluations, they have not proven to be consistently useful (Adebayo et al., 2020; Denain & Steinhardt, 2023; Casper et al., 2022b; Hase et al., 2023; Durmus et al., 2024). Some of the most promising research directions are to use interpretability methods to achieve things that would be hard or impossible to achieve without them (Schut et al., 2023). Unless interpretability methods demonstrate that they are competitive with alternative approaches to achieve engineering goals, then the act of demonstrating their usefulness may lead to a bias toward developing methods that only perform well in best-case scenarios and on simple tasks, rather than those that can handle worst-case scenarios and practical challenges.

Interpretability researchers have historically faced challenges in adequately validating their hypotheses due to the high costs in terms of time and cognitive labor. In the following section, we explore two potential solutions that could simplify the validation process: model organisms and interpretability benchmarks.

'Model organisms' facilitate hypothesis validation. Although interpretability is sometimes motivated by achieving engineering goals, it is often also approached through the perspective of the natural sciences (Olah et al., 2020b). In certain natural sciences, such as genomics and neuroscience, it is common for researchers to investigate a few extensively studied species known as 'model organisms' or 'model systems'. By conducting in-depth studies on a select group of organisms, like *E. coli*, fruit flies, mice, and macaque monkeys, researchers can leverage the insights and tools gained from those organisms and apply them to other species. For example, imaging specific types of neural activity in mice is more tractable due to existing hypotheses about which proteins should be fluorescently labeled in order to identify specific types of neurons. The use of model organisms allows for cross-checking results with previous work, enabling stronger validation of hypotheses.

Currently, interpretability researchers lack consensus on which networks should serve as model organisms. Essentially, what should the *Drosophila melanogaster* of mechanistic interpretability be? In mechanistic interpretability, an ideal model organism should be open source, easy and cheap to use, representative of a broad range of systems and phenomena, have a replicable training process with open source training data, and have multiple instances with different random seeds, among other criteria (Sharkey et al., 2022a). Thus far, researchers have mostly used model organisms that possess only some of these criteria, such as a transformer than can perform modular addition (Nanda et al., 2023a) or GPT-2 (Radford et al., 2018).

Model organisms not only support cross-validation of hypotheses, but also facilitate the progressive construction of experimental infrastructure by providing a reliable foundation for experiment design. This simplifies the process of rigorous hypothesis testing, thus helping prevent oversimplification and 'interpretability illusions'.

Studying solely model organisms, instead of more directly pursuing engineering goals, risks merely making true statements about neural network structure, rather generating insights that are of immediate practical benefit. For mechanistic interpretability to make the fastest and most substantial progress toward engineering goals, both scientific and engineering wins should be pursued in parallel.

Furthermore, certain choices made while studying model organisms risk steering the field in suboptimal directions. For instance, interpretability research is often motivated by the engineering goal of understanding state-of-the-art models thoroughly enough to make assurances of their safety (Bereska & Gavves, 2024; Tegmark & Omohundro, 2023; Dalrymple et al., 2024). However, limiting its focus by studying small toy models (e.g. Nanda et al. (2023a)) or how larger models accomplish select subtasks (Arditi et al., 2024), risks incentivizing research and methods that fail to generalize to more safety-relevant real-world settings.

Validating interpretability methods using benchmarks. Beyond validating individual hypotheses, we may wish to validate entire interpretability methods. Benchmarking is a proven approach to making incremental improvements in other areas of machine learning, with several approaches to benchmarking interpretability methods being developed in recent years.

One desideratum for interpretability benchmarks is to evaluate interpretations against ground truth explanations (Freiesleben & König, 2023; Zhou et al., 2022). Benchmarks can be established using models with known ground truth explanations. Such models can be created by compiling simple programs into weights of models that exactly implement the known program (Lindner et al., 2023; Weiss et al., 2021; Thurnherr & Scheurer, 2024; Gupta et al., 2024). Alternatively, predetermined explanations can be enforced at training time in conventional models using Interchange Intervention Training (Gupta et al., 2024; Geiger et al., 2022). Other interpretability benchmarks that evaluate specific steps in the interpretability pipelines also exist, such as model decomposition (Huang et al., 2024a; Makelov et al., 2024), generating descriptions of network component functions (Schwettmann et al., 2023) , or testing natural language explanations (Huang et al., 2023).

2.2 Concept-based interpretability: Identifying components for given roles

2.2.1 Concept-based probes

When attempting to localize a human-interpretable concept within the network, an intuitive approach is to 'probe' for it (Köhn, 2015; Gupta et al., 2015; Alain & Bengio, 2017; Ettinger et al., 2016). A concept-based probe is a classifier trained to predict a concept from the hidden representation of another model (Hupkes & Zuidema, 2018). Probing requires a labeling function that assigns classification labels to input data, indicating the 'value' of the concept on that data (Note: A binary value indicating the presence or absence of a concept is a special case of this approach). Once the labels are assigned, a probe, which is a simple parameterized model, is trained to predict concept labels based on hidden activations. If the probe is a linear model, then we have localized the concept as a vector in latent space.

Probes were first introduced in NLP (Köhn, 2015; Gupta et al., 2015) and have since been extensively explored in the field (Conneau et al., 2018; Tenney et al., 2019; Rogers et al., 2020; Gurnee et al., 2023; Peters et al., 2018; Burns et al., 2023; Marks & Tegmark, 2024). They have also been applied in vision (Alain & Bengio, 2017; Kim et al., 2018b) and deep reinforcement learning (McGrath et al., 2022; Forde et al., 2023). Probing also includes concept activation vectors (Kim et al., 2018a), information-theoretic probing (Voita & Titov, 2020), and structural probing (Hewitt & Manning, 2019).

Although relatively simple to implement, probing has two main challenges (Ravichander et al., 2021; Belinkov, 2022b): (1) The need for carefully chosen data for well-defined concepts, and (2) probes detect correlations instead of causal variables in hidden activations.

2.2.2 Probes need carefully chosen data for well-defined concepts

Concept-based probing requires a labeling function that assigns labels to input data. Obtaining a labeling function is not always trivial and may require substantial human effort to define a data set for a single concept. Moreover, it is only possible to identify concepts that we have defined precisely enough to create high-quality data. This limitation implies that concept-based probing can only identify concepts that we were already looking for, rather than reveal unexpected features in the network. Another approach, known as Contrast-Consistent Search (CCS), probes not for single concepts but an axis in activation space that corresponds to positive or negative propositions by enforcing probabilistic consistency conditions (Burns et al., 2023). Despite being more unsupervised than standard concept-based probing, even CCS requires the construction of data sets with clear positive and negative cases. A potential path forward for concept-based decomposition is to develop methods that automatically develop data sets for probing and concept localization (Shaham et al., 2025) (Section 2.4).

2.2.3 Probes detect correlations, rather than causal variables, in hidden activations

Probes are tools for a correlational analysis, measuring if hidden activations serve as signals for a given concept. Indeed, from an information-theoretic point of view, an arbitrarily powerful probe measures the mutual information between a hidden representation and a concept (Hewitt & Liang, 2019; Pimentel et al., 2020).

However, training a probe to associate a concept with specific hidden activations does not necessarily imply that those activations causally mediate how that concept is used by the network, or even if the network uses the concept at all (Ravichander et al., 2021; Geiger et al., 2024b; Elazar et al., 2021; Belinkov, 2022b). Probes can be successfully trained on hidden activations that lack any causal connection to the output, only localizing correlated hidden activation vectors. For this reason, probing should be used only to generate hypotheses about which network components might be causally linked to a concept. Confirming such hypotheses requires further investigation using causal interventions or other probing methods.

To improve the causal relevance of probe vectors, one approach is to use counterfactual data, which involves intervening on the concept of interest (for instance, observing the resulting output if the dog in an image was changed to a cat) (Elazar et al., 2021; Mueller, 2024; Geiger et al., 2024b). Methods include distributed alignment search (Geiger et al., 2024c; Wu et al., 2023; Huang et al., 2024a), causal probing (Guerner et al., 2024), using attribution methods to measure the effect of concept vectors on network predictions (Kim et al., 2018b), and various concept erasure methods (Ravfogel et al., 2020; 2022; Elazar et al., 2021; Belrose et al., 2023b;b). While these methods can identify causal mediators of concepts in hidden representations, they require more specialized data than probes do.

At times, it might be acceptable for probes to identify merely correlated hidden activations if the correlations generalize to the test distribution. However, probing approaches face a considerable risk of not only discovering correlated features, but also spurious correlations due to the high dimensionality of hidden activations. Validating probes is therefore essential to avoid overfitting. This includes evaluating them on out-of-distribution test data that varies along task-specific dimensions to ensure that general purpose features have been found. One question that remains unanswered is how to use regularization to achieve good probe generalization.

2.2.4 Concept-based intrinsic interpretability

Although probes begin with a trained network and search for specific concepts within it, it is also feasible to leverage the concepts in the network training process itself, such as in the case of concept bottleneck models (Koh et al., 2020). This is beneficial as it is more likely that the components to which concepts are assigned are causally relevant. Instead of specific concepts, networks can also be trained to use particular causal structures (Geiger et al., 2022). While it may not be possible to prespecify all relevant concepts or structures, integrating this approach with methods for disentangling concepts could prove useful (Chen et al., 2018). Cloud et al. (2024) incentivize modularity via applying data-dependent, weighted masks to gradients during backpropagation.

2.3 Proceduralizing mechanistic interpretability into circuit discovery pipelines: A case study

How should we codify the process of mechanistic interpretability to yield the deepest possible insights? To form a complete pipeline by combining various methods, several methodological choices regarding decomposition, description, and validation, must be made. Circuit discovery has emerged as a prominent pipeline in recent mechanistic interpretability research (Wang et al., 2023; Hanna et al., 2023; Heimersheim & Janiak, 2023). Its objective is to describe how a neural network performs a task of interest while making specific choices for network decomposition, component description, and hypothesis validation. In this section, we look at the typical choices in each step in greater depth and discuss how this popular pipeline could be improved.

The ‘circuit discovery’ pipeline takes the following steps:

1. **Task Definition**. For a given model we want to study, we select a task that the model can perform, and a dataset on which the network performs that task. This is a concept-based step, since the definition of the task was based on how human researchers define a task distribution.

2. **Decomposition**. During the decomposition step, it is common to think of the neural network as a directed acyclic graph (DAG), where activations are represented by nodes and the "abstract weights" between them represented by edges. Most work thus far has selected architectural components (Section 2.1.2), such as attention heads and MLP layers, to be the nodes. However, more recent work has also used SDL latents for nodes (Marks et al., 2024)

3. **An initial description step: Identify task-relevant vs. -irrelevant subgraphs.** The circuit discovery procedure then identifies task-relevant nodes and edges. Typically, causal interventions are used, drawing samples from some "clean" and "counterfactual" data sets. Circuit discovery methods are generally based on iterative activation patching (Wang et al., 2023; Chan et al., 2022a; Lieberum et al., 2023) or integrated gradients (Marks et al., 2024).

4. **An iterative description-validation loop**. After obtaining a task-relevant subgraph, the next step involves describing the function of each node or edge individually. This step is less formulaic than previous steps. Researchers rely on their intuition, attempting to create testable hypotheses for the function of a component or edge of the circuit, and then design custom experiments to validate or invalidate their hypothesis. Only after several iterations of hypothesis testing through experimentation are researchers finally satisfied with their explanation. In research papers, this loop is rarely made explicit, as only the final description is presented. However, Chan et al. (2022b) detail this process for understanding the induction task, and Nanda (2023b)provides another description of such a loop (building on work by Li et al. (2023)).

5. **Final Validation**. Circuits are commonly evaluated based on three attributes (Wang et al., 2023): *faithfulness*, which refers to how closely the circuit approximates the entire network's behavior, *minimality*, which assesses if nodes in the subgraph are unnecessary, and *completeness*, which determines whether any nodes not included in the subgraph are important for task behavior. Additional ad hoc validation methodologies also exist. For example, Wang et al. (2023) generate adversarial examples for their task based on their mechanistic understanding, while Shi et al. (2024) devise a suite of formal statistical hypothesis tests for circuit efficacy.

This circuit discovery procedure has yielded valuable insight, but falls short using current methods. The pipeline has several issues:

Task definition is concept-based. Defining circuits has thus far been with respect to tasks defined by humans. Miller et al. (2024) demonstrate that the within-task variance of model performance across the distribution of data points in a task is large, implying that the circuit provides a good approximation of the average case performance on the dataset, but a poor one for any individual data point. This suggests that the process of first selecting a task and then studying how the model performs it may not be an effective approach to achieve "reverse engineering" -style interpretability. Thus, it might be worth learning the task decomposition instead (Haani et al., 2024).

Network decomposition methods are flawed. Perhaps most importantly, prior circuit discovery work has attempted to decompose models in either architectural bases (Wang et al., 2023; Conmy et al., 2023) or sparse autoencoder latents (Marks et al., 2024; Huben et al., 2024), which are imperfect ways to decompose neural networks for mechanistic interpretability (Section 2.1.2). Future work could locate circuits in improved decompositions or simultaneously learn both network decompositions and circuits.

Circuit faithfulness is low. Simple early circuits were found to be unfaithful (Chan et al., 2022b; 2023). Miller et al. (2024) show that existing measures of faithfulness depend on the causal intervention implementation used, and further demonstrate that such metrics are misleading when applied to several complex end-to-end circuits. Makelov et al. (2023) argue that subspace activation patching via distributed alignment search may lead to interpretability illusion mechanisms, although these findings are contested by Wu et al. (2024c).

Scalable methods are only approximate. Identifying relevant components through individual interventions is costly when there are many components. Attribution patching Syed et al. (2024) was designed to identify potential relevant candidates for further testing through intervention, which becomes more important as the number of components expands significantly through sparse dictionary learning (Marks et al., 2024). However, attribution patching uses gradients, which only yield a first-order approximation of the effect of ablating components (Wu et al., 2024c; Molchanov et al., 2017), leaving it unclear whether this method and any improvements on it (Kramár et al., 2024) produce adequate approximations.

Circuit discovery algorithms struggle with backup and negative behavior. Additional challenges for circuit analysis arise from the effects of "backup" and "negative" behavior (Wang et al., 2023; McGrath et al., 2023; McDougall et al., 2024), which actively suppress task performance and are thus not captured by maximizing task performance metrics. Despite this, they remain important factors to consider; Mueller (2024) provides further discussion of these issues.

Streetlight Interpretability: The tasks studied so far have been deliberately selected to be simple to define and study mechanistically (Wang et al., 2023). This gives a misleading impression of the level of difficulty involved in implementing circuit discovery for any arbitrary task that a network implements. Indeed, attempts to study arbitrary circuits have proceeded less successfully (Nanda et al., 2023b).

Solving issues with current mechanistic interpretability pipelines remains an open challenge that promises significant benefit. Upon establishing reasonable procedures, automating the overall pipeline will become more feasible. However, some individual steps in mechanistic interpretability can already be fruitfully automated, as discussed in the next section (Section 2.4).

2.4 Automating steps in mechanistic interpretability research

Historically, mechanistic interpretability research has required considerable manual researcher effort, though it typically studies models that are smaller than those at the frontier. To make interpretability useful for downstream use cases, scalable approaches are crucial. In this section, we discuss *automated interpretability* methods. We will explore previous cases where manual tasks in mechanistic interpretability have been successfully automated and address open problems in further automation.

Automating feature description and validation. A task that is amenable to automation is 'describing the functional role of model components' (Section 2.1.3). With the increasing sophistication of language models, researchers have generated descriptions of the functional role of neurons in image models (Hernandez et al., 2022), neurons in language models (Bills et al., 2023), and sparse autoencoder latents in language models (Huben et al., 2024; Bricken, 2023; Juang et al., 2024) using highly activating data set examples. These interpretations are validated by assessing how effectively a human or model can use them to predict the activation of a feature in a given data set example, or predict where a feature is active within a single image or text excerpt. The success of these predictions can be used as a quantitative measurement of 'interpretability'. This was previously used to measure progress toward a decomposition method that carves networks at the joints of their generalization structure, assuming that such a decomposition would be maximally interpretable (Section 2.1.2). While imperfect, these methods for interpretation hypothesis generation and validation might be improved by automating the generation of inputs to test the interpretation hypotheses by ensuring that generations activate the interpreted feature (Huang et al., 2023), or defining more rigorous statistical tests (Bloom & Lin, 2024b). Automatic component labeling could expand in the future to include descriptions of feature effects, relationships between features (Bussman et al., 2024), or how components interact during runtime produce behavior.

Automating circuit discovery pipelines. An approach called Automated Circuit DisCovery' (ACDC) automates part of the pipeline discussed in Section 2.3 to identify computational subgraphs involved in particular tasks (Conmy et al., 2023). Several works have since improved upon and accelerated this process (Syed et al., 2024; Kramár et al., 2024; Marks et al., 2024). Note that ACDC-like approaches in general assist in identifying relevant subgraphs for a pre-defined task, but do not automate important subsequent steps, such as describing the functional role of subgraph components.

While significant progress has been made toward automating steps of mechanistic interpretability pipelines, fully automating current pipelines would not yield satisfactory explanations of model behavior[6]. Further methodological progress is required for fully automated neural network interpretability to be capable of generating the quality of interpretations necessary to achieve our goals.

3 Open problems in applications of mechanistic interpretability

Ultimately, we need mechanistic interpretability methods that enable us to solve concrete scientific and engineering problems (Figure 6). While predicting the impact of fundamental science in advance is difficult, having concrete goals in mind during research is usually beneficial. We want mechanistic interpretability methods to help us achieve various outcomes, such as monitoring and auditing AI systems more effectively (Section 3.2), controlling of AI system behavior more precisely (Section 3.2.2), predicting AI system outcomes more accurately (Section 3.3), enhancing AI system capabilities (Section 3.4), and extracting knowledge from AI systems (Section 3.5). We should also anticipate that mechanistic interpretability will uncover "unknown problems" present in systems, revealing that the true realm of challenges and possibilities is greater than what we currently perceive it to be.

As highlighted in the previous section, progress in mechanistic interpretability methods is multifaceted. Each axis of methodological advancement leads to varying degrees of progress toward different goals. Before we discuss open questions in its applications, we identify distinct axes of methodological progress that lead to different amounts of progress toward different goals.

3.1 Axes of mechanistic interpretability progress.

Decomposition vs. description of network components: Improvements in network decomposition versus component description methods offer varying benefits for different goals. Decomposition methods vary in their efficacy at carving networks at the joints of their generalization structure, while description methods can yield descriptions that vary in depth. Deeper descriptions of a component are typically more causal or mechanistic, whereas shallower descriptions may rely more on correlations and only connect to inputs or outputs without referencing intermediate causes or effects (Figure 5). Deeper descriptions thus attempt to explain more about how the component interacts with other components within the network's algorithm. Certain goals can be achieved with minimal or no progress in decomposition or description, while others may demand substantial progress.

Extent of network decomposition or description: The extent of network decomposition or description needed may vary depending on the goal. Certain goals only require an understanding of specific network components (such as an individual features or a circuit), while others might require enumerating or understanding of larger circuits or the entire model (as in 'enumerative safety').

Extent of task distribution analyzed: The scope of task distribution analysis also depends on the intended goal. For instance, monitoring a model for a single kind of behavior might only require decomposing or understanding the model only over a narrow task distribution, while others, such as formal verification, might demand understanding over the entire distribution of tasks.

Mechanistic understanding post vs. during training: Understanding the mechanisms of a fixed model might suffice for some goals, but more ambitious goals might require an understanding of not only the models' mechanisms, but also how they change during the learning process.

In this section, we'll discuss how mechanistic interpretability has been used or could be leveraged to further the field's various goals. We'll assess the progress made thus far, and identify the advancements along different axes of methodological progress that will be most crucial to success.

[6]For one attempt at this using leading decomposition and description methods, see Marks et al. (2024)

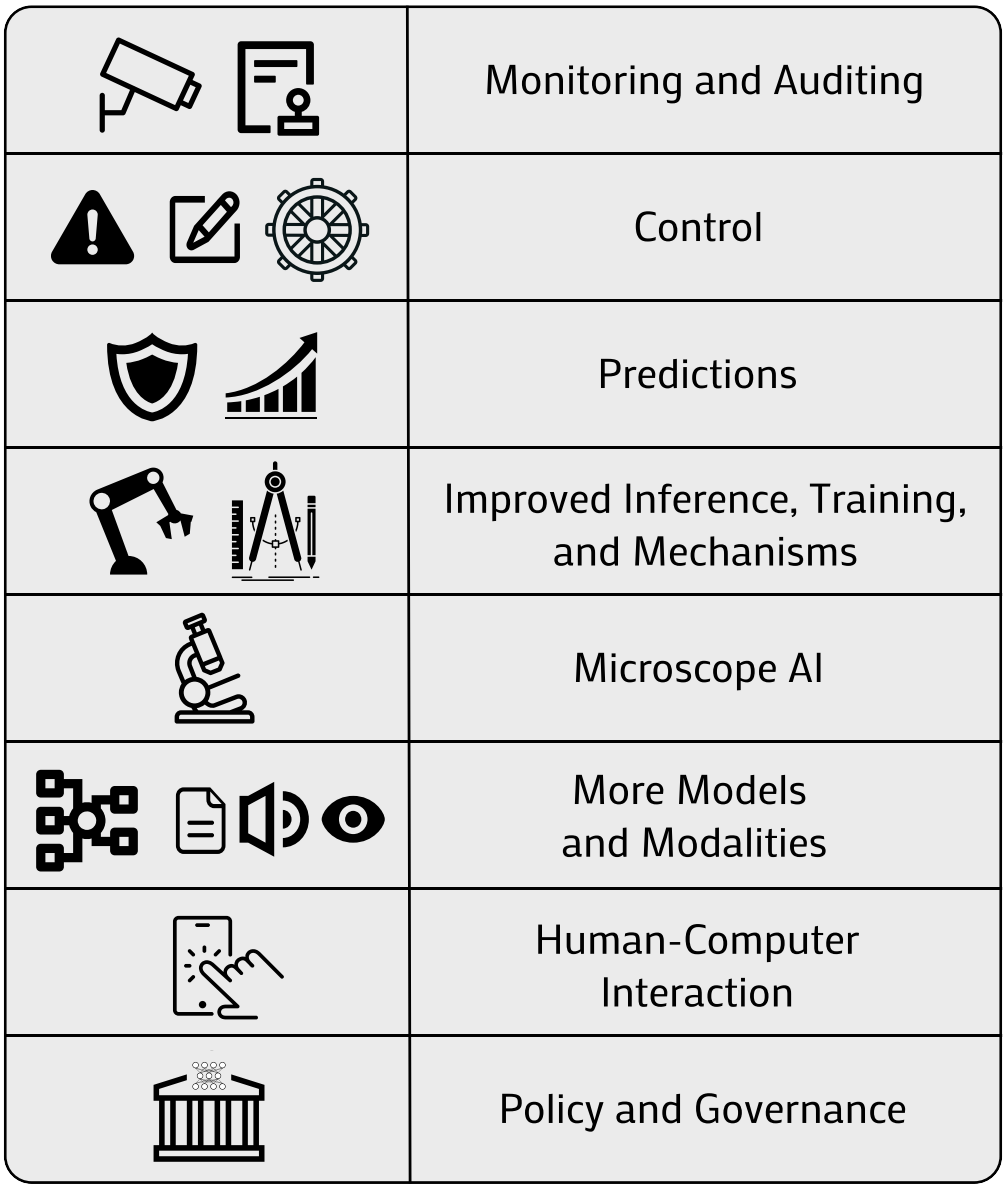

Figure 6: A summary of problem areas for applications of mechanistic interpretability.

3.2 Using mechanistic interpretability for better monitoring and auditing of AI systems for potentially unsafe cognition

3.2.1 Mechanistic interpretability-based evaluations could help us detect unsafe or unethical AI cognition

Currently, we rely on "black box" evaluations to understand a model's capabilities, but studying input-output behavior alone may not reveal all dangerous behaviors. Such behaviors include deceiving users (Park et al., 2023b; Ward et al., 2023; Scheurer et al., 2024; Meinke et al., 2024); for instance, by intentionally underperforming on evaluations ("sandbagging", van der Weij et al. (2024)), leveraging situational awareness (Laine et al., 2024); or giving dishonest responses tailored to match the user's beliefs ("sycophancy"; Sharma et al. (2024b)). Interpretability techniques could be used to uncover the mechanisms underlying these potentially harmful behaviors and thus help to detect and characterize them. This becomes increasingly

important as the capabilities of models increases, especially when using training methods that incentivize models to feign particular properties for the purpose of passing evaluation.

Using interpretability methods to identify internal signs of concern (also known as "white-box" evaluations (Casper et al., 2024) or "understanding-based" evaluations (Hubinger, 2023)) is therefore an important problem. White-box evaluation methods could serve as tools to detect potential biases that arise when models learn to use spurious correlations (Gandelsman et al., 2024a; Casper et al., 2022a; Abid et al., 2022). However, human judgment might be required to determine which features are 'supposed' to be relevant to the task (Marks et al., 2024; Kim et al., 2018a; Goyal et al., 2022).

Despite current shortcomings in decomposition and description, white-box evaluations are likely feasible today. Even shallow, correlation-based descriptions could signal potentially concerning cognition. For example, developing new methods that reliably distinguish between features that merely recognize deceptive behavior vs. mechanisms that cause deceptive behavior may be challenging. However, a correlation-based method that flags both can facilitate catching the latter. To be useful, it may not even be necessary to decompose or describe the entire network; having descriptions for components that are used on concerning subdistributions of model behavior might suffice. For instance, imperfect interpretability methods may help evaluators develop hypotheses about how models will behave, thus guiding further inquiry. Meanwhile, recent work proposes incorporating SDL (Section 2.1.2) into safety cases for advanced AI systems. By monitoring internal representations, it could aid in detecting potential sabotage or deceptive behavior before deployment (Grosse, 2024). While such approaches show promise, they have difficulty in validating whether learned features capture all concerning patterns of reasoning reliably.

Evaluations for unsafe cognition may be a particularly important use case as it plays well to the comparative advantages of mechanistic interpretability relative to the other areas of machine learning. The majority of other areas of machine learning already focus on controlling or steering the behavior of AI systems to alter input-output behavior. It is therefore unclear that this is to mechanistic interpretability's comparative advantage. On the other hand, interpretability is perhaps the only research area that attempts to understand the mechanisms of model cognition. This implies that it might be particularly fruitful for interpretability researchers to tackle problems that become easier to solve through improving such understanding: auditing for unsafe cognition, debugging unexpected behavior, and monitoring systems in deployment.

Enabling real time monitoring of AI systems for potentially unsafe cognition Beyond white-box evaluations, interpretability has further applications in monitoring. For instance, internals could be used to passively monitor the system during deployment, much like content moderation systems currently in use today. Alternatively, internals could be used to flag when a model takes an action for abnormal reasons even in absence of satisfactory descriptions (Section 2.1.3), known as "mechanistic anomaly detection" (Christiano, 2022; Johnston et al., 2024), which may be a sign of suspicious behavior. Mechanistic anomaly detection primarily requires progress in decomposition methods (Section 2.1.2) as it is necessary to be confident about what constitutes an individual mechanism within the network. Current SDL methods identify active latents, but not active mechanisms, which are implemented by network parameters. It may not be necessary to have deep descriptions of the function of individual mechanisms as long as anomalies can be detected.

Improving our ability to red-team AI systems and elicit unsafe outputs Beyond white-box evaluations, leveraging interpretability could improve our ability to conduct adversarial attacks, red-team, or jailbreak AI systems (Casper et al., 2024). This process is beneficial as it exhibits failure modes models may display in the wild, when facing adversarial pressure from wide deployment or malicious actors, thereby enabling developers to effectively preempt and address them. Furthermore, it may form a significant element of safety cases (Clymer et al., 2024; Balesni et al., 2024a; Grosse, 2024; Goemans et al., 2024) for AI systems, by providing assurances of form: We tried hard to red-team the system, yet failed to exhibit concerning behavior despite having *more* affordances than users may have. One reasonable assumption is that developers may have white-box access to models, while users may not. Although many existing red-teaming methods (Perez et al., 2022; Zou et al., 2023b) already require gradient access, we could additionally leverage interpretability insights to accelerate red-teaming. For instance, Arditi et al. (2024) discovered a universal "refusal direction" in chat-finetuned language models, which is causally important for models engaging in

the behavior of refusing harmful requests. Lin et al. (2024) used this to red-team models by optimizing for inputs that minimize the projection of the residual stream onto this direction during the forward pass. This approach may be more efficient than optimizing over the whole model, as previous methods did (Zou et al., 2023b). Mechanistic interpretability techniques also promise to improve our ability to attribute model outputs to their corresponding inputs section 2.1.3. This could enhance human red-teamers' ability to find key input features responsible for bad behavior in models, thus speeding up iteration cycles. Though such approaches are possible today with only feature-based understanding, they might improve with more crisp mechanism-based understanding.

3.2.2 Using mechanistic interpretability for better control of AI system behavior

Ensuring the safe deployment of AI first requires effective control over their behavior. Currently, the techniques used for this purpose are mostly unrelated to interpretability (e.g. Christiano et al. (2017); Rafailov et al. (2023); inter alia), but sometimes inspired by it (Rimsky et al. (2024), Zou et al. (2024), Kirch et al. (2024); inter alia). Mechanistic interpretability could assist in interpreting (Lee et al., 2025) and improving (Conmy & Nanda, 2024) these control methods, or in developing new ones. In this section, we outline interpretability-inspired control methods, and envision future possibilities with further progress in interpretability methods.

One new control method derived from mechanistic interpretability insights is **activation steering** (a.k.a. activation addition). A fixed activation vector, hypothesized to linearly represent a model concept, is added to an intermediate activation of a model at inference time (Li et al., 2024a; Turner et al., 2024; Zou et al., 2023a; Rimsky et al., 2024). Turner et al. (2024) introduced activation steering, directly inspired by the Linear Representation Hypothesis (discussed in Section 2.1.2). This technique results from the hypothesis, and its success can be thought of as evidence for the hypothesis. Moderate success can be achieved in steering using basic decomposition and description methods. Mechanistic interpretability decomposition methods enable the steering of models toward a narrower range of behaviors with fewer side effects (Chalnev et al., 2024). Advancements in mechanistic interpretability methods are likely to result in improved steering capabilities, such as activating entire mechanisms instead of individual features.

Machine unlearning was originally defined as the problem of scrubbing the influence of particular data points on a trained machine learning model (Cao & Yang, 2015). In the context of modern generative models, machine unlearning is more broadly defined as removing particular undesirable knowledge or capabilities ('unlearning targets') from models, while preserving model performance on tasks involving non-targets (Liu et al., 2024a). Targets for unlearning that are of particular interest include sensitive private or copyrighted data, model biases (Liu et al., 2024b), and hazardous knowledge that could be misused by malicious actors (Li et al., 2024b); for instance, information regarding the creation of bioweapons. A better understanding of how knowledge or capabilities are implemented within model internals can help in the development of new techniques for machine unlearning (Belrose et al., 2023b; Zou et al., 2024; Guo et al., 2024; Pochinkov & Schoots, 2024; Ashuach et al., 2024), as well as to better evaluate unlearning efficacy through white-box, non-behavioral, techniques (Lynch et al., 2024; Deeb & Roger, 2024; Hong et al., 2024). Thus far, mechanistic interpretability methods that modify intermediate activations (but not weights) for unlearning have yet to yield competitive results (Farrell et al., 2024).

Unlearning falls under the broader aim of **model (knowledge) editing**, which seeks to make precise modifications to a machine learning model that incorporates specific knowledge with desirable generalization properties, while minimizing the impact on other knowledge (Wang et al., 2024c). By attempting to carve neural networks at their joints, mechanistic interpretability could improve our ability to make interventions on knowledge with few side effects. Meng et al. (2022a) make initial progress toward interpretability-based model editing with their ROME technique. However, Thibodeau (2022) and Hase et al. (2023) highlight flaws in the technique, indicating that mechanistic interpretability has not yet found appropriate model components to intervene on (Section 2.1.2). With better comprehension of neural networks, we should anticipate more surgical model editing techniques in the future.

Editing any given capability or piece of knowledge presents a greater challenge than deleting them. Meaningful progress in unlearning and editing methods may depend on improved network decomposition methods,

as it would require isolating the individual mechanisms that correspond to specific knowledge or capabilities. Progress in mechanistic interpretability may elucidate the structure of knowledge and capabilities in AI models, leading to a better understanding of what kinds of model edits possibilities are realistic in future. Knowledge and capabilities could, in fact, be part of large mechanisms that overlap with each other, making it challenging to isolate them into discrete components. For mechanistic interpretability to effectively guide editing, strong description methods will be necessary to understand how to modify specific targets without affecting others.

Finally, mechanistic interpretability may provide tools to rigorously **understand how finetuning alters models**. This may assist in debugging instances in which finetuning leads to undesired and spurious effects (Casper et al., 2023b). Recent work (Jain et al., 2024; Prakash et al., 2024; Lee et al., 2025) suggests that existing finetuning methodologies primarily make shallow edits to existing model representations and circuitry. Importantly, this suggests that harmlessness training (which trains models to refuse to answer harmful requests) may be cheaply undone. Empirical evidence supports this claim, both with further finetuning (Gade et al., 2024; Lermen et al., 2024), as well as with causal interventions on the forward pass (Arditi et al., 2024). Separately, localizing knowledge and capabilities within models may improve the sample efficiency of finetuning, by selectively modifying only relevant parameters (as in, e.g. Wu et al. (2024b)). Further advancement in tools for comparing feature-level differences between models (such as Lindsey et al. (2024)) may accelerate our ability to debug finetuning or other control methods (Bricken et al., 2024). Mechanistic interpretability work has thus yielded several insights into how finetuning changes models and how to improve it, and may provide further insights and improvements in the future.

3.3 Using mechanistic interpretability for better predictions about AI systems

Accurately predicting model behavior in new scenarios or regimes is difficult (arguably impossible) without understanding model internals. Interpretability could hopefully facilitate two kinds of predictions:

- Predicting model behavior in novel situations
- Predicting capabilities that arise during training or finetuning

3.3.1 Predicting behavior in novel situations

In order to determine whether an AI system may potentially underperform poorly or pose a safety risk in new situations, the ability to predict its behavior in untested settings is imperative. A model's behavior, which may only become apparent in unforeseen circumstances, cannot be fully captured by its performance on a finite set of behavioral evaluations.

By understanding the mechanisms of jailbreaking, we can anticipate the means through which a user might bypass existing safeguards (Lee et al., 2025; Arditi et al., 2024). Similarly, if models have "trojans", backdoors (Hubinger et al., 2024), adversarial examples, or biases, comprehending a model's internal mechanisms could improve our ability to predict when models will display undesirable behavior, even if these scenarios were not encountered during standard training or behavioral evaluations. Casper et al. (2023a) benchmark feature synthesis tools through their ability to aid developers in identifying trojans, while interpretability assisted in identifying cases of adversarial examples (Gandelsman et al., 2024b; Mu & Andreas, 2020; Wang et al., 2023; Kissane et al., 2024a), and SDL was used to uncover biases based on spurious correlations in an LLM-based classifier (Marks et al., 2024). Beyond specific failures, interpretability methods can also be used to gain a broader understanding of network behavior. For example, prior work identified signatures in model internals that predict a model's likelihood of hallucinating (Yu et al., 2024) or its knowledge of particular facts (Gottesman & Geva, 2024).

Generally, being able to predict an AI system's behavior in advance is more challenging – but also more desirable – than merely being able to monitor its behavior and cognition. Mechanistic interpretability could allow us to make a certain type of claim, namely, "there exists no mechanisms that would cause the model to deliberately behave undesirably" (Olah, 2023). For a strong version of this claim, substantial progress in both decomposition and description methods is necessary. However, weaker versions of the claim, addressing

specific undesirable behaviors, might be more feasible with near-term methods. For instance, if it is possible to decompose networks and identify all components, even basic description methods might let us recognize that there are no mechanisms relating to bioweapons or illicit substances within the network, thus letting us predict that models are probably not capable of instructing users how to fabricate bioweapons or illicit substances (a possibility sometimes referred to as "enumerative safety" (Elhage et al., 2021; Olah, 2023)). As AI systems become increasingly agentic, claims about even more general behavior may be possible. Understanding their values or goals (or, less anthropomorphically, 'the internal mechanisms that determine their action plans and actions') should enable us to better predict their behavior across a broad range of contexts (Colognese & Jose, 2023).

When deploying AI in high-stakes scenarios, rigorous and reliable predictions are necessary, much like those demanded of safety-critical software applications. Sometimes, such software is formally verified, thereby ensuring certain safety-critical aspects of its behavior are guaranteed, since its compliance with specific properties is mathematically proven. In the context of mechanistic interpretability, the equivalent would be formal verification of AI systems (Dalrymple et al., 2024; Tegmark & Omohundro, 2023; Critch & Krueger, 2020): mathematically proving that an AI system's behavior will satisfy a desired property on any input in a given distribution. Formal verification of AI systems remains an unresolved issue at present. The level of understanding of AI necessary to enable formal verification of large, general AI systems for nontrivial properties is well beyond the current capabilities of mechanistic interpretability. However, some recent studies using toy models provide a glimpse into what solving formal verification of AI might look like. Approaches inspired by mechanistic interpretability have been used to prove accuracy bounds on a single-layer transformer trained on a synthetic task, albeit with great difficulty (Gross et al., 2024). Program synthesis through mechanistic analysis offers an alternative approach by converting simple trained neural networks into more interpretable, controllable, and verifiable programs (Michaud et al., 2024).

Several open questions remain about the tractability of scaling these approaches from toy models to frontier systems. For instance, for program synthesis, it is uncertain to what extent computations within real-world neural networks can be reduced to operations that can be cleanly represented in symbolic code, or what the total length of such code would be. Ensuring the safety of agentic systems with formal guarantees is further complicated by the need to model a system's interactions in an arbitrarily complex environment that might not be formalizable (Seshia et al., 2022; Wongpiromsarn et al., 2023; Dalrymple et al., 2024).

3.3.2 Predicting capabilities that arise during training or finetuning

The most competitive methods of AI development systems result in uninterpretable systems that often fail in ways that surprise their developers (OpenAI et al., 2024; Team et al., 2024; Anthropic, 2024). Applying mechanistic interpretability to alleviate this issue is a key area for future research.

Improved mechanistic understanding of model training could enhance the ability to predict when certain capabilities will appear. For instance, it has been observed that new model capabilities can emerge as a function of scale (Wei et al. (2022), though also see Schaeffer et al. (2023)). Evidence suggests that new capabilities may be learned in a somewhat discrete Michaud et al. (2023) or stagewise (Wang et al., 2024b) fashion, and that in synthetic data settings, the emergence of new capabilities coincides with abrupt changes in the trajectory of model parameters (Park et al., 2024a).

Other work shows a correlation between in-context learning capabilities and the emergence of induction heads, an attention-based circuit mechanism (Olsson et al., 2022). By connecting these threads of research, the long-term hope for mechanistic interpretability research is to link small-scale mechanistic structure to larger-scale structure, such as the evolving shape of the loss landscape during model scaling (Olah, 2023). To make progress toward this goal, research needs to move beyond simply improving 'decomposition' (Section 2.1.2) or 'description' (Section 2.1.3) quality and instead be capable of describing the dynamic changes in the mechanistic structure of networks throughout the learning process.

We may also want to link the emergence of capabilities to specific properties of the training data set. Through mechanistic interpretability, we can create data sets that facilitate training models to demonstrate desirable attributes and predict their behavior. By attributing model outputs to specific training examples, influence functions have been applied to LLMs (Koh & Liang, 2017; Grosse et al., 2023) to predict limitations in

their generalization abilities, such as a lack of robustness when the order of certain phrases was flipped (Berglund et al., 2024). Other work examines how data set composition shapes the emergence of in-context and weights-based learning (Reddy, 2024).

A related problem of interest involves predicting which model capabilities, that may not be present in a given model, can be "elicited" with sufficiently advanced prompting or finetuning strategies (Greenblatt et al., 2024). Prakash et al. (2024) find evidence that finetuning improves capabilities primarily by enhancing existing circuits, rather than developing fundamentally new mechanisms. Relatedly, (Jain et al., 2024) and Lee et al. (2025) show that finetuning can mask capabilities present in a base model in a way that can easily be reversed via simple changes to the model. Improved mechanistic understanding of finetuning could help reveal capabilities obscured in this fashion. Since capabilities are behaviors that often span multiple sequential steps, it may be necessary to have mechanistic interpretability methods that examine mechanisms spanning multiple time steps. However, current mechanistic interpretability research is primarily focused on understanding mechanisms involved in predictions at a single time step.

3.4 Using mechanistic interpretability to improve our ability to perform inference, improve training, and make better use of learned mechanisms

A mechanistic understanding of AI models could be leveraged to improve their utility, from faster inference and better training, to enhancing and manipulating representations.

By understanding the internal generation process of AI models, we could accelerate their inference. For example, it could help identify which parts of the computation could be skipped without changing the model's final output (Voita et al., 2019; Din et al., 2024; Voita et al., 2024; Gromov et al., 2024). The ability to inspect the information or functions implemented in a model could facilitate the development of more effective distillation methods by recognizing gaps that should be distilled (Gottesman & Geva, 2024) and discovering novel ways to distill them (Zhang et al., 2024).

Another aspect that mechanistic interpretability could enhance is model training. Interpreting how the model processes specific examples and using this information to influence its predictions (Koh & Liang, 2017; Grosse et al., 2023) may inform the selection of better training data to improve the model's capabilities in desired ways. Moreover, better monitoring of the training process can be achieved by correlating certain drops in training loss with capability gains (Olsson et al., 2022; Wang et al., 2024a) or identifying a general order in which specific capabilities emerge during training (Michaud et al., 2023). In addition, identifying the contributing components of a given task could help to devise novel, parameter-efficient training methods. Finally, being able to decompose networks into their functional components presents possibilities to build components that lend themselves to learning computational structures that we better understand (Crowson et al., 2022; Fu et al., 2023).

Mechanistic interpretability has the potential to not only accelerate AI inference and training, but also enhance its utility. Intervening in the model's computation has the potential to remove unwanted bugs in its reasoning abilities, and achieve better balance between its knowledge recall process and latent reasoning (Yu et al., 2023; Jin et al., 2024; Biran et al., 2024; Balesni et al., 2024b). More broadly, understanding the inner workings of different models could lead to better recombination of what they have learned, such as combining model parameters (Wortsman et al., 2022) and transferring representations across models (Ghandeharioun et al., 2025; Csiszárik et al., 2021).

3.5 Using mechanistic interpretability for 'microscope AI'

Current approaches for knowledge discovery from data involve statistical or causal analysis, dimensionality reduction, or using machine learning models that are inherently interpretable. These techniques can be valuable, but are influenced by human priors, typically assume linear relationships between variables, and cannot handle massive multimodal data. On the other hand, neural networks can do these things. Deep learning models are capable of encoding complex, non-linear relationships and extracting meaningful features from massive data sets without human intervention. Historically, these abilities had limited scientific value,

as without methods to interpret these models, we could not understand the patterns they found. However, with ongoing advancements in interpretability research, this is beginning to change.

Applying interpretability for knowledge discovery is sometimes called **microscope AI**. This approach involves training a neural network to model a data set, then applying interpretability techniques to the model to gain insight into any (potentially novel) predictors it discovers. In this way, the superhuman pattern matching skills of deep neural networks can serve as a tool to parse complex data sets.

Current methods allow for versions of microscope AI, depending on the kind of insights that we want to learn. Some examples of these applications include extracting novel chess concepts from AlphaZero and teaching them to top grandmasters (Schut et al., 2023) using a CNN trained on defendant mugshots and judge decisions to reveal how facial features affect judgments (Ludwig & Mullainathan, 2023), transforming psychology articles into a causal graph with an LLM to enable link prediction and produce expert-level hypotheses (Tong et al., 2024b), and analyzing a CNN to learn previously unknown morphological features for predicting immune cell protein expression (Cooper et al., 2022), among several other studies (O'Brien et al., 2023b; Hicks et al., 2021; Narayanaswamy et al., 2020; Korot et al., 2021). As interpretability methods improve along various axes, deeper insights in and across more domains will become possible.

Currently, the majority of scientists are unable to access microscope AI due to the need for specialized expertise in machine learning, interpretability, and domain knowledge to recognize significant new patterns, a combination of skills that is rare in many fields. This may change as interpretability research becomes more widely adopted in the sciences and as interpretability becomes increasingly automated and accessible.

3.6 Mechanistic interpretability on a broader range of models and model families

The vast majority of mechanistic interpretability research to date has focused on just three model families: CNN-based image models (e.g. Erhan et al. (2009); Nguyen et al. (2016c); Olah et al. (2020b)), BERT-based text models (e.g. Devlin (2018); Rogers et al. (2020)) and GPT-based text models (e.g. Elhage et al. (2021); Wang et al. (2023); Nanda et al. (2023a)). The degree of generalizability of these findings to other models and contexts is currently a somewhat open question. Given that future frontier models may use architectures that differ from the current state-of-the-art, and are expected to be multimodal by default, interpretability researchers may need to expand the range of models and modalities that they study and try to identify universal approaches that can be applied to all of them.

Assessing how well interpretability methods apply to architectures beyond those for which they were developed, and whether we can develop techniques that generalize effectively across architectures remain open questions. This is especially important due to the recent success of other competitive architectures as alternatives to CNNs and transformers. Notable alternatives include diffusion models (Sohl-Dickstein et al., 2015; Rombach et al., 2022) and Vision Transformers (Dosovitskiy et al., 2021) for image generation/classification and RWKV (Peng et al., 2023) and later state space models (SSMs) (Gu & Dao, 2024) for language modeling. Recent studies show that certain methods transfer from CNNs to SSMs (Paulo et al., 2024), and from transformers to some SSMs e.g. (Meng et al. (2022b) vs. Sharma et al. (2024a)) and (Wang et al. (2023) vs. Ensign & Garriga-Alonso (2024)).

Beyond the transferability of interpretability methods, a related open question concerns the transferability of conclusions across model families. The overwhelming majority of mechanistic interpretability research focuses on the transformer model family and therefore does not distinguish between observations that are model-specific and those that are not. Consequently, we may be overlooking valuable insights that could be obtained by comparing the results across multiple model families. These insights may, for example, evince or refute the 'universality hypothesis', which states that (Li et al., 2015; Olah et al., 2020b) different neural networks learn similar features and circuits to one another.

3.7 Human computer interaction with model internals

As we saw in Section 3.3, the ability to control and understand neural networks is tightly linked. Thus, mechanistic interpretability has great potential to facilitate new types of human-AI interaction. Systems amenable to human comprehension and control would allow diverse users to intuitively manipulate and

interact with them based on their preferences, greatly broadening their utility. As a starting point, AI engineers who build and test AI systems have an obvious interest in working with the internal workings of neural nets. If experts could visualize and interact with internal representations, it would unlock obvious benefits for scientific research.

However, interactive tooling has a much broader constituency, including policy-focused AI auditors, as well as end users. Consider an auditor looking for bias or safety issues in a neural net. With a way to probe the network directly instead of relying on testing behavior, the auditor can be much more successful in discovering potential low-probability but high-stakes errors. For end users, transparency into a network facilitates appropriate calibration of trust. One recent idea proposes a dashboard that can show, in real time, the internal features that influence a chatbot's answers during a text chat (Chen et al., 2024; Zou et al., 2023a; Viégas & Wattenberg, 2023). Such a dashboard might help users spot AI errors, or display warnings if safety-relevant features activate. More generally, results from mechanistic interpretability could be used to blend direct-manipulation interfaces with text interfaces, providing users with a richer palette of controls (Carter & Nielsen, 2017).

4 Open socio-technical problems in mechanistic interpretability

Effective practical application of mechanistic interpretability brings both technical challenges and complex social ramifications. It could enable us to act on AI policy and governance, presenting a valuable opportunity to implement regulatory standards and social ideals through technical means (Section 4.1). Such consequential impacts inevitably come with important social and philosophical considerations, which likewise require rigorous inquiry if we are to fully realize the potential benefits of AI (Section 4.2).

4.1 Translating technical progress in mechanistic interpretability into levers for AI policy and governance

Current frontier AI governance efforts rarely specify concrete ways in which a mechanistic understanding of AI models might be used to help implementation. For example, OpenAI's Preparedness Framework commits to mitigating biological risks posed by its AI systems (OpenAI, 2023), but the framework lacks details on specific measures that might be taken. Progress in interpretability could potentially enable the removal of any knowledge from the model which could aid users in creating biological weapons (Li et al., 2024b). However, it remains uncertain how technical progress will translate into better AI governance, largely due to the numerous open technical problems in mechanistic interpretability. However, there are several promising routes toward better levers for AI policy and governance.

These avenues include assisting companies and governments to identify risks through evaluations and enhancing forecasts about new AI developments; more thorough oversight of AI systems in deployment; simplifying how AI systems operate within existing liability law through clearer explanations of AI decisions; enabling governments to establish risk mitigation regulations and companies to commit to concrete mitigation commitments; and protecting copyright law.

An understanding of model internals can help AI labs assess the risks from frontier models (Chang et al., 2024; Shevlane et al., 2023; Casper et al., 2024) and thus better fulfill their obligations under the EU AI Act to "perform model evaluation ... with a view to identifying and mitigating systemic risk" (Council of the European Union, 2024). More specifically, a mechanistic understanding of models could help evaluators elicit dangerous capabilities via improved finetuning (United Kingdom AI Safety Institute, 2024), guide their adversarial red-team attempts (Tong et al., 2024a), and ensure that AI systems are not intentionally underperforming evaluations (van der Weij et al., 2024). Advancements in mechanistic interpretability could also assist companies and governments in anticipating when or if AI models will obtain specific dangerous capabilities. Improved forecasting capabilities could enhance threat modeling by reducing the "reasonable disagreement amongst experts over which risks to prioritize" (Anthropic, 2024). For instance, it may help build consensus on whether language models are just stochastic parrots (Bender et al., 2021) or if they have coherent world models (Li et al., 2023). We also might be able to use interpretability to build evidence supporting or refuting different models of catastrophic threat, such as determining the validity of

mesa-optimization (Hubinger et al., 2021) or inner alignment concerns (Carlsmith, 2023). It would also give additional time for companies to prepare adequate risk mitigation measures, (OpenAI, 2023), and for governments to establish appropriate guidance or regulation (UK Government, 2022).

The EU AI Act mandates that developers of General Purpose AI models with systemic risk have obligations to report incidents involving their system to the AI office. Interpretability tools have the potential to continuously monitor AI inference and detect incidents that require reporting. Compared to incidents in other domains (for instance, nuclear security), AI systems allow us to log all inputs and system states, even those that may lead to catastrophic harm. Access to a small number of such data points may greatly improve our ability to mitigate similar future failures (Greenblatt & Shlegeris, 2024). For example, interpretability could be used to investigate the critical "features" of the input that led to the AI incident. This could improve our ability to further red-team the system (Section 3.3) and generate more similar data points that could result in similar incidents. This, in turn, may help us reduce the likelihood of future incidents (Chan et al., 2024). The incident could be utilized more directly in the construction of test-time monitors to detect similar future incidents (see Roger (2023)). More speculatively, interpretability could be used to verify companies' compliance with domestic regulation (O'Brien et al., 2023a), or to authenticate states' compliance with future international treaties regarding the use of AI (Aarne et al., 2024).

Leveraging a mechanistic understanding of model internals could also make decision rationales for AI model outputs more easily obtainable. This could aid in enforcing citizens' rights under the EU General Data Protection Regulation "to obtain an explanation of the decision reached" by a system "based solely on automated processing" (GDP, 2016; Gilpin et al., 2019). Model editing tools could also resolve problems regarding the copyright status of existing generative models (Grynbaum & Mac, 2023). According to the US Copyright Act (United States Congress), for any copyrighted work, an artifact "from which the work can be perceived, reproduced, or otherwise communicated. . . with the aid of a machine or device" is considered a copy of the work (Lee et al., 2024). Interpretability tools could help detect and remove memorized works that can be reproduced verbatim by generative models.

4.2 Social and philosophical problems in mechanistic interpretability

The ability to interpret advanced AI systems holds immense potential to advance the science of AI and increase our ability to control it (Critch & Krueger, 2020; Tegmark & Omohundro, 2023; Dalrymple et al., 2024). In this paper, we provide an overview of various interpretability tools that offer novel insights. Nonetheless, interpretability research has thus far produced few tools that are used to make state-of-the-art systems safer in the real world (Rauker et al., 2023). Modern AI systems are still generally trained, evaluated, monitored, and debugged using techniques that do not rely on understanding their internal workings.

The absence of paradigmatic clarity is a major socio-technical factor for this. Questions such as which goals the field of interpretability should pursue, how success should be graded, and how we should define interpretability warrant more thoughtful answers than exist at present. In interpretable AI research, the motivations and methods employed are often described as "diverse and occasionally discordant" (Lipton, 2018). At times, the objective of AI interpretability research is articulated as advancing a fundamental "understanding" or "uncovering the true essence" (Christensen & Cheney, 2015) of what is happening inside black-box models. The intrinsic validity of this paradigm deserves philosophical inquiry.

However, from an engineer's perspective, pursuing "understanding" without a practical downstream application misses an engineer's objective. Proponents of this view may contend that quantifiable benchmarks linked to concrete practical goals are accurate measures of the success of interpretability. This motivation is concrete and useful, but some have criticized interpretability research as artificially limiting the solution space to engineering problems. When interpretability tools are studied with motivations such as fairness or safety, Krishnan (2020) argues that, "Since addressing these problems need not involve something that looks like an 'interpretation' (etc.) of an algorithm, the focus on interpretability artificially constrains the solution space by characterizing one possible solution as the problem itself." Thus, some argue that interpretability research has failed to produce competitive techniques (Rauker et al., 2023; Casper et al., 2022b), omitted non-interpretability baselines (Rudin, 2019; Krishnan, 2020), and graded interpretability tools on their own curve (Doshi-Velez & Kim, 2017; Miller, 2019; Rauker et al., 2023). In all safety-relevant applications of

mechanistic interpretability, it is important to assess the usefulness of interpretability against alternative methodologies. Failing to do so or misrepresenting these comparisons can lead to follow-up work that rests on false assumptions (Lipton & Steinhardt, 2019; Leech et al., 2024).This is particularly problematic when it impacts the efforts of critical safety work.

Despite these concerns, it is not always imperative for mechanistic interpretability to strictly outperform uninterpretable baselines: It may still be helpful to develop methods that offer interpretability-based advantages, along with the benefits of more competitive uninterpretable methods (e.g. Rimsky et al. (2024)). While interpretability-based methods do not currently outperform black-box baselines, if they perform in a similar ballpark, further progress in mechanistic interpretability could soon lead to better methods, especially in problems which we believe might disproportionately benefit from improved mechanistic understanding (Section 3).

Another potential reason for lack of clarity is that models are often studied in a vacuum: Smart & Kasirzadeh (2024) emphasize that the usefulness or correctness of model interpretations can depend on the broader context of the model's development or deployment. For example, it may be hard to identify representations of fairness within models, since understanding this requires an understanding of broader contexts. The same data may lead to different conclusions under different definitions of fairness.

Finally, the interpretability community must exercise caution in their communication to minimize potential abuses of the results of its work. Unfortunately, selective transparency can be used to actively mislead (Ananny & Crawford, 2018). Furthermore, interpretability is at risk of being used for purposes that might serve corporate interests at the potential expense of safety. The field of AI interpretability is highly influenced by research teams in the private sector. On one hand, industry resources and research contributions have significantly advanced interpretability research. On the other hand, compared to academia, corporations publish selectively, often have financial conflicts of interest, and may provide limited transparency. Meanwhile, the recent definite — but ultimately modest — progress in mechanistic interpretability has been used to lobby against specific AI regulation by falsely claiming that, "Although advocates for AI safety guidelines often allude to the 'black box' nature of AI models, where the logic behind their conclusions is not transparent, recent advancements in the AI sector have resolved this issue, thereby ensuring the integrity of open-source code models." (Andreessen Horowitz, 2023).

5 Conclusion

While mechanistic interpretability has made meaningful progress in both methods and applications, significant challenges remain before we can achieve many of the field's ambitious goals.

The path forward requires progress along multiple axes. We would benefit from stronger theoretical foundations for decomposing neural networks at the joints of their generalization structure. Current methods like sparse dictionary learning, while promising, face both practical limitations in scaling to larger models and deeper conceptual challenges regarding their underlying assumptions. We must also develop more robust methods for validating our interpretations of model behavior, moving beyond correlation-based descriptions to capture true causal mechanisms. Additionally, we need better techniques to understand how mechanisms evolve during training and how they interact to produce complex behaviors.

These methodological advances could unlock several promising applications. Improved interpretability methods could enable more effective monitoring of potential risks, better control over model behavior, and more accurate predictions of capabilities. For AI capabilities, mechanistic understanding could lead to more efficient architectures, better training procedures, and more targeted ways to enhance model performance. In various scientific domains, microscope AI approaches could help extract valuable insights from model internals. To achieve these diverse goals, the field must ensure a focus on generating insights that have real-world utility. This will involve establishing better benchmarks and comparing interpretability-based approaches to non-interpretability baselines. However, fastest progress will likely come from the field pursuing both scientific and engineering goals simultaneously, rather than one at the expense of the other.

The practical impact of progress in mechanistic interpretability extends beyond technical achievements. Interpretability tools could provide crucial mechanisms for governance and oversight. They could help verify

compliance with safety standards, detect potential risks before deployment, and provide clearer attribution of model decisions. However, realizing these benefits will require careful attention to the risks of potential misuse and of giving false assurance about AI safety.

Looking toward the future, many expect current AI capabilities to be only a foretaste of what is to come. As AI capabilities advance, the need for a mechanistic understanding of their decision-making processes becomes increasingly urgent. While the black-box nature of AI models remains unresolved, the untapped potential of mechanistic interpretability is what makes it such an exciting research area, and highlights the importance of solving its many open research problems.

References

Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance), 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679.

Onni Aarne, Tim Fist, and Caleb Withers. Secure, governable chips. Report, Center for a New American Security, 2024. URL https://www.cnas.org/publications/reports/secure-governable-chips.

Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 66–88. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/abid22a.html.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 700–712. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/075b051ec3d22dac7b33f788da631fd4-Paper.pdf.

Micah Adler and Nir Shavit. On the complexity of neural computation in superposition, 2024. URL https://arxiv.org/abs/2409.15318.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models, 2024. URL https://arxiv.org/abs/2402.04614.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4699–4711. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018. doi: 10.1177/1461444816676645. URL https://doi.org/10.1177/1461444816676645.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 39–48, 2016.

Andreessen Horowitz. Written evidence to the UK House of Lords Communications and Digital Select Committee inquiry: Large language models (LLM0114), 2023. URL https://committees.parliament.uk/writtenevidence/127070/pdf/. Written evidence submitted to the UK Parliament.

Anthropic. Giving claude a role with a system prompt, 2024. URL https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/system-prompts.

Anthropic. Responsible scaling policy. Policy document, Anthropic, 2024. URL https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=oVTkOs8Pka. Survey Certification, Expert Certification.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Tomer Ashuach, Martin Tutek, and Yonatan Belinkov. Revs: Unlearning sensitive information in language models via rank editing in the vocabulary space, 2024. URL https://arxiv.org/abs/2406.09325.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *ACL*, 2023. URL https://arxiv.org/abs/2305.18029.

Kola Ayonrinde, Michael T. Pearce, and Lee Sharkey. Interpretability as compression: Reconsidering sae explanations of neural activations with mdl-saes, 2024. URL https://arxiv.org/abs/2410.11179.

Mikita Balesni, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy Scheurer, Charlotte Stix, Rusheb Shah, Nicholas Goldowsky-Dill, Dan Braun, Bilal Chughtai, Owain Evans, Daniel Kokotajlo, and Lucius Bushnaq. Towards evaluations-based safety cases for ai scheming, 2024a. URL https://arxiv.org/abs/2411.03336.

Mikita Balesni, Tomek Korbak, and Owain Evans. The two-hop curse: Llms trained on a->b, b->c fail to learn a–>c, 2024b. URL https://arxiv.org/abs/2411.16353.

Randall Balestriero and richard baraniuk. A spline theory of deep learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 374–383. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/balestriero18b.html.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907375117. URL https://www.pnas.org/content/early/2020/08/31/1907375117.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48 (1):207–219, March 2022a. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48 (1):207–219, March 2022b. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

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, 2023a. URL https://arxiv.org/abs/2303.08112.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d066d21c619d0a78c5b557fa3291a8f4-Paper-Conference.pdf.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries, 2024. URL https://arxiv.org/abs/2406.12775.

Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. Interpreting neural networks through the polytope lens, 2022. URL https://arxiv.org/abs/2211.12312.

Joseph Bloom and Johnny Lin. Understanding sae features with the logit lens - ai alignment forum, Mar 2024a. URL https://www.alignmentforum.org/posts/qykrYY6rXXM7EEs8Q/understanding-sae-features-with-the-logit-lens.

Joseph Isaac Bloom and Johnny Lin. Understanding SAE features with the logit lens. *Alignment Forum*, March 2024b. URL https://www.alignmentforum.org/posts/qykrYY6rXXM7EEs8Q/understanding-sae-features-with-the-logit-lens.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert, 2021. URL https://arxiv.org/abs/2104.07143.

Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QO9-y8also-.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7txPaUpUnc.

Leo Breiman. *Classification and regression trees*. Routledge, 1984.

Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfMWhAqYQ.

Trenton Bricken, Oct 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Trenton Bricken, Siddharth Mishra-Sharma, Jonathan Marcus, Adam Jermyn, Christopher Olah, Kelley Rivoire, and Thomas Henighan. Stage-wise model diffing. *Transformer Circuits*, 2024. URL https://transformer-circuits.pub/2024/model-diffing/index.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Lucius Bushnaq and Jake Mendel. Circuits in superposition: Compressing many small neural networks into one. *Alignment Forum*, oct 2024. URL https://www.alignmentforum.org/posts/roE7SHjFWEoMcGZKd/circuits-in-superposition-compressing-many-small-neural.

Lucius Bushnaq, Stefan Heimersheim, Nicholas Goldowsky-Dill, Dan Braun, Jake Mendel, Kaarel Hänni, Avery Griffin, Jörn Stöhler, Magdalena Wache, and Marius Hobbhahn. The local interaction basis: Identifying computationally-relevant and sparsely interacting features in neural networks, 2024. URL https://arxiv.org/abs/2405.10928.

Bart Bussman, Michael Pearce, Patrick Leask, Joseph Bloom, Lee Sharkey, and Neel Nanda. Showing sae latents are not atomic using meta-saes, Aug 2024. URL https://www.lesswrong.com/posts/TMAmHh4DdMr4nCSr5/showing-sae-latents-are-not-atomic-using-meta-saes.

S.R. Cajal. *Estructura de los centros nerviosos de las aves (1888)*. Jiménez y Molina, 1924. URL https://books.google.co.uk/books?id=SXr8sgEACAAJ.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Joe Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power?, 2023. URL https://arxiv.org/abs/2311.08379.

Shan Carter and Michael Nielsen. Using artificial intelligence to augment human intelligence. *Distill*, 2017. doi: 10.23915/distill.00009. https://distill.pub/2017/aia.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust feature-level adversaries are interpretability tools. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33093–33106. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d616a353c711f11c722e3f28d2d9e956-Paper-Conference.pdf.

Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust feature-level adversaries are interpretability tools. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33093–33106. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d616a353c711f11c722e3f28d2d9e956-Paper-Conference.pdf.

Stephen Casper, Tong Bu, Yuxiao Li, Jiawei Li, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. Red teaming deep neural networks with feature synthesis tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=Od6CHhPM7I.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Stephen Casper, Kaivalya Hariharan, and Dylan Hadfield-Menell. Diagnostics for deep neural networks with automated copy/paste attacks, 2023c. URL https://arxiv.org/abs/2211.10024.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, volume 35 of *FAccT '24*, pp. 2254–2272. ACM, June 2024. doi: 10.1145/3630106.3659037. URL http://dx.doi.org/10.1145/3630106.3659037.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features, 2024. URL https://arxiv.org/abs/2411.02193.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into ai agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 958–973, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658948. URL https://doi.org/10.1145/3630106.3658948.

Lawrence Chan, Adrià Garriga-alonso, and Nicholas Goldowsky-Dill. Causal scrubbing: A method for rigorously testing interpretability hypotheses [redwood research] - ai alignment forum, Oct 2022a. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Lawrence Chan, Adrià Garriga-alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Tao Lin, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: Results on induction heads - ai alignment forum, Dec 2022b. URL https://www.alignmentforum.org/posts/j6s9H9SHrEhEfuJnq/causal-scrubbing-results-on-induction-heads.

Lawrence Chan, Adri'a Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Tao Lin, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: results on a paren balance checker. *Alignment Forum*, 2023. URL https://www.alignmentforum.org/s/h95ayYYwMebGEYN5y/p/kjudfaQazMmC74SbF.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders, 2024. URL https://arxiv.org/abs/2409.14507.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: self-interpretation of large language model embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational ai, 2024. URL https://arxiv.org/abs/2406.07882.

Lars Thøger Christensen and George Cheney. Peering into transparency: Challenging ideals, proxies, and organizational practices. *Communication Theory*, 25(1):70–90, February 2015. doi: 10.1111/comt.12052. URL https://doi.org/10.1111/comt.12052.

Paul F Christiano. Can we efficiently distinguish different mechanisms?, Dec 2022. URL https://www.alignmentforum.org/posts/JLyWP2Y9LAruR2gi9/can-we-efficiently-distinguish-different-mechanisms.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Bilal Chughtai and Lucius Bushnaq. Activation space interpretability may be doomed. *LessWrong*, 1 2025. URL https://www.lesswrong.com/posts/gYfpPbww3wQRaxAFD/activation-space-interpretability-may-be-doomed.

Bilal Chughtai and Yeu-Tong Lau. Understanding positional features in layer 0 SAEs. *LessWrong*, 2024. URL https://www.lesswrong.com/posts/ctGeJGHg9pbc8memF/understanding-positional-features-in-layer-0-saes.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Mark M. Churchland and Krishna V. Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of Neurophysiology*, 97(6):4235–4257, 6 2007. doi: 10.1152/jn.00095.2007. URL https://journals.physiology.org/doi/full/10.1152/jn.00095.2007. PubMed ID: 17376854.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks, 2024. URL https://arxiv.org/abs/2410.04332.

Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. Safety cases: How to justify the safety of advanced ai systems, 2024. URL https://arxiv.org/abs/2403.10462.

Paul Colognese and Arun Jose. High-level interpretability: detecting an AI's objectives. *Alignment Forum*, September 2023. URL https://www.alignmentforum.org/posts/tFYGdq9ivjA3rdaS2/high-level-interpretability-detecting-an-ai-s-objectives.

Arthur Conmy and Neel Nanda. Activation steering with SAEs. *Alignment Forum*, 2024. URL https://www.alignmentforum.org/posts/C5KAZQib3bzzpeyrg/full-post-progress-update-1-from-the-gdm-mech-interp-team#Activation_Steering_with_SAEs.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Jack Cooper, In Hwa Um, Ognjen Arandjelović, and David J Harrison. Lymphocyte classification from hoechst stained slides with deep learning. *Cancers*, 14(23):5957, Dec 2022. doi: 10.3390/cancers14235957. URL https://doi.org/10.3390/cancers14235957. PMID: 36497439; PMCID: PMC9738034.

Council of the European Union. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2024. URL https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

Andrew Critch and David Krueger. Ai research considerations for human existential safety (arches), 2020. URL https://arxiv.org/abs/2006.04948.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 88–105, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19836-6.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=aedFIIRRfXr.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7uVcpu-gMD.

Róbert Csordás, Christopher Potts, Christopher D. Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations, 2024. URL https://arxiv.org/abs/2408.10920.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024. URL https://arxiv.org/abs/2405.06624.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33016309. URL https://doi.org/10.1609/aaai.v33i01.33016309.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893.

Adam Davies and Ashkan Khakzar. The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms, 2024. URL https://arxiv.org/abs/2408.05859.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata, 2023. URL https://arxiv.org/abs/2307.03637.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3243–3255, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.262. URL https://aclanthology.org/2020.emnlp-main.262.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights?, 2024. URL https://arxiv.org/abs/2410.08827.

Jean-Stanislas Denain and Jacob Steinhardt. Auditing visualizations: Transparency methods struggle to detect anomalous behavior, 2023. URL https://arxiv.org/abs/2206.13498.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations, 2024. URL https://arxiv.org/abs/2303.09435.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4a-Paper.pdf.

Jonathan Donnelly and Adam Roegiest. On interpretability and feature representations: An analysis of the sentiment neuron. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (eds.), *Advances in Information Retrieval*, pp. 795–802, Cham, 2019. Springer International Publishing. ISBN 978-3-030-15712-8.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=J6zHcScAo0.

Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, Oliver Rausch, Saffron Huang, Sam Bowman, Stuart Ritchie, Tom Henighan, and Deep Ganguli. Evaluating feature steering: A case study in mitigating social biases, 2024. URL https://anthropic.com/research/evaluating-feature-steering.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 03 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00359. URL https://doi.org/10.1162/tacl_a_00359.

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.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. Softmax linear units. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/solu/index.html.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear, 2024a. URL https://arxiv.org/abs/2405.14860.

Joshua Engels, Logan Riggs, and Max Tegmark. Decomposing the dark matter of sparse autoencoders, 2024b. URL https://arxiv.org/abs/2410.14670.

Danielle Ensign and Adrià Garriga-Alonso. Investigating the indirect object identification circuit in mamba, 2024. URL https://arxiv.org/abs/2407.14008.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, University of Montreal*, 01 2009.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 134–139, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2524. URL https://aclanthology.org/W16-2524.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models, 2024. URL https://arxiv.org/abs/2410.19278.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1407. URL https://aclanthology.org/D18-1407.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models, 2024. URL https://arxiv.org/abs/2405.00208.

Jakob N. Foerster, Justin Gilmer, Jascha Sohl-Dickstein, Jan Chorowski, and David Sussillo. Input switched affine networks: an rnn architecture designed for interpretability. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1136–1145. JMLR.org, 2017.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. doi: 10.1109/iccv.2017.371. URL http://dx.doi.org/10.1109/ICCV.2017.371.

Jessica Zosa Forde, Charles Lovering, George Konidaris, Ellie Pavlick, and Michael L. Littman. Where, when & which concepts does alphazero learn? lessons from the game of hex. Unpublished manuscript, 2023.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Timo Freiesleben and Gunnar König. Dear xai community, we need to talk! In Luca Longo (ed.), *Explainable Artificial Intelligence*, pp. 48–65, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44064-9.

Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models, 2024. URL https://openreview.net/forum?id=v675Iyu0ta.

Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=COZDy0WYGg.

Pranav Gade, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Badllama: cheaply removing safety fine-tuning from llama 2-chat 13b, 2024. URL https://arxiv.org/abs/2311.00117.

Ella M. Gale, Nicholas Martin, Ryan Blything, Anh Nguyen, and Jeffrey S. Bowers. Are there any 'object detectors' in the hidden layers of cnns trained to identify objects or scenes? *Vision Research*, 176: 60–71, 2020. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2020.06.007. URL https://www.sciencedirect.com/science/article/pii/S0042698920301140.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=5Ca9sSzuDp.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip, 2024b. URL https://arxiv.org/abs/2406.04341.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 163–173, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL https://aclanthology.org/2020.blackboxnlp-1.16/.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

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. In *Machine Learning Research*, 2024a. URL https://arxiv.org/abs/2301.04709.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024b. Curran Associates Inc. ISBN 9781713845393.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024c. URL https://proceedings.mlr.press/v236/geiger24a.html.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Roland S. Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un)reliability of feature visualizations. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In Wanxiang Che and Ekaterina Shutova (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 12–21, Abu Dhabi, UAE, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.2. URL https://aclanthology.org/2022.emnlp-demos.2/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *ICML*, 2024. URL https://arxiv.org/abs/2401.06102.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: a unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5922–5932. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/41c542dfe6e4fc3deb251d64cf6ed2e4-Paper.pdf.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, Jul. 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://ojs.aaai.org/index.php/AAAI/article/view/4252.

Leilani H. Gilpin, Cecilia Testart, Nathaniel Fruchter, and Julius Adebayo. Explaining explanations to society. In *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*, 2019. URL https://arxiv.org/abs/1901.06560.

Arthur Goemans, Marie Davidsen Buhl, Jonas Schuett, Tomek Korbak, Jessica Wang, Benjamin Hilton, and Geoffrey Irving. Safety case template for frontier ai: A cyber inability argument, 2024. URL https://arxiv.org/abs/2411.08088.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023. URL https://arxiv.org/abs/2304.05969.

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.232. URL https://aclanthology.org/2024.emnlp-main.232/.

Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 70–88, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533074. URL https://doi.org/10.1145/3531146.3533074.

Ryan Greenblatt and Buck Shlegeris. Catching AIs red-handed. *Alignment Forum*, January 2024. URL https://www.alignmentforum.org/posts/i2nmBfCXnadeGmhzW/catching-ais-red-handed.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=zzOOqD6R1b.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers, 2024. URL https://arxiv.org/abs/2403.17887.

Jason Gross, Rajashree Agrawal, Thomas Kwa, Euan Ong, Chun Hei Yip, Alex Gibson, Soufiane Noubir, and Lawrence Chan. Compact proofs of model performance via mechanistic interpretability. In *ICML Workshop on Mechanistic Interpretability*, 2024. URL https://arxiv.org/abs/2406.11779.

Roger Grosse. Three sketches of ASL-4 safety case components, November 2024. URL https://alignment.anthropic.com/2024/safety-cases/.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large language model generalization with influence functions, 2023. URL https://arxiv.org/abs/2308.03296.

Michael M. Grynbaum and Ryan Mac. The Times sues OpenAI and Microsoft over A.I. use of copyrighted work. *The New York Times*, December 2023. URL https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Clément Guerner, Anej Svete, Tianyu Liu, Alexander Warstadt, and Ryan Cotterell. A geometric notion of causal probing, 2024. URL https://arxiv.org/abs/2307.15054.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Robust unlearning via mechanistic localizations. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=06pNzrEjnH.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 12–21, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1002. URL https://aclanthology.org/D15-1002.

Rohan Gupta, Iván Arcuschin, Thomas Kwa, and Adrià Garriga-Alonso. Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=R9gR9MPuD5.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Kaarel Haani, Rio Popper, and Jake Mendel. A starting point for making sense of task structure (in machine learning). *LessWrong*, February 2024. URL https://www.lesswrong.com/posts/exp4JGPJu46g6sdRp/a-starting-point-for-making-sense-of-task-structure-in.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1135–1143, Cambridge, MA, USA, 2015. MIT Press.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=p4PckNQR8k.

Kaarel Hänni, Jake Mendel, Dmitry Vaintrob, and Lawrence Chan. Mathematical models of computation in superposition. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=0cVJP8kClR.

Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL https://doi.org/10.1080/00437956.1954.11659520.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EldbUlZtbd.

Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Xu Owen He. Mixture of a million experts, 2024. URL https://arxiv.org/abs/2407.04153.

Stefan Heimersheim. You can remove gpt2's layernorm by fine-tuning, 2024. URL https://arxiv.org/abs/2409.13710.

Stefan Heimersheim and Jett Janiak. A circuit for Python docstrings in a 4-layer attention-only transformer. *Alignment Forum*, February 2023. URL https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/.

Tom Henighan. Caloric and the utility of incorrect theories. *Transformer Circuits*, 2024. URL https://transformer-circuits.pub/2024/april-update/index.html#caloric-theory.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep features. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=NudBMY-tzDr.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019. URL https://arxiv.org/abs/1909.03368.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419.

John Hewitt, John Thickstun, Christopher Manning, and Percy Liang. Backpack language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9103–9125, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.506. URL https://aclanthology.org/2023.acl-long.506/.

Steven A. Hicks, Jonas L. Isaksen, Vajira Thambawita, et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific Reports*, 11:10949, 2021. doi: 10.1038/s41598-021-90285-5. URL https://doi.org/10.1038/s41598-021-90285-5.

Jacob Hilton, Nick Cammarata, Shan Carter, Gabriel Goh, and Chris Olah. Understanding rl vision. *Distill*, 2020. doi: 10.23915/distill.00029. https://distill.pub/2020/understanding-rl-vision.

Geoffrey Hinton. Shape representation in parallel systems. *Proceedings of teh Seventh International Joint Conference on Artificial Intelligence*, 1981. URL https://www.cs.toronto.edu/~hinton/absps/shape81.pdf.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 07 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.7.1527. URL https://doi.org/10.1162/neco.2006.18.7.1527.

Oskar John Hollinsworth, Curt Tigges, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks*

for NLP, pp. 58–87, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5. URL https://aclanthology.org/2024.blackboxnlp-1.5/.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic evaluation of unlearning using parametric knowledge traces, 2024. URL https://arxiv.org/abs/2406.11614.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. The developmental landscape of in-context learning, 2024. URL https://arxiv.org/abs/2402.02364.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance, 2021. URL https://arxiv.org/abs/1905.03151.

Xiyang Hu, Cynthia Rudin, and Margo I. Seltzer. Optimal sparse decision trees. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:139104649.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 317–331, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.24. URL https://aclanthology.org/2023.blackboxnlp-1.24/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Xinting Huang, Madhur Panwar, Navin Goyal, and Michael Hahn. Inversionview: A general-purpose method for reading information from neural activations. In *ICML Interpretability Workshop*, 2024b. URL https://arxiv.org/abs/2405.17653.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Evan Hubinger. Towards understanding-based safety evaluations. *Alignment Forum*, March 2023. URL https://www.alignmentforum.org/posts/uqAdqrvxqGqeBHjTP/towards-understanding-based-safety-evaluations.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2021. URL https://arxiv.org/abs/1906.01820.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Dieuwke Hupkes and Willem Zuidema. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 5617–5621. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/796. URL https://doi.org/10.24963/ijcai.2018/796.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Alon Jacovi. Trends in explainable ai (xai) literature, 2023. URL https://arxiv.org/abs/2301.05433.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minne@inproceedingskatz-etal-2024-backward, title = "Backward Lens: Projecting Language Model Gradients into the Vocabulary Space", author = "Katz, Shahar and Belinkov, Yonatan and Geva, Mor and Wolf, Lior", editor = "Al-Onaizan, Yaser and Bansal, Mohit and Chen, Yun-Nung", booktitle = "Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing", month = nov, year = "2024", address = "Miami, Florida, USA", publisher = "Association for Computational Linguistics", url = "https://aclanthology.org/2024.emnlp-main.142/", doi = "10.18653/v1/2024.emnlp-main.142", pages = "2390–2422", abstract = "Understanding how Transformer-based Language Models (LMs) learn and recall information is a key goal of the deep learning community. Recent interpretability methods project weights and hidden states obtained from the forward pass to the models' vocabularies, helping to uncover how information flows within LMs. In this work, we extend this methodology to LMs' backward pass and gradients. We first prove that a gradient matrix can be cast as a low-rank linear combination of its forward and backward passes' inputs. We then develop methods to project these gradients into vocabulary items and explore the mechanics of how new information is stored in the LMs' neurons." sota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Jett Janiak, Chris Mathwin, and Stefan Heimersheim. Polysemantic attention head in a 4-layer transformer. LessWrong Blog, 2023. URL https://www.lesswrong.com/posts/nuJFTS5iiJKT5G5yh/polysemantic-attention-head-in-a-4-layer-transformer.

Adam Jermyn, Chris Olah, and Tom Henighan. Attention head superposition. *Transformer Circuits*, 2023. URL https://transformer-circuits.pub/2023/may-update/index.html#attention-superposition.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1193–1215, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.70.

David Johnston, Arkajyoti Chakraborty, and Nora Belrose. Mechanistic anomaly detection research update, Aug 2024. URL https://blog.eleuther.ai/mad_research_update/.

Caden Juang, Gonccedilalo Paulo, Jacob Drori, and Nora Belrose. Open source automated interpretability for sparse autoencoder features, Jul 2024. URL https://blog.eleuther.ai/autointerp/.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin

Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Subhash Kantamneni, Josh Engels, Senthooran Rajamanoharan, and Neel Nanda. SAE probing: What is it good for? absolutely something! *AI Alignment Forum*, nov 2024. URL https://www.lesswrong.com/posts/NMLq8yoTecAF44KX9/sae-probing-what-is-it-good-for-absolutely-something.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. Backward lens: Projecting language model gradients into the vocabulary space. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2390–2422, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.142. URL https://aclanthology.org/2024.emnlp-main.142/.

Dmitrii Kharlapenko, neverix, Neel Nanda, and Arthur Conmy. Self-explaining SAE features. *Alignment Forum*, August 2024. URL https://www.alignmentforum.org/posts/8ev6coxChSWcxDy8/self-explaining-sae-features.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/kim18d.html.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018b. URL https://proceedings.mlr.press/v80/kim18d.html.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pp. 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Nathalie Maria Kirch, Severin Field, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks, 2024. URL https://arxiv.org/abs/2411.03343.

Louis Kirsch, Julius Kunze, and David Barber. Modular networks: learning to decompose neural computation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2414–2423, Red Hook, NY, USA, 2018. Curran Associates Inc.

Connor Kissane, Arthur Conmy, and Neel Nanda. Attention output saes improve circuit analysis - ai alignment forum, Jun 2024a. URL https://www.alignmentforum.org/posts/EGvtgB7ctifzxZg6v/attention-output-saes-improve-circuit-analysis.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders, 2024b. URL https://arxiv.org/abs/2406.17759.

Connor Kissane, Robert Krzyzanowski, Neel Nanda, and Arthur Conmy. SAEs are highly dataset dependent: a case study on the refusal direction. *Alignment Forum*, nov 2024c. URL https://www.alignmentforum.org/posts/rtp6n7Z23uJpEH7od/saes-are-highly-dataset-dependent-a-case-study-on-the.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Arne Köhn. What's in an embedding? analyzing word embeddings through multilingual evaluation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2067–2073, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1246. URL https://aclanthology.org/D15-1246.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Edward Korot, Nikolas Pontikos, Xiaoxuan Liu, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Scientific Reports*, 11:10286, 2021. doi: 10.1038/s41598-021-89743-x. URL https://doi.org/10.1038/s41598-021-89743-x.

Eliza Kosoy, Emily Rose Reagan, Leslie Lai, Alison Gopnik, and Danielle Krettek Cobb. Comparing machines and children: Using developmental psychology experiments to assess the strengths and weaknesses of lamda responses, 2023. URL https://arxiv.org/abs/2305.11243.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745.

Mukund Krishnan. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502, 2020. doi: 10.1007/s13347-019-00372-9. URL https://doi.org/10.1007/s13347-019-00372-9.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pp. 84–90, may 2012. doi: 10.1145/3065386.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=UnWhcpIyUC.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 507–514, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Matthew L. Leavitt and Ari Morcos. Towards falsifiable interpretability research, 2020. URL https://arxiv.org/abs/2010.12016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning, May 2015. URL https://www.nature.com/articles/nature14539.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: a case study on dpo and toxicity. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'bout ai generation: Copyright and the generative-ai supply chain. *Journal of the Copyright Society of the USA*, 2024. URL https://arxiv.org/abs/2309.08133.

Gavin Leech, Juan J. Vazquez, Misha Yagudin, Niclas Kupper, and Laurence Aitchison. Questionable practices in machine learning, 2024. URL https://arxiv.org/abs/2407.12220.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2024. URL https://arxiv.org/abs/2310.20624.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2024a. URL https://arxiv.org/abs/2306.03341.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024b.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL https://proceedings.mlr.press/v44/li15convergent.html.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla, 2023. URL https://arxiv.org/abs/2307.09458.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. Towards understanding jailbreak attacks in LLMs: A representation space analysis. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7067–7085, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.401. URL https://aclanthology.org/2024.emnlp-main.401/.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Alexander Rives, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 3 2023. doi: 10.1126/science.ade2574.

David Lindner, Janos Kramar, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=tbbId8u7nP.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html. * Equal contribution.

Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://doi.org/10.1145/3236386.3241340.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, February 2019. ISSN 1542-7730. doi: 10.1145/3317287.3328534. URL https://doi.org/10.1145/3317287.3328534.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models, 2024a. URL https://arxiv.org/abs/2402.08787.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Machine unlearning in generative ai: A survey, 2024b. URL https://arxiv.org/abs/2407.20516.

Ziming Liu, Eric Gan, and Max Tegmark. Seeing is believing: Brain-inspired modular training for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2305.08746.

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

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. Working Paper 31017, National Bureau of Economic Research, March 2023. URL http://www.nber.org/papers/w31017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms, 2024. URL https://arxiv.org/abs/2402.16835.

Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. Interpretability needs a new paradigm, 2024. URL https://arxiv.org/abs/2405.05386.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014. URL https://arxiv.org/abs/1412.0035.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL https://arxiv.org/abs/2311.17030.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=MHIX9H8aYF.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2013. URL https://arxiv.org/abs/1312.5663.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503:78–84, 2013. doi: 10.1038/nature12742. URL https://doi.org/10.1038/nature12742.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

Chris Mathwin, Dennis Akar, and Lee Sharkey. Gated attention blocks: Preliminary progress toward removing attention head superposition. *LessWrong*, apr 2024. URL https://www.lesswrong.com/posts/kzc3qNMsP2xJcxhGn/gated-attention-blocks-preliminary-progress-toward-removing-1.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=5Hd6813x3U.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022. doi: 10.1073/pnas.2206625119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2206625119.

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.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024. URL https://arxiv.org/abs/2412.04984.

Jake Mendel. SAE feature geometry is outside the superposition hypothesis. *Alignment Forum*, 2024. URL https://www.alignmentforum.org/posts/MFBTjb2qf3ziWmzz6/sae-feature-geometry-is-outside-the-superposition-hypothesis.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022a. arXiv:2202.05262.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=-h6WAS6eE4.

Eric J Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3tbTw2ga8K.

Eric J. Michaud, Isaac Liao, Vedang Lad, Ziming Liu, Anish Mudide, Chloe Loughridge, Zifan Carl Guo, Tara Rezaei Kheirkhah, Mateja Vukelić, and Max Tegmark. Opening the ai black box: program synthesis via mechanistic interpretability, 2024. URL https://arxiv.org/abs/2402.05110.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit evaluation metrics are not robust. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zSf8PJyQb2.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9:2383, 2018. doi: 10.1038/s41467-018-04316-3. URL https://doi.org/10.1038/s41467-018-04316-3.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJGCiw5gl.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models, 2021. URL https://arxiv.org/abs/2007.04131.

Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery*, 38:2903–2941, 2024. doi: 10.1007/s10618-022-00901-9. URL https://doi.org/10.1007/s10618-022-00901-9.

Mordvintsev. Inceptionism: Going deeper into neural networks, 2015. URL https://research.google/blog/inceptionism-going-deeper-into-neural-networks/.

Marius Mosbach, Vagrant Gautam, Tomás Vergara-Browne, Dietrich Klakow, and Mor Geva. From insights to actions: The impact of interpretability and analysis research on nlp, 2024. URL https://arxiv.org/abs/2406.12618.

Michael C Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/07e1cd7dca89a1678042477183b7ac3f-Paper.pdf.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17153–17163. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c74956ffb38ba48ed6ce977af6727275-Paper.pdf.

Aaron Mueller. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks, 2024. URL https://arxiv.org/abs/2407.04690.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability, 2024. URL https://arxiv.org/abs/2408.01416.

Neel Nanda. Attribution patching: Activation patching at industrial scale, Mar 2023a. URL https://www.neelnanda.io/mechanistic-interpretability/attribution-patching.

Neel Nanda. Othello-GPT: Reflections on the research process. *Alignment Forum*, March 2023b. URL https://www.alignmentforum.org/posts/TAz44Lb9n9yf52pv8/othello-gpt-reflections-on-the-research-process.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=9XFSbDPmdW.

Neel Nanda, Senthooran Rajamanoharan, J'anos Kram'ar, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level (Post 1). *Alignment Forum*, December 2023b. URL https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall.

Arunachalam Narayanaswamy, Subhashini Venugopalan, Dale R. Webster, Lily Peng, Greg S. Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Michael Brenner, Philip C. Nelson, and Avinash V. Varadarajan. Scientific discovery by generating counterfactuals using image translation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 273–283, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3395–3403, Red Hook, NY, USA, 2016a. Curran Associates Inc. ISBN 9781510838819.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks, 2016b. URL https://arxiv.org/abs/1602.03616.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks, 2016c. URL https://arxiv.org/abs/1602.03616.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug and play generative networks: Conditional iterative generation of images in latent space, 2017. URL https://arxiv.org/abs/1612.00005.

Nostalgebraist. Interpreting gpt: The logit lens - ai alignment forum, Aug 2020. URL https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Joe O'Brien, Shaun Ee, and Zoe Williams. Deployment corrections: An incident response framework for frontier ai models, 2023a. URL https://arxiv.org/abs/2310.00328.

Thomas O'Brien, James Stremmel, Lola Pio-Lopez, Paul McMillen, Cody Rasmussen-Ivey, and Michael Levin. Machine learning for hypothesis generation in biology and medicine: Exploring the latent space of neuroscience and developmental bioelectricity. Preprint, September 2023b. URL https://doi.org/10.31219/osf.io/269e5.

Chris Olah. Interpretability dreams. *Transformer Circuits*, May 2023. URL https://transformer-circuits.pub/2023/interpretability-dreams/index.html. An informal note on future goals for mechanistic interpretability.

Chris Olah, 2024. URL https://transformer-circuits.pub/2022/mech-interp-essay/index.html.

Chris Olah and Adam Jermyn. What is a linear representation? what is a multidimensional feature? *Transformer Circuits*, 2024. URL https://transformer-circuits.pub/2024/july-update/index.html#linear-representations.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11), Nov 2017a. doi: 10.23915/distill.00007.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017b. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 2020a. doi: 10.23915/distill.00024.002. https://distill.pub/2020/circuits/early-vision.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020b. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL `https://arxiv.org/abs/2209.11895`.

OpenAI. OpenAI preparedness framework. Framework, OpenAI, 2023. URL `https://cdn.openai.com/openai-preparedness-framework-beta.pdf`. Beta version.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner,

Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=owuEcT6BTl.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023a. URL https://openreview.net/forum?id=T0PoOJg8cK.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024b. URL https://openreview.net/forum?id=KXuYjuBzKo.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions, 2023b. URL https://arxiv.org/abs/2308.14752.

Gonçalo Paulo, Thomas Marshall, and Nora Belrose. Does transformer interpretability transfer to rnns?, 2024. URL https://arxiv.org/abs/2404.05971.

Michael T. Pearce, Thomas Dooms, and Alice Rigg. Weight-based decomposition: A case for bilinear mlps, 2024. URL https://arxiv.org/abs/2406.03947.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936/.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.225. URL https://aclanthology.org/2022.emnlp-main.225/.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1179. URL https://aclanthology.org/D18-1179/.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420.

Nicholas Pochinkov and Nandi Schoots. Dissecting language models: Machine unlearning via selective pruning, 2024. URL https://arxiv.org/abs/2403.01267.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *Proceedings of the 2024 International Conference on Learning Representations*, 2024. arXiv:2402.14811.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to deceive with attention-based explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4782–4793, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.432. URL https://aclanthology.org/2020.acl-main.432/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2018. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2024. URL https://arxiv.org/abs/2407.02646.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=zLBlin2zvW.

David Raposo, Matthew T. Kaufman, and Anne K. Churchland. A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, 17:1784–1792, 2014. doi: 10.1038/nn.3865. URL https://doi.org/10.1038/nn.3865.

Tilman Rauker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks . In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 464–483, Los Alamitos, CA, USA, February 2023. IEEE Computer Society. doi: 10.1109/SaTML54575.2023.00039. URL https://doi.ieeecomputersociety.org/10.1109/SaTML54575.2023.00039.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Logan Riggs, Sam Mitchell, and Adam Kaufman. Finding sparse linear connections between features in LLMs. *AI Alignment Forum*, dec 2023. URL https://www.lesswrong.com/posts/7fxusXdkMNmAhkAfc/finding-sparse-linear-connections-between-features-in-llms.

Mattia Rigotti, Omri Barak, Melissa Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497:585–590, 2013. doi: 10.1038/nature12160. URL https://doi.org/10.1038/nature12160.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press, May 2022. ISBN 9781316519332. doi: 10.1017/9781009023405. URL http://dx.doi.org/10.1017/9781009023405.

Fabien Roger. Coup probes: Catching catastrophes with probes trained off-policy. *Alignment Forum*, November 2023. URL https://www.alignmentforum.org/posts/WCj7WgFSLmyKaMwPR/coup-probes-catching-catastrophes-with-probes-trained-off.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11 (2):305–345, 1999. doi: 10.1162/089976699300016674.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Naomi Saphra and Sarah Wiegreffe. Mechanistic?, 2024. URL https://arxiv.org/abs/2410.09087.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure, 2024. URL https://arxiv.org/abs/2311.07590.

Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the human-ai knowledge gap: Concept discovery and transfer in alphazero, 2023. URL https://arxiv.org/abs/2310.16410.

Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. FIND: A function description benchmark for evaluating interpretability methods. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=mkSDXjX6EM.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. doi: 10.1007/s11263-019-01228-7.

Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry. Toward verified artificial intelligence. *Commun. ACM*, 65(7):46–55, June 2022. ISSN 0001-0782. doi: 10.1145/3503914. URL https://doi.org/10.1145/3503914.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

L. Shapley. *7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317.*, pp. 69–79. Princeton University Press, Princeton, 1997. ISBN 9781400829156. doi: doi: 10.1515/9781400829156-012. URL https://doi.org/10.1515/9781400829156-012.

Lee Sharkey. A technical note on bilinear layers for interpretability, 2023. URL https://arxiv.org/abs/2305.03452.

Lee Sharkey, Sid Black, and Beren Millidge. Current themes in mechanistic interpretability research, Nov 2022a. URL https://www.lesswrong.com/posts/Jgs7LQwmvErxR9BCC/current-themes-in-mechanistic-interpretability-research#Study_model_systems_in_depth.

Lee Sharkey, Dan Braun, and Beren. [interim research report] taking features out of superposition with sparse autoencoders - ai alignment forum, Dec 2022b. URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition.

Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in mamba. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=yoVRyrEgix.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=tvhaxkMKAn.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

C. S. Sherrington. Observations on the scratch-reflex in the spinal dog. *The Journal of Physiology*, 34(1-2): 1–50, 3 1906. doi: 10.1113/jphysiol.1906.sp001139. URL https://physoc.onlinelibrary.wiley.com/doi/10.1113/jphysiol.1906.sp001139. 185 citations as of [insert current date].

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. Model evaluation for extreme risks, 2023. URL https://arxiv.org/abs/2305.15324.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in llms, 2024. URL https://arxiv.org/abs/2410.13032.

Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. Language models are better than humans at next-token prediction, 2024. URL https://arxiv.org/abs/2212.11281.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014a. URL https://arxiv.org/abs/1312.6034.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014b. URL https://arxiv.org/abs/1312.6034.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 62–75. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/009c434cab57de48a31f6b669e7ba266-Paper.pdf.

Andrew Smart and Atoosa Kasirzadeh. Beyond model interpretability: socio-structural explanations in machine learning. *AI and SOCIETY*, September 2024. ISSN 1435-5655. doi: 10.1007/s00146-024-02056-1. URL http://dx.doi.org/10.1007/s00146-024-02056-1.

Lewis Smith. The "strong" feature hypothesis could be wrong, Aug 2024a. URL https://www.lesswrong.com/posts/tojtPCCRpKLSHBdpn/the-strong-feature-hypothesis-could-be-wrong.

Lewis Smith. Feature is overloaded terminology. LessWrong, 2024b. Available at: https://www.lesswrong.com/posts/9Nkb389gidsozY9Tf/lewis-smith-s-shortform?commentId=fd64ALuWK8rXdLKz6.

P. Smolensky. *Neural and conceptual interpretation of PDP models*, pp. 390–431. MIT Press, Cambridge, MA, USA, 1986. ISBN 0262631105.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Alex Tamkin, Mohammad Taufeeque, and Noah D. Goodman. Codebook features: sparse and discrete interpretability for neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar,

Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen

Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama,

Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Max Tegmark and Steve Omohundro. Provably safe systems: the only path to controllable agi, 2023. URL https://arxiv.org/abs/2309.01933.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Jacques Thibodeau. But is it really in rome? an investigation of the rome model editing technique - ai alignment forum, Dec 2022. URL https://www.alignmentforum.org/posts/QL7J9wmS6W2fWpofd/but-is-it-really-in-rome-an-investigation-of-the-rome-model.

Hannes Thurnherr and Jérémy Scheurer. Tracrbench: Generating interpretability testbeds with large language models. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=vNubZ5zK8h.

Demian Till. Do sparse autoencoders find "true features"? *Less Wrong*, 2024. URL https://www.lesswrong.com/posts/QoR8noAB3Mp2KBA4B/do-sparse-autoencoders-find-true-features.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG. arXiv:2310.15213.

Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024a. Curran Associates Inc.

Siyuan Tong, Kaiwen Mao, Zhe Huang, et al. Automating psychological hypothesis generation with ai: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11:896, 2024b. doi: 10.1057/s41599-024-03407-5. URL https://doi.org/10.1057/s41599-024-03407-5.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

UK Government. National AI strategy. Government strategy, Government of the United Kingdom, 2022. URL https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version.

United Kingdom AI Safety Institute. Early lessons from evaluating frontier AI systems. Technical report, United Kingdom AI Safety Institute, 2024. URL https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems.

United States Congress. Copyright law of the united states (title 17) and related laws contained in title 17 of the united states code. URL https://www.copyright.gov/title17/. This publication contains the text of Title 17 of the United States Code, including all amendments enacted by Congress through

December 23, 2022. It includes the Copyright Act of 1976 and all subsequent amendments to copyright law; the Semiconductor Chip Protection Act of 1984, as amended; and the Vessel Hull Design Protection Act, as amended.

Guillermo Valle-Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rye4g3AqFm.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. Ai sandbagging: Language models can strategically underperform on evaluations, 2024. URL https://arxiv.org/abs/2406.07358.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Fernanda Viégas and Martin Wattenberg. The system model and the user model: Exploring ai dashboard design, 2023. URL https://arxiv.org/abs/2305.02469.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1288–1301, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.75.

Chelsea Voss, Nick Cammarata, Gabriel Goh, Michael Petrov, Ludwig Schubert, Ben Egan, Swee Kiat Lim, and Chris Olah. Visualizing weights. *Distill*, 2021. doi: 10.23915/distill.00024.007. https://distill.pub/2020/circuits/visualizing-weights.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024a. URL https://openreview.net/forum?id=ns8IH5Sn5y.

George Wang, Matthew Farrugia-Roberts, Jesse Hoogland, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape geometry reveals stagewise development of transformers. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024b. URL https://openreview.net/forum?id=2JabyZjM5H.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3), November 2024c. ISSN 0360-0300. doi: 10.1145/3698590. URL https://doi.org/10.1145/3698590.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review, 2024d. URL https://arxiv.org/abs/2403.10415.

Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. Honesty is the best policy: Defining and mitigating ai deception. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 2313–2341. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/06fc7ae4a11a7eb5e20fe018db6c036f-Paper-Conference.pdf.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040/.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009.

Matthew Watkins. Mapping the semantic void: Strange goings-on in gpt embedding spaces, Dec 2023. URL https://www.lesswrong.com/posts/c6uTNm5erRrmyJvvD/mapping-the-semantic-void-strange-goings-on-in-gpt-embedding.

David S. Watson. Conceptual challenges for interpretable machine learning. *Synthese*, 200:65, 2022. doi: 10.1007/s11229-022-03485-5. URL https://doi.org/10.1007/s11229-022-03485-5.

Donglai Wei, Bolei Zhou, Antonio Torrabla, and William Freeman. Understanding intra-class knowledge inside cnn, 2015. URL https://arxiv.org/abs/1507.02379.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 10473–10486, 2023. doi: 10.1109/TNNLS.2022.3167409.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11080–11090. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/weiss21a.html.

Tichakorn Wongpiromsarn, Mahsa Ghasemi, Murat Cubuktepe, Georgios Bakirtzis, Steven Carr, Mustafa O. Karabag, Cyrus Neary, Parham Gohari, and Ufuk Topcu. Formal methods for autonomous systems. *Foundations and Trends® in Systems and Control*, 10(3–4):180–407, 2023. ISSN 2325-6826. doi: 10.1561/2600000029. URL http://dx.doi.org/10.1561/2600000029.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Wilson Wu, Louis Jaburi, Jacob Drori, and Jason Gross. Unifying and verifying mechanistic interpretations: A case study with group operations, 2024a. URL https://arxiv.org/abs/2410.07476.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=nRfClnMhVX.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=fykjplMc0V.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to makelov et al. (2023)'s "interpretability illusion" arguments, 2024c. URL https://arxiv.org/abs/2401.12631.

Keith Wynroe and Lee Sharkey. Decomposing the QK circuit with bilinear sparse dictionary learning. *AI Alignment Forum*, jul 2024. URL https://www.lesswrong.com/posts/2ep6FGjTQoGDRnhrq/decomposing-the-qk-circuit-with-bilinear-sparse-dictionary.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=WHqVVk3UHr.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30378–30392. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c402501846f9fe03e2cac015b3f0e6b1-Paper-Conference.pdf.

Adam Yedidia. Gpt-2's positional embedding matrix is a helix, July 2023. URL https://www.lesswrong.com/posts/qvWP3aBDBaqXvPNhS/gpt-2-s-positional-embedding-matrix-is-a-helix.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9615–9625, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.840.

Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015. URL http://arxiv.org/abs/1506.06579.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7943–7956, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.466. URL https://aclanthology.org/2024.findings-emnlp.466/.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9924–9959, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL https://aclanthology.org/2023.emnlp-main.615.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić (eds.), *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 1–10, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.deelio-1.1. URL https://aclanthology.org/2021.deelio-1.1/.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Enyan Zhang, Michael A. Lepori, and Ellie Pavlick. Instilling inductive biases with subnetworks, 2024. URL https://arxiv.org/abs/2310.10899.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In *Proceedings of the USENIX Security Symposium (Security)*, 2020.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. In *ICLR*, 2015. URL http://arxiv.org/abs/1412.6856.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9623–9633, Jun. 2022. doi: 10.1609/aaai.v36i9.21196. URL https://ojs.aaai.org/index.php/AAAI/article/view/21196.

Roland Simon Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S. A. Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of CNN activations? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vLPqnPf9k0.

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, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b. URL https://arxiv.org/abs/2307.15043.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL https://arxiv.org/abs/2406.04313.

A Appendix

B Summary of open questions

B.1 Open problems in mechanistic interpretability methods and foundations

B.1.1 Reverse engineering: Identifying the roles of network components

Reverse engineering step 1: Neural network decomposition

1. How should we decompose networks into more interpretable constituent parts?

   (a) What isomorphism or what approximation of a neural network (or parts of it) is the best way to express it for the purposes of interpreting it?
   (b) How should we coarse grain neural networks?
   (c) How should we build higher level abstractions on top of low-level network components?

2. How true is the linear representation hypothesis?
   (a) To what extent do models encode concepts linearly in their representations?
   (b) How should we characterize representations that are not linearly represented in neural networks?
   (c) What properties of a concept, or of the training distribution, result in a particular concept becoming encoded linearly (or not)?

3. Is the combination of the linear representation hypothesis and superposition the right frame for thinking about computation in neural networks?
   (a) Can we fully determine the causes of feature superposition and polysemanticity within neural networks?
   (b) How should we understand superposition in attention blocks?
   (c) How should we understand cross-layer superposition?
   (d) What new theoretical insights can be gleaned from considering how networks perform computation natively in superposition, rather than treating superposition purely as a compression strategy?

4. Can the problems with SDL be overcome?
   (a) What lies in SDL reconstruction errors? Will the errors converge to zero with methodological progress?
   (b) Is sparsity the correct proxy for interpretability?
   (c) Can the approach be scaled to the largest models?
   (d) Does SDL make sense if we don't believe in the linear representation hypothesis?
   (e) Is sparsity the best possible proxy for interpretability?
   (f) Is it correct to think of SDL features as 'bags of features', or is there important information contained within the geometry of representation space?
   (g) If SDL finds compositions of the "true" features, is this a problem?
   (h) Can SDL be applied to all architectural components to fully decompose networks?
   (i) How can we better measure the success of SDL techniques?
   (j) How should we connect sparsely activating features into circuits? Will this be practically possible, and the best possible description of network mechanisms?
   (k) Can we develop new methods that address conceptual and practical issues with SDL?

5. How important is the geometry of activation space for explaining neural network behavior?
   (a) How can we identify the underlying functional structure of networks (which defines why activations are located in particular geometric arrangements in activation space)?
   (b) Must we understand global feature geometry or only local feature geometry in order to understand computation in neural networks?

6. Can we connect theories for how neural networks generalize to interpretability?
   (a) Can we distinguish parts of networks that underlie generalization from parts that underlie memorization?
   (b) What mechanisms underlie the relationship between interpretability and generalization?
   (c) Are there connections between adversarial robustness and superposition?
   (d) Can we connect interpretability to theories of deep learning like SLT?

7. Can we build intrinsically more interpretable models at low performance cost? How helpful is this?

(a) Interpretability training: Can we train networks that are interpretable by default at low performance cost?
(b) Interpretable inference: Can we convert already-trained models into forms that are much easier to completely interpret at little performance cost?
(c) How can we train large-scale models such that the concepts they use are naturally understandable to humans?
(d) How can we design training objectives so that the model is incentivized to use known specific abstractions?
(e) How can we localize concepts we want to control (e.g., long-term plans) during training?

Reverse engineering step 2: Identifying the functional role of components

1. Can we improve on max-activating input data set examples for understanding the causes of network component activations?
   (a) How can we avoid imposing human bias to explanations?
   (b) Can we progress toward deeper descriptions based on internal mechanisms?
   (c) How might we develop interpretation methods that can recognize and work with unfamiliar concepts - computational patterns that don't map cleanly to human intuitions?
2. How can we develop attribution methods that faithfully and efficiently compute which network components are important for some downstream metric?
   (a) How can we develop attribution methods that capture higher-order effects beyond first-order approximations of model behavior?
   (b) Is it possible to create perturbation-based methods that don't force models to operate outside their training distribution?
   (c) Can we develop hybrid approaches that combine the strengths of different attribution methods while mitigating their individual weaknesses?
3. How can we better measure the downstream effects of model components?
   (a) How can we reliably distinguish between true causal pathways and compensatory effects like the "Hydra effect" when performing interventions?

Reverse engineering step 3: Validation of descriptions

1. Can we improve our ability to validate mechanistic explanations for model behavior in ways that do not depend on researcher intuition and are computationally tractable to use?
   (a) Can we improve on methodologies for evaluating hypotheses through their predictive power on activations of network components?
   (b) Can we develop methodologies for evaluating hypotheses through their predictive power for model behavior (e.g. unusual failures or adv examples)?
   (c) Can we develop methodologies for handcoding weights that faithfully represent some hypotheses, as a drop in replacement for subnetworks we claim to understand?
   (d) Can we develop a wider suite of networks with known ground truth explanations to validate techniques against?
   (e) Can we use mechanistic explanations to achieve engineering goals?
   (f) Can we use mechanistic explanations to achieve engineering goals in a way that improves upon black box baselines?
2. Can we develop "model organisms" as a community, which are understood deeply, and seen as a test-bed for new unproven interpretability methodologies to be tested?

3. Can we establish standardized baselines and benchmarks for comparing different interpretability approaches on real-world, non-cherry-picked tasks, where the ground truth is known?
4. What would constitute a comprehensive set of "stress tests" for interpretability hypotheses that could reliably detect interpretability illusions?
5. How might we design evaluation frameworks that assess interpretability methods on their average case and worst-case performance rather than just best-case scenarios?
6. How can we ensure that our understanding of internals generalizes to out-of-distribution inputs?

B.1.2 Concept-based interpretability: Identifying network components for given roles

1. How can we reliably distinguish causal from merely correlated features when probing neural networks?
2. Can we develop automated systems to generate high-quality probing data sets, reducing the current heavy reliance on human effort?
3. What regularization and validation techniques can be used to prevent spurious correlations while ensuring probes find generalizable features?
4. How can we improve probing for concepts that may not have clear positive/negative examples?

B.1.3 Proceduralizing mechanistic interpretability into circuit discovery pipelines

1. Can we develop techniques that build on lower level methods that provide deeper or more complete insights about neural networks?
2. How much can we learn from further work in the existing circuit discovery paradigm?
   (a) Should we expect circuit discovery to benefit from further methodological progress in decomposing neural networks? Will faithfulness go up and explanation description length go down?
   (b) Can we remove the constraint that task definition for circuit discovery is inherently concept-based, which may be complicating mechanistic analysis?
   (c) Can we get around the practical issues of negative and backup behavior?
   (d) Will circuit discovery provide insights into arbitrary tasks, or will it only be helpful in cases where we are able to crisply define tasks?

B.1.4 Automating steps in mechanistic interpretability research

1. Can we improve on AI automated feature description and validation methods?
   (a) Through automating the generation and testing of arbitrary hypotheses?
   (b) Through describing differences between features?
   (c) Through descriptions of how components interact?
2. Can we improve on ACDC-like circuit discovery methods?
3. Can we automate other parts of the mechanistic interpretability pipeline?
   (a) Conceptual interpretability research?
   (b) Decomposition method discovery?
   (c) More ad hoc validation of hypotheses?
4. Should we take steps to mitigate potentially misaligned AI systems sabotaging AI automated interpretability?

B.2 Open problems in applications of mechanistic interpretability

B.2.1 Using mechanistic interpretability for better monitoring and auditing of AI systems for potentially unsafe cognition

1. Can we effectively use interpretability for safety evaluations?
   (a) Can we develop robust "white box" evaluations that detect concerning internal patterns without needing to understand the entire network?
   (b) Can we reliably distinguish between features that merely recognize deceptive behavior versus mechanisms that generate deceptive behavior?
   (c) How can we validate that learned features capture all concerning patterns of reasoning?
   (d) Can we reliably identify which features are appropriately versus spuriously relevant to a given task?
2. Can we leverage interpretability to enhance red-teaming and system testing?
   (a) Can we use interpretability insights to make red-teaming more efficient than current methods?
   (b) How can we best use feature attribution to help human red-teamers identify problematic input patterns?
3. Can we develop effective test-time monitoring systems based on interpretability?
   (a) Can we get mechanistic anomaly detection to work?
   (b) Can we create passive monitoring systems based on model internals that effectively flag concerning internal patterns during deployment?
   (c) Can we develop monitoring systems that work with only feature-level understanding rather than requiring deep mechanical insights?

B.2.2 Using mechanistic interpretability for better control of AI system behavior

1. Can we improve steering methods through interpretability?
   (a) How can we make activation steering more precise and reduce its side effects?
   (b) Can we develop methods to steer entire mechanisms rather than just single features?
2. Can we achieve reliable model unlearning and editing?
   (a) Will carving the network at its true joints help us improve on model unlearning and editing?
   (b) Can mechanistic interpretability help us develop better methods for evaluating unlearning efficacy?
   (c) Can mechanistic interpretability help us determine which classes of model edit are even possible, without damaging generalization in undesirable ways?
3. Can we better understand and improve finetuning through interpretability?
   (a) Can we make finetuning more sample-efficient by targeting specific parameters?
   (b) Can we develop better tools for analyzing feature-level or mechanism-level differences between model versions?

B.2.3 Using mechanistic interpretability for better predictions about AI systems

1. Can we predict model behavior in novel situations outside of the distribution of inputs we have access to with mechanistic understanding?
   (a) Can we reliably predict when and how jailbreaking or safety bypasses might occur?
   (b) How can we identify internal signatures that predict specific failure modes like hallucination?
   (c) Can we develop methods to predict model behavior without requiring behavioral evaluations?

   (d) Can we find "values" or "goals" in systems that might be indicative of behavior in more generality?
   (e) Is it possible to prove the absence of specific dangerous capabilities through mechanistic analysis?

2. Can we develop formal verification methods for AI systems?
   (a) Can current toy model verification approaches scale to frontier systems?
   (b) How much of neural computation can be reduced to verifiable symbolic operations?
   (c) Can we create formal guarantees about system behavior in complex, non-formalizable environments?
   (d) What level of mechanistic understanding is necessary for meaningful formal verification?

3. Can we make rigorous claims about model safety?
   (a) Can we definitively prove the absence of specific dangerous mechanisms?
   (b) How can we verify claims about model values and goals in a rigorous way?
   (c) What types of safety claims are possible with current interpretability methods?
   (d) Can we develop "enumerative safety" approaches that reliably identify all relevant mechanisms?

4. Can we better predict AI capability development through interpretability?
   (a) Can we identify early signatures that predict emergent capabilities?
   (b) How do model mechanisms evolve dynamically during training?
   (c) Can we map the connection between small-scale circuits and large-scale capabilities?
   (d) How does the loss landscape's structure relate to capability emergence?

5. Can we understand the relationship between training data and capabilities?
   (a) How do specific training examples influence the development of model mechanisms?
   (b) Can we predict model limitations based on training data composition?
   (c) Can we design training data sets to reliably produce specific desired capabilities?
   (d) How does data set structure affect the balance between in-context and weights-based learning?

6. Can we predict latent or maskable capabilities?
   (a) How can we identify capabilities that could be 'unlocked' through prompting or finetuning?
   (b) Can we detect when finetuning has masked rather than removed capabilities?
   (c) How do we analyze mechanisms that span multiple timesteps or sequential behaviors?
   (d) Can we predict which model capabilities are fundamental versus superficially trained?

B.2.4 Using mechanistic interpretability to improve our ability to perform inference, improve training and make use of learned representations

1. Can we use interpretability to make inference more efficient?
   (a) How can we identify skippable computations without affecting outputs?
   (b) Can we create more effective distillation methods through mechanistic understanding?
   (c) How can we optimize model architecture based on component function analysis?
   (d) Can we identify and optimize critical computational pathways?

2. Can we improve training through mechanistic insights?
   (a) Can we better select training data by understanding example influence?
   (b) How can we monitor and optimize capability emergence during training?
   (c) Can we develop more parameter-efficient training methods through component analysis?
   (d) Can we create better architectures through component understanding?

   (e) Can we identify and enhance components with specific functionalities?

3. Can we instill capabilities directly into networks?
   (a) Can we design better inductive biases based on mechanistic insights?
   (b) Is it possible to create modular architectures with swappable components?
   (c) Can we develop reliable methods for combining model parameters?
   (d) Is it possible to transfer specific capabilities between models?

B.2.5 Using mechanistic interpretability for 'microscope AI'

1. Can we leverage AI models for scientific discovery?
   (a) How can we extract novel patterns and predictors that models have found?
   (b) Can we make microscope AI techniques accessible to domain experts?
   (c) How do we validate scientific insights derived from model interpretability?
   (d) Can we extend microscope AI beyond current simple correlational discoveries?

2. Can we develop better knowledge extraction methods?
   (a) How can we detect when models have found genuinely novel patterns?
   (b) Can we automate the process of finding scientific insights in model weights?
   (c) How do we bridge the gap between model features and scientific concepts?
   (d) Can we make these techniques usable without deep machine learning expertise?

B.2.6 Mechanistic interpretability on a broader range of models and model families

1. Can interpretability methods generalize across architectures?
   (a) Do current interpretability methods (SDL, circuit analysis) transfer to SSMs? Or, like the transition from CNNs to transformers, are new approaches necessary?
   (b) Which insights are model-specific versus universal?
   (c) How can we adapt methods for multimodal models?

2. How do different models trained on similar data compare mechanistically?
   (a) Is the "universality hypothesis" true across models? To what extent do neural networks learn similar features and circuits to each other (and to humans?)
   (b) Do different architectures learn fundamentally different features?
   (c) How do mechanisms of particular tasks differ between transformers, CNNs, and SSMs?
   (d) Are there insights we can gain from comparing architectures?

3. Can we future-proof interpretability research?
   (a) How can we prepare for interpreting novel architectures?
   (b) Should we focus on architecture-specific or general methods?
   (c) Can we identify truly fundamental interpretability principles?
   (d) Will current methods work on future frontier models?

B.2.7 Human computer interaction with model internals

1. Can we create interfaces that use mechanistic understanding to enhance human-neural network interaction?
   (a) How can we visualize model internals in an intuitive way?
   (b) Can we develop real-time interpretability dashboards?
   (c) What's the right balance between simplicity and depth in these interfaces?

   (d) How do we make complex model mechanisms understandable to non-experts?

2. Can we develop interpretability tools to help auditors?
   (a) How can we help auditors find potential failure modes directly?
   (b) Can we develop tools to detect bias at the mechanism level?
   (c) What interfaces would make auditing more efficient and thorough?
   (d) How can we present technical findings to policy makers?

3. Can we improve end-user interaction with AI?
   (a) How can transparency features help users calibrate trust?
   (b) Can we create intuitive controls based on model mechanisms?
   (c) Can we create intuitive ways to steer model behavior?

B.2.8 Governance

1. Can mechanistic analysis help identify and prevent failures?
   (a) Can we identify specific mechanisms that caused AI failures?
   (b) How do we map the causal chain of mechanisms leading to incidents?
   (c) Can we detect when similar mechanisms are about to activate?
   (d) Is it possible to isolate and modify failure-causing mechanisms?

2. Can we study mechanism patterns related to governance?
   (a) Can we identify mechanisms responsible for specific dangerous capabilities?
   (b) How do we detect deceptive or evasive mechanisms?
   (c) Can we map the mechanisms involved in model decision-making?
   (d) Is it possible to verify the absence of specific harmful mechanisms?

3. Can mechanistic insights verify compliance?
   (a) How can we trace decision mechanisms to explain model outputs?
   (b) Can we identify mechanisms that process copyrighted content?
   (c) Is it possible to detect mechanisms that encode specific knowledge?
   (d) How do we verify modifications to problematic mechanisms?

B.2.9 Open socio-technical problems in mechanistic interpretability

Translating technical progress in mechanistic interpretability into levers for AI policy and governance

1. Can we use a mechanistic understanding to better evaluate AI capabilities?
   (a) How can we use interpretability to improve capability elicitation?
   (b) Can we use interpretability to reliably detect when models are strategically underperforming capabilities evaluations?

2. Can we use a mechanistic understanding to improve our ability to forecast when or whether new capabilities will arise ahead of time?

3. How can we use interpretability to better estimate the likelihood of different threat models?

4. Can we use interpretability to prevent AI incidents?
   (a) Can we use interpretability to construct reliable test-time monitors to detect AI incidents?

   (b) Can we use reliably prevent similar incidents in the future, by using interpretability to design new evaluation tasks on incident scenarios?

5. Can interpretability help verify which workloads GPUs are being used for?

6. How should interpretability inform copyright law?

7. How can mechanistic understanding help resolve copyright challenges in generative AI, particularly regarding the detection and removal of memorized copyrighted works?

Social and philosophical challenges in mechanistic interpretability

1. What is interpretability?
   (a) What are the goals of the field?
   (b) How should success be graded?
   (c) Should we treat interpretability as a science or an engineering discipline? What implications does this have on what research should be done?

2. How can we mitigate downside risks of interpretability research?
   (a) How can we communicate the results of our research such that the risk of their misuse is minimized?