

From Interpretability Methods to Interpretable Models

Julien Colin^{1,2}, Nuria Oliver^{1*} and Thomas Serre^{2,3*}

¹ ELLIS Alicante, Spain

² Carney Institute for Brain Science, Brown University, USA

³ Department of Cognitive and Psychological Sciences, Brown University, USA

Abstract. More than a decade in, explainable AI (XAI) for computer vision has assembled a mature toolbox: attribution, feature visualization, concept-based, and circuit-based methods. Yet almost all of the field’s effort has gone into building and comparing these *methods*, and little into the question they were meant to answer—how interpretable are our *models*, and are we making progress as they evolve? We argue for shifting the field’s focus from methods to models, along two complementary lines. One is already within reach: existing tools let us characterize and compare what different models represent and compute. The other is harder, and largely neglected: whether a model can actually be understood by the humans who rely on it—the independent evaluators on whom trust and certification depend, not the experts confirming what they already expect. It can only be measured, not inferred. We review why the toolbox is mature enough to support both, survey the thin body of work comparing models, draw a parallel to systems neuroscience, and close with a model-centric XAI agenda.

1 Introduction

Artificial intelligence now drives decisions in high-stakes domains: medical image interpretation [1,2,3], autonomous driving [4,5], and robotic control [6]. These systems are accurate enough to deploy, yet remain largely black boxes to the people who rely on them. Understanding them nonetheless matters: to catch failures before they cause harm [7,8], to meet emerging legal expectations of explanation [9], and to learn from models that make new scientific discoveries [10,11]. As models grow more capable and more widely deployed, the gap between what they do and what we understand widens [12], making interpretability a pressing need [13]. These pressures are beginning to acquire regulatory teeth: interpretability is increasingly named in certification frameworks—most explicitly by EASA, which lists “AI explainability” as a trustworthiness building block for aviation [14], and by the EU AI Act, whose high-risk provisions require systems transparent enough for humans to interpret and oversee their outputs [15]—an expectation that, while not yet a hard requirement, rises in proportion to risk.

* Equal senior contribution. E-mail: julien.colin@brown.edu

What interpretability actually means, though, has never been settled: the term has been used for many distinct properties [16,17], and what they share is that it is ultimately human-centric—an explanation counts only if a person can actually understand it [18]. In computer vision, one effort has been to make models interpretable by design [19]⁴: building transparency into the architecture itself—prototype networks that classify by comparison to visualizable exemplars [20], concept bottleneck models that route predictions through human-named concepts [21], or self-explaining networks that emit an explanation as part of the forward pass [22]—trading architectural flexibility for transparency. A separate effort—explainable AI (XAI) [23], the focus of this paper—instead treats a trained model as a black box and explains its predictions after the fact. The line between the two is blurrier than it looks: designing a model to be explainable does not, on its own, make it interpretable.

Over the past decade and a half, XAI has produced several complementary families of methods—attribution, feature visualization, concept-based, and, more recently, circuit-based. Much of the community’s effort, including work using human evaluations to ask how well a method lets us understand a model [24,25,26], went into making these methods better and comparing them against one another, rather than into using them to explain and compare models.

This paper takes a different position. We do not claim that XAI methods are perfect, nor that method development should stop—only that the toolbox is now rich enough to redirect much of the field’s effort from *how well a method explains a model* to *how interpretable a model is, and how that changes as models evolve*. That shift has two sides: comparing what different models represent and compute, which existing tools already support, and measuring whether an independent evaluator can actually understand them—which expert analysis alone cannot establish. The moment is right: methods are mature enough to build on (Section 2), the model landscape is consolidating around a few families (Section 3), and—as systems neuroscience showed, building decades of insight with far cruder instruments—progress follows from studying the object rather than waiting for perfect tools. Yet remarkably little work turns the toolbox on models themselves (Section 4), leaving the field’s original question open; we close with an agenda for a model-centric XAI (Section 5).

2 The XAI toolbox has matured

The case for studying models rather than methods rests on a premise: the methods must be good enough to build on. This section argues that they are. We review the toolbox through three established families of XAI methods in computer vision, together with a fourth, more recent one that targets a model’s computation rather than its representations. Each answers a different interpretability

⁴ We retain *interpretable by design* as it is the term used in the literature. But since we take interpretable to refer to a model that can be understood by humans, these methods are, in our view, closer to explainable by design: they guarantee that an explanation is available, not that the model can be understood.

question, and each has matured enough to serve as a reliable instrument for comparing models, even if none is a complete account of how a model reasons.

Attribution methods. Given a model, an input, and its prediction, which pixels drove this decision? This is perhaps the most fundamental interpretability question in computer vision, and the first to be systematically addressed by XAI through attribution methods [27]. Their goal is to identify the image regions most responsible for a model’s prediction, and a wide range of approaches has been developed over the past decade to do so: gradient-based [27,28,29,30], activation-based [31,32], perturbation-based [33,34,35,36]. The same question extends naturally beyond the final output to internal activations, where attribution instead identifies which inputs drive a given intermediate computation [37,38].

Attribution methods have arguably received the most scrutiny of the four families, from builders and critics alike. Early sanity checks revealed that several popular methods behaved no differently from edge detectors and were insensitive to model or label randomization [39]. Subsequent work exposed additional shortcomings related to reliability [40], robustness and manipulability [41,42], evaluation methodology [43,44], and the very definition of faithfulness these methods are meant to satisfy [45]. This sustained criticism has progressively narrowed the field toward methods with stronger theoretical grounding, such as those satisfying axiomatic properties [30,46] or based on principled statistical foundations [36,47]. Consequently, recent advances have tended to refine this core set of methods rather than introduce fundamentally new approaches.

Although attribution methods remain imperfect and continue to improve, they have already proven useful for helping humans understand model behavior. In particular, users can leverage attribution maps to reliably predict a model’s behavior on unseen inputs [24], as well as to improve performance in visual search and human-AI teaming tasks [25,48]. Taken together, this evidence suggests that attribution methods are sufficiently mature to serve as reliable tools for comparing models, even if they are not perfect explanations of model reasoning.

Feature visualization methods. While attribution methods are useful, they have an intrinsic limitation: they only explain *where* the model pays attention, not *what* it pays attention to there. Feature visualization methods offer a complementary way to address this gap [49,50,51,52], either by showing natural images that highly activate a target element, or by generating a synthetic image to maximize its activation. The latter approach offers more flexibility, as it goes beyond dataset examples, but this flexibility comes at a cost: without properly constraining the optimization process, it tends to produce either adversarial-looking patterns or stimuli that might be visually appealing but unfaithful to what is actually encoded [50,53,54]. Recent work has largely addressed this weakness by constraining the optimization to a small set of parameters and imposing a human-derived prior directly on the input, allowing the approach to scale to today’s leading vision models [55]. Both variants have since been evaluated for how well they let humans understand what a unit encodes, with encouraging

results [56,57,58,59]. A complementary line forgoes synthesized images and instead describes what a unit encodes directly in natural language, in vision and language models alike [60,61,62].

Concept-based methods. Attribution and feature visualization can explain any target, but historically, that target was individual neurons in a given layer. This proved too narrow, or too entangled, a basis for many units [63,51,64]. TCAV [65] proposed a different basis instead: directions derived from human-defined concepts. One major weakness of this method is that it limits the vocabulary of explanations to whichever concepts are tested, which may not match those that the model relies on. ACE [66] addressed this by automating concept discovery directly from the model’s representations, an approach later extended by ICE [67], CRAFT [68], and sparse autoencoders (SAE) [69]. All of these automated methods can be viewed through the unifying lens of dictionary learning [70], with different instantiations making different assumptions about how a model organizes its concepts [71,72]. Nevertheless, these methods are already useful: they support direct intervention on models [73], clarify which concepts different tasks recruit [74], expose the systematic blind spots of other models [75], and, more broadly, help humans understand models [68].

Circuit-based methods. The newest family moves from *what* a model represents to *how* it computes—reverse-engineering the specific circuits, the connected sets of features and weights, that implement a given behavior [76]. It was pioneered on vision models by the Circuits program, which traced curve detectors and other mechanisms through InceptionV1 [77,78,79,80], and more work has followed since [81,82,83,84,85], though the approach has received comparatively more attention in language models. There, sparse autoencoders decompose representations into interpretable features at scale [69,86,87], and circuit analysis has isolated reusable, human-legible mechanisms: induction heads that copy patterns in context [88], a circuit for indirect-object identification [89], and localized, editable factual associations [90]. The multimodal case sits in between: recent work frames the failures of vision-language models through the binding problem [91] and reads their symbol-like mechanisms directly from activations [92]. In vision proper, the approach is still comparatively niche, but it completes the toolbox: where attribution, feature visualization, and concepts describe a model’s *representations*, circuit-based methods aim to explain its *computation*.

3 Why Now

Imperfect tools, cumulative science. The use of imperfect but informative tools is a recurring practice in science. We take neuroscience as an example. As early as the late 1950s, Lettvin and colleagues read bug- and dimming-detectors directly off the frog’s optic nerve [93]; Hubel and Wiesel mapped orientation-selective receptive fields in cat visual cortex with single electrodes and slits of light cast from a slide projector [94,95], a strikingly simple setup by modern

standards; and, further up the ventral stream, Gross and colleagues found inferotemporal neurons so selective they were discovered almost by accident, including a cell tuned to a hand and, famously, one that fired when the experimenters waved an actual toilet brush in front of the animal [96]. Later work uncovered highly selective neurons in the medial temporal lobe [97,98,99] in addition to distributed representations for faces and objects in ventral temporal cortex [100]. The instruments sharpened in step, from hand-crafted stimuli to reverse correlation for mapping receptive fields [101], and on to causal tools such as microstimulation [102] and lesion-based double dissociations [103] that intervene on a circuit rather than only observing it. Throughout, progress came from using whatever instrument was available and refining models as new findings exposed their limits.

This precedent comes with a nuance. Single-electrode recordings, sampling one neuron at a time, favored a serial feedforward view of vision. Later work showed the cortex to be densely recurrent, with feedback from higher areas shaping even V1 responses [104] and recurrent processing necessary for computations a feedforward model cannot explain [105,106]. This finding did not make the feedforward account wrong, as rapid feedforward processing remains a real and central part of vision, but certainly enriched the knowledge. The lesson for XAI is encouraging: imperfect tools can reveal a durable core of insight that better tools later refine rather than discard, so there is no need to wait for a perfect method before studying models.

Consolidation of models. The space of computer vision models is shrinking in architectural diversity as it grows in raw number. Computer vision has evolved from a fragmented landscape of convolutional architectures [107,33,108,109] to a smaller number of dominant, general-purpose backbones and training paradigms: the Vision Transformer [110], language-supervised contrastive models such as CLIP and its sigmoid-loss successors [111,112], the self-supervised DINO family, now in its third generation [113,114,115], masked image modeling (MAE) and its successors [116], and emerging predictive world-model objectives such as I-JEPA [117]. Although these methods differ in supervision, objectives and downstream capabilities, they increasingly share a common transformer backbone and produce broadly comparable visual representations. This convergence shifts the bottleneck for XAI. Instead of developing new explanation methods for ever-changing architectures, it is now feasible to use existing XAI tools to systematically compare how different training objectives and datasets shape internal representations and computations.

Tools for comparing models. Comparing models is hard: each model’s representation lives in its own idiosyncratic basis. Prior work has worked around this with aggregate comparisons, via the number of unique concept detectors [63,64] or average human interpretability scores across a set of features [58,59], or with soft matching of functionally analogous features [118]. More recently, methods have emerged that support direct model comparison instead of aggregate or functional proxies, such as crosscoders and universal

sparse autoencoders [119,120,121,122], which learn a concept space shared across models and allow comparison at a finer granularity. This converges with a fast-growing literature on measuring representational alignment across systems [123,124,125,126,127,128]. Concept-based interpretability and representational alignment have largely developed as parallel tracks, but recent work shows they can be reconciled [129], making that literature particularly well suited to comparing models directly.

4 Interpretability methods to study models

A small but growing body of work already turns interpretability tools on models themselves. At the scale of one model, a detailed circuit analysis of InceptionV1’s early layers [78,77,130] gave a qualitative account of what a vision network represents, surfacing features like curve detectors; an approach later turned on CLIP to reveal multimodal neurons that fire across both images and text [131]; feature visualization scaled to modern architectures [55] coupled with concept-based methods produced an explorable atlas of the concepts a ResNet50 uses across every class of ImageNet [132]; and a 32,000-concept SAE dictionary for DINOv2 [74] helped characterize the specific concepts recruited by different tasks: “elsewhere” concepts for classification, boundary concepts for segmentation, and three families of monocular depth cues for depth estimation that mirror classical visual neuroscience. The same lens has since been turned on the models now in widest use: CLIP’s image representation has been decomposed head by head and labeled in text [133], the artifact tokens of vision transformers reverse-engineered as internal computation registers [134], and sparse autoencoders used to read interpretable, causally testable features out of both self-supervised backbones [135] and text-to-image diffusion models [136].

Across models, the same tools address comparative questions. Interpretability dissociates from scale [58,137] and, more generally, from capability [63,58,137,118]. It behaves as a stable, measurable property of a model, though how to steer it remains unclear [118]. Direct representational comparison has its own long-standing toolkit (e.g., centered kernel alignment [125] and model stitching [138]) used, for instance, to ask whether vision transformers “see” like convolutional networks [139]; more recent work mines the units common to an entire model zoo [140] and labels neurons automatically to compare concept coverage across networks [61]. Whether models converge on the same concepts gets a mixed answer [64,141,142,143,121], echoing a broader literature on representational convergence [144,145] and human–model alignment [146] that interpretability can both draw on and contribute to.

The human side of the question has its own, equally mature instrument: psychophysics. A growing body of work measures whether explanations improve human understanding of models, including forced-choice tasks that test whether people can identify which images a unit prefers [56,57], studies examining whether interpretability methods help users predict model behavior or improve human–AI team performance [25,24,48,68], and

broader evaluations of how well people grasp machine generated explanations [34,66,147,148,149,150,26,68,151,152]. To date, however, these methods have been used almost entirely to compare XAI methods. Turning on models and asking which model a person can understand best is the natural, and still largely untaken, next step.

Part of the reason why this body of work remains limited may be structural. Machine learning venues tend to reward new methods and benchmark gains rather than careful, systematic evaluations of models. Tellingly, some of the most ambitious efforts to reverse-engineer what a network computes—the Circuits program and the work on monosemanticity—appeared as Distill articles and Transformer Circuits Thread posts rather than in peer-reviewed proceedings [77,80,86,153]. This is beginning to change, but the incentive to publish a systematic interpretability analysis of a model, or a head-to-head comparison across models, remains weak.

5 A Research Agenda for Model-Centric XAI

We close by laying out what taking this shift would require.

First, agree on what understanding means. The field has never converged on a single definition of interpretability: the term is used to refer to different properties [16,17], united primarily by the premise that it is ultimately human-centric [18]. This is not a semantic quibble: without agreement on what it means for a person to understand a model, there is no stable target for the comparisons, metrics, and proxies the rest of this agenda calls for. Establishing what constitutes understanding, and it should be demonstrated, is the prerequisite for everything that follows, and it is why automating its measurement before defining it (as discussed below) risks optimizing the wrong objective rather than a useful shortcut.

Model-centric study is what drives method progress. Studying models directly is not a departure from developing new methods, it is a complementary way to identify what new developments are needed. A prime example of that is the deep dive into the neurons learned by InceptionV1 [77], which surfaced polysemantic units that resisted any single-neuron characterization, motivating the need for concept-based methods [154,68,69]. The same cross-validated study of a model that builds cumulative insight, as argued in Section 3, is also what exposes where our tools fall short. A model-centric agenda should treat this as a mechanism to use deliberately: sustained study of specific models is the most reliable way to discover what the field’s toolbox is still missing.

Where do models actually stand? This is the question interpretability was meant to answer, yet it remains largely unresolved. Despite a mature and increasingly sophisticated toolbox, relatively little work has applied these methods systematically across models (Section 4). As a result, we still do not know how much

progress either approach—interpretable-by-design or post hoc interpretability—has made toward the original goal, nor which aspects of the problem remain unsolved. Answering this question is essential for directing future research. If the bottleneck lies in our interpretability methods, the priority should be developing better tools; if it lies in the models, the focus should shift toward designing and training models that are inherently more understandable.

Design and post hoc explanation remain uncomparing. Interpretability-by-design and post hoc explanation have rarely been compared head to head, particularly with respect to the understanding they produce in human observers. By-design methods, ranging from B-cos networks [155] to label-free concept bottlenecks [156], explicitly impose structure meant to be interpretable, yet it remains largely unknown whether this structure leads to greater human understanding than post hoc explanations of unconstrained “black-box” models. Indeed, recent work suggests that the distinction between the two paradigms may be less fundamental than it appears: both can instantiate the same geometric object, differing only in how the object is identified [157]. Whether either approach ultimately produces models that people understand better therefore remains an open empirical question which belongs squarely within a model-centric agenda.

What predicts interpretability is still unresolved. Interpretability is dissociated from both scale and task performance [58,137,118], which raises an obvious question: what properties predict it? Alignment with human perceptual or semantic organization has emerged as one candidate [158,159,160,118], and robustness as another [161,162,163], but neither currently offers a compelling causal explanation. We are accumulating correlations about which models are more interpretable faster than we are developing principled accounts of why. If alignment does turn out to drive interpretability, it becomes more than a diagnostic: deliberately shaping model representations toward human ones—using perception, psychophysics, and neuroscience as reference points—could be a path to building interpretable models rather than only explaining them, closing the loop between measuring understanding and engineering it

Structural insight is the tractable place to start. Some of this agenda can be pursued without a human in the loop, making structural comparison the natural place to start. It is currently the most tractable direction and the best suited to broad, coordinated progress. Crosscoders and universal sparse autoencoders [119,121] now make it possible to compare representations and features across models directly, rather than relying on aggregate proxies or indirect functional matching [63,64,118]. The same tools also enable systematic comparisons between XAI methods that make different assumptions about how models organize concepts [72], allowing disagreement between methods to become a source of insight rather than noise to be averaged away. More broadly, they provide a direct way to address the convergence question raised in Section 4 directly: whether differently trained models converge on the same concepts is fundamen-

tally a question of structural comparison, and the tools to answer it at scale are now within reach.

Human understanding remains the goal, and the bottleneck. Structural agreement between models is not the same as human understanding, and closing that gap is both the harder challenge and the ultimate objective of this agenda. Large-scale psychophysics has begun to shed light on what makes a representation understandable to human observers [58,137,118], but scaling these evaluations remains prohibitively expensive. This limitation is becoming increasingly acute because tools such as crosscoders and universal sparse autoencoders can generate far more candidate representations than any human study can realistically evaluate. Developing scalable, validated proxies for human understanding, proxies trusted as substitutes for direct human evaluation without replacing it entirely, is a tractable and underexplored direction. It is also what would complete the arc of this paper: transforming the same interpretability toolbox that originally served only to explain models into an instrument for building models that people can understand.

But it is too early to automate the human measure. The bottleneck above invites an obvious shortcut: replace the human with a model. The first serious attempt does exactly this: the Machine Interpretability Score automates the human forced-choice task [56] by scoring how consistently a learned perceptual-similarity model groups a unit’s most activating images, and validates that score against human judgments before scaling it to millions of units [137]. The scale is impressive, but the measure presupposes what it sets out to establish: the automated judge is itself a model of human perception, so interpretability collapses into agreement with a fixed proxy for what people find similar, which is precisely what we do not yet understand. The concern is not that such proxies lack value, but that they risk being adopted before the underlying construct is well defined. The field still has yet to answer more fundamental questions: what psychophysical task best captures understanding? What metric should determine whether a person truly understands a model? Encoding provisional answers into an automated evaluator risks prematurely fixing the definition of interpretability and ranking models by agreement with the proxy rather than by human understanding. The order matters: establish the human measure first, through principled psychophysics, and automate only once we know what we are automating. Even then, any automated proxy should remain anchored to, and regularly recalibrated against, direct human evaluation instead of replacing it.

6 Conclusion

Explainable AI set out to make machine learning models understandable to humans, and for the past fifteen years it has largely focused on developing the tools to do so. We have argued that those tools are finally mature enough to return to the question that motivated them: not whether a given method explains a model,

but whether our models are becoming interpretable and to whom. It is a subtle shift in perspective with far-reaching consequences. It recasts interpretability as a property of models, instead of methods, that can be measured, compared, and steered as models evolve.

The implications extend beyond interpretability research. As AI systems become more capable and broadly used, interpretability is fast becoming the foundation for trust, oversight, and certification. It is also our best opportunity to learn from these systems in return—to discover how they represent vision, language, and other domains where they now rival or surpass human performance. Systems neuroscience built a science around an object it could barely observe. We have the opposite predicament, models are fully accessible, yet they often remain opaque. Bridging the gap between complete access and genuine understanding is, we believe, the defining challenge of the next decade. The instruments are finally ready. What remains is the decision to turn them on the models and on the people they are ultimately meant to help.

References

1. Jiao, L., Hao, J., Li, R., et al.: Foundation Models Meet Medical Image Interpretation. *Research* **9**, 1024 (2026)
2. D’Antonoli, T.A., Bluethgen, C., Cuocolo, R., et al.: Foundation Models for Radiology: Fundamentals, Applications, Opportunities, Challenges, Risks, and Prospects. *Diagnostic and Interventional Radiology* (2025)
3. Baharoon, M., Qureshi, W., Ouyang, J., et al.: Evaluating General Purpose Vision Foundation Models for Medical Image Analysis: An Experimental Study of DINOv2 On Radiology Benchmarks. *arXiv preprint arXiv:2312.02366* (2023)
4. Gao, H., Wang, Z., Li, Y., et al.: A Survey for Foundation Models in Autonomous Driving. In: 2025 6th International Conference on Computer Vision and Data Mining (ICCVDM). pp. 63–71. IEEE (2025)
5. Jiang, S., Huang, Z., Qian, K., et al.: A Survey on Vision-language-action Models for Autonomous Driving. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4524–4536 (2025)
6. Kim, M.J., Pertsch, K., Karamcheti, S., et al.: OpenVLA: An Open-Source Vision-Language-Action Model. In: *Conference on Robot Learning*. pp. 2679–2713. PMLR (2025)
7. Geirhos, R., Jacobsen, J.H., Michaelis, C., et al.: Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
8. D’Amour, A., Heller, K., Moldovan, D., et al.: Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research* **23**(226), 1–61 (2022)
9. Kaminski, M.E.: *The Right to Explanation, Explained*. In: *Research Handbook on Information Law and Governance*. Edward Elgar Publishing (2021)
10. Jumper, J., Evans, R., Pritzel, A., et al.: Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **596**(7873), 583–589 (2021)
11. Fawzi, A., Balog, M., Huang, A., et al.: Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning. *Nature* **610**(7930), 47–53 (2022)
12. Hewitt, J., Geirhos, R., Kim, B.: We can’t understand ai using our existing vocabulary. *arXiv preprint arXiv:2502.07586* (2025)

13. Kazmierczak, R., Berthier, E., Frehse, G., et al.: Explainability and Vision Foundation Models: A Survey. *Information Fusion* **122**, 103184 (2025)
14. European Union Aviation Safety Agency: EASA Artificial Intelligence Concept Paper Issue 2: Guidance for Level 1 & 2 Machine Learning Applications. Tech. rep., European Union Aviation Safety Agency (EASA) (Mar 2024)
15. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union* (2024), adopted 13 June 2024; entered into force 1 August 2024
16. Lipton, Z.C.: The Mythos of Model Interpretability. In: Workshop on Human Interpretability in Machine Learning, Proceedings of the International Conference on Machine Learning (ICML) (2016)
17. Doshi-Velez, F., Kim, B.: Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608 (2017)
18. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* **267**, 1–38 (February 2019)
19. Rudin, C.: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* **1**, 206–215 (2019)
20. Chen, C., Li, O., Tao, C., et al.: This Looks Like that: Deep Learning for Interpretable Image Recognition. In: Advances in Neural Information Processing Systems (NIPS) (2019)
21. Koh, P.W., Nguyen, T., Tang, Y.S., et al.: Concept Bottleneck Models. In: International Conference on Machine Learning. pp. 5338–5348. PMLR (2020)
22. Alvarez-Melis, D., Jaakkola, T.S.: Towards Robust Interpretability with Self-Explaining Neural Networks. In: Advances in Neural Information Processing Systems (NIPS) (2018)
23. Gilpin, L.H., Bau, D., Yuan, B.Z., et al.: Explaining Explanations: An Overview of Interpretability of Machine Learning. In: Proceedings of the IEEE International Conference on data science and advanced analytics (DSAA) (2018)
24. Colin, J., Fel, T., Cadène, R., et al.: What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. *Advances in Neural Information Processing Systems (NeurIPS)* pp. 2832–2845 (2022)
25. Kim, S.S.Y., Meister, N., Ramaswamy, V.V., et al.: HIVE: Evaluating the Human Interpretability of Visual Explanations. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV) (2022)
26. Sixt, L., Schuessler, M., Popescu, O.I., et al.: Do Users Benefit from Interpretable Vision? A User Study, Baseline, and Dataset. *Proceedings of the International Conference on Learning Representations (ICLR)* (2022)
27. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: Workshop, Proceedings of the International Conference on Learning Representations (ICLR) (2013)
28. Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences. In: Proceedings of the International Conference on Machine Learning (ICML) (2017)
29. Smilkov, D., Thorat, N., Kim, B., et al.: SmoothGrad: removing noise by adding noise. In: Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML) (2017)

30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: Proceedings of the International Conference on Machine Learning (ICML) (2017)
31. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
32. Chattopadhyay, A., Sarkar, A., Howlader, P., et al.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2018)
33. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-scale Image Recognition. In: International Conference on Learning Representations (ICLR) (2015), arXiv:1409.1556
34. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier . In: Knowledge Discovery and Data Mining (KDD) (2016)
35. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
36. Fel, T., Cadene, R., Chalvidal, M., et al.: Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
37. Dhamdhere, K., Sundararajan, M., Yan, Q.: How Important Is a Neuron? In: International Conference on Learning Representations (ICLR) (2019), arXiv:1805.12233
38. Leino, K., Sen, S., Datta, A., et al.: Influence-Directed Explanations for Deep Convolutional Networks. In: IEEE International Test Conference (ITC) (2018), arXiv:1802.03788
39. Adebayo, J., Gilmer, J., Muelly, M., et al.: Sanity Checks for Saliency Maps. In: Advances in Neural Information Processing Systems (NIPS) (2018)
40. Kindermans, P.J., Hooker, S., Adebayo, J., et al.: The (UN)Reliability of Saliency Methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science, vol. 11700, pp. 267–280. Springer (2019)
41. Ghorbani, A., Abid, A., Zou, J.: Interpretation of Neural Networks Is Fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2017)
42. Dombrowski, A.K., Alber, M., Anders, C.J., et al.: Explanations Can Be Manipulated and Geometry Is to Blame. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
43. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., et al.: On the (in)Fidelity and Sensitivity for Explanations. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
44. Hooker, S., Erhan, D., Kindermans, P.J., et al.: A Benchmark for Interpretability Methods in Deep Neural Networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
45. Jacovi, A., Goldberg, Y.: Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
46. Lundberg, S., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NIPS) (2017)

47. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *Public Library of Science (PloS One)* (2015)
48. Taesiri, M.R., Nguyen, G., Nguyen, A.: Visual Correspondence-based Explanations Improve AI Robustness and Human-AI Team Accuracy. *Advances in Neural Information Processing Systems (NeurIPS)* (2022)
49. Mahendran, A., Vedaldi, A.: Understanding Deep Image Representations by Inverting Them. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)
50. Nguyen, A., Yosinski, J., Clune, J.: Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned by Each Neuron in Deep Neural Networks. *arXiv preprint arXiv:1602.03616* (2016)
51. Olah, C., Mordvintsev, A., Schubert, L.: Feature Visualization. *Distill* (2017)
52. Nguyen, A., Yosinski, J., Clune, J.: Understanding Neural Networks via Feature Visualization: A survey. *arXiv preprint arXiv:1904.08939* (2019)
53. Nguyen, A., Dosovitskiy, A., Yosinski, J., et al.: Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks. *Advances in neural information processing systems* **29** (2016)
54. Nguyen, A., Clune, J., Bengio, Y., et al.: Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4467–4477 (2017)
55. Fel, T., Boissin, T., Boutin, V., et al.: Unlocking Feature Visualization for Deeper Networks with Magnitude Constrained Optimization. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 36, pp. 37813–37826 (2023)
56. Borowski, J., Zimmermann, R.S., Schepers, J., et al.: Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021)
57. Zimmermann, R.S., Borowski, J., Geirhos, R., et al.: How Well Do Feature Visualizations Support Causal Understanding of CNN Activations? *Advances in Neural Information Processing Systems (NeurIPS)* **34**, 11730–11744 (2021)
58. Zimmermann, R.S., Klein, T., Brendel, W.: Scale Alone Does Not Improve Mechanistic Interpretability in Vision Models. *Advances in Neural Information Processing Systems (NeurIPS)* **36** (2023)
59. Colin, J., Goetschalckx, L., Fel, T., et al.: Choosing the Right Basis for Interpretability: Psychophysical Comparison Between Neuron-based and Dictionary-based Representations. *arXiv preprint arXiv:2411.03993* (2024)
60. Hernandez, E., Schwettmann, S., Bau, D., et al.: Natural Language Descriptions of Deep Visual Features. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2022)
61. Oikarinen, T., Weng, T.W.: CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2023)
62. Bills, S., Cammarata, N., Mossing, D., et al.: Language Models Can Explain Neurons in Language Models. *OpenAI Blog* (2023)
63. Bau, D., Zhou, B., Khosla, A., et al.: Network Dissection: Quantifying Interpretability of Deep Visual Representations. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 6541–6549 (2017)

64. Fong, R., Vedaldi, A.: Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8730–8738 (2018)
65. Kim, B., Wattenberg, M., Gilmer, J., et al.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: International conference on machine learning. Proceedings of the International Conference on Machine Learning (ICML) (2018)
66. Ghorbani, A., Wexler, J., Zou, J.Y., et al.: Towards Automatic Concept-based Explanations. In: Advances in Neural Information Processing Systems. pp. 9273–9282 (2019)
67. Zhang, R., Madumal, P., Miller, T., et al.: Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11682–11690 (2021)
68. Fel, T., Picard, A., Bethune, L., et al.: CRAFT: Concept Recursive Activation FacTorization For Explainability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
69. Gao, L., la Tour, T.D., Tillman, H., et al.: Scaling and Evaluating Sparse Autoencoders. arXiv preprint arXiv:2406.04093 (2024)
70. Mairal, J., Bach, F., Ponce, J., et al.: Sparse Modeling for Image and Vision Processing. Foundations and Trends® in Computer Graphics and Vision **8**(2-3), 85–283 (2014)
71. Fel, T., Boutin, V., Béthune, L., et al.: A holistic approach to unifying automatic concept extraction and concept importance estimation. Advances in Neural Information Processing Systems **36**, 54805–54818 (2023)
72. Hindupur, S.S.R., Lubana, E.S., Fel, T., et al.: Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry. In: Belgrave, D., Zhang, C., Lin, H., et al. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 11649–11699. Curran Associates, Inc. (2025)
73. Dhimoïla, G., Fel, T., Boutin, V., et al.: Cross-Modal Redundancy and the Geometry of Vision–Language Embeddings. In: The Fourteenth International Conference on Learning Representations (ICLR) (2026), arXiv:2602.06218
74. Fel, T., Wang, B., Lepori, M.A., et al.: Into the Rabbit Hull: From Task-Relevant Concepts in DINO To Minkowski Geometry. In: The Fourteenth International Conference on Learning Representations (2026)
75. Bohacek, M., Fel, T., Agrawala, M., et al.: Uncovering Conceptual Blindspots in Generative Image Models Using Sparse Autoencoders. arXiv preprint arXiv:2506.19708 (2025)
76. Elhage, N., Nanda, N., Olsson, C., et al.: A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread (2021)
77. Olah, C., Cammarata, N., Schubert, L., et al.: Zoom in: An Introduction to Circuits. Distill (2020)
78. Olah, C., Cammarata, N., Schubert, L., et al.: An Overview of Early Vision in InceptionV1. Distill (2020)
79. Cammarata, N., Goh, G., Carter, S., et al.: Curve Detectors. Distill (2020)
80. Cammarata, N., Carter, S., Goh, G., et al.: Thread: Circuits. Distill (2020)
81. Hamblin, C., Konkle, T., Alvarez, G.: Pruning for feature-preserving circuits in cnns. arXiv preprint arXiv:2206.01627 (2022)
82. Rajaram, A., Chowdhury, N., Torralba, A., Andreas, J., Schwettmann, S.: Automatic discovery of visual circuits. arXiv preprint arXiv:2404.14349 (2024)

83. Kwon, D., Lee, S., Choi, J.: Granular concept circuits: Toward a fine-grained circuit discovery for concept representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2356–2365 (2025)
84. Żukowska, N., Stammer, W., Schiele, B., Fischer, J.: Seeing through circuits: Faithful mechanistic interpretability for vision transformers. arXiv preprint arXiv:2604.14477 (2026)
85. Thasarathan, H., Kowal, M., Fel, T., Derpanis, K.G.: Cross-model circuit discovery. In: Mechanistic Interpretability Workshop at ICML 2026
86. Bricken, T., Templeton, A., Batson, J., et al.: Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. Transformer Circuits Thread (2023)
87. Cunningham, H., Ewart, A., Riggs, L., et al.: Sparse Autoencoders Find Highly Interpretable Features in Language Models. In: International Conference on Learning Representations (ICLR) (2024), arXiv:2309.08600
88. Olsson, C., Elhage, N., Nanda, N., et al.: In-context Learning and Induction Heads. Transformer Circuits Thread (2022)
89. Wang, K., Variengien, A., Conmy, A., et al.: Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. In: International Conference on Learning Representations (ICLR) (2023), arXiv:2211.00593
90. Meng, K., Bau, D., Andonian, A., et al.: Locating and Editing Factual Associations in GPT. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
91. Campbell, D., Rane, S., Giallanza, T., et al.: Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem. In: Advances in Neural Information Processing Systems (NeurIPS) (2024), arXiv:2411.00238
92. Assouel, R., Campbell, D., Bengio, Y., et al.: Visual Symbolic Mechanisms: Emergent Symbol Processing in Vision Language Models. arXiv preprint arXiv:2506.15871 (2025), arXiv:2506.15871
93. Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., et al.: What the Frog’s Eye Tells the Frog’s Brain. Proceedings of the IRE **47**(11), 1940–1951 (1959)
94. Hubel, D.H., Wiesel, T.N.: Receptive Fields of Single Neurones in the Cat’s Striate Cortex. The Journal of physiology **148**(3), 574 (1959)
95. Hubel, D.H.: Eye, Brain, and Vision. Scientific American Library/Scientific American Books (1995)
96. Gross, C.G., Rocha-Miranda, C.E., Bender, D.B.: Visual Properties of Neurons in Inferotemporal Cortex of the Macaque. Journal of Neurophysiology **35**(1), 96–111 (1972)
97. Quiroga, R.Q., Reddy, L., Kreiman, G., et al.: Invariant Visual Representation by Single Neurons in the Human Brain. Nature **435**(7045), 1102–1107 (2005)
98. Quiroga, R.Q., Kreiman, G., Koch, C., et al.: Sparse but Not ‘Grandmother-cell’ Coding in the Medial Temporal Lobe. Trends in cognitive sciences **12**(3), 87–91 (2008)
99. Quiroga, R.Q., Fried, I., Koch, C.: Brain Cells for Grandmother. Scientific American **308**(2), 30–35 (2013)
100. Haxby, J.V., Gobbini, M.I., Furey, M.L., et al.: Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science **293**(5539), 2425–2430 (2001)
101. Ringach, D., Shapley, R.: Reverse Correlation in Neurophysiology. Cognitive Science **28**(2), 147–166 (2004)

102. Salzman, C.D., Britten, K.H., Newsome, W.T.: Cortical Microstimulation Influences Perceptual Judgements of Motion Direction. *Nature* **346**(6280), 174–177 (1990)
103. Goodale, M.A., Milner, A.D.: Separate Visual Pathways for Perception and Action. *Trends in Neurosciences* **15**(1), 20–25 (1992)
104. Hupé, J.M., James, A.C., Payne, B.R., et al.: Cortical Feedback Improves Discrimination Between Figure and Background by V1, V2 and V3 Neurons. *Nature* **394**(6695), 784–787 (1998)
105. Felleman, D.J., Van Essen, D.C.: Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex* **1**(1), 1–47 (1991)
106. Lamme, V.A.F., Roelfsema, P.R.: The Distinct Modes of Vision Offered by Feed-forward and Recurrent Processing. *Trends in Neurosciences* **23**(11), 571–579 (2000)
107. Krizhevsky, A., Hinton, G., et al.: Learning Multiple Layers of Features from Tiny Images (2009)
108. Szegedy, C., Liu, W., Jia, Y., et al.: Going Deeper with Convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
109. He, K., Zhang, X., Ren, S., et al.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
110. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
111. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 8748–8763. PmLR (2021)
112. Zhai, X., Mustafa, B., Kolesnikov, A., et al.: Sigmoid Loss for Language Image Pre-training. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 11975–11986 (2023)
113. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
114. Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024), arXiv:2304.07193
115. Siméoni, O., Vo, H.V., Seitzer, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
116. He, K., Chen, X., Xie, S., et al.: Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
117. Assran, M., Duval, Q., Misra, I., et al.: Self-supervised Learning from Images with a Joint-embedding Predictive Architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5563–5574 (2023)
118. Colin, J., Goetschalckx, L., Oliver, N., et al.: Capability $\neq$ Interpretability: Human Interpretability of Vision Foundation Models. arXiv preprint arXiv:2605.20337 (2026)
119. Lindsey, J., Templeton, A., Marcus, J., et al.: Sparse Crosscoders for Cross-Layer Features and Model Diffing. *Transformer Circuits Thread* (2024)

120. Gorton, L.: Group crosscoders for mechanistic analysis of symmetry. arXiv preprint arXiv:2410.24184 (2024)
121. Thasarathan, H., Forsyth, J., Fel, T., et al.: Universal Sparse Autoencoders: Interpretable Cross-model Concept Alignment. In: Forty-second International Conference on Machine Learning (2025)
122. Kassem, A., Jiralerspong, T., Rostamzadeh, N., Farnadi, G.: Delta-crosscoder: Robust crosscoder model diffing in narrow fine-tuning regimes. arXiv preprint arXiv:2603.04426 (2026)
123. Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2**, 249 (2008)
124. Morcos, A., Raghu, M., Bengio, S.: Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems* **31** (2018)
125. Kornblith, S., Norouzi, M., Lee, H., et al.: Similarity of Neural Network Representations Revisited. In: International Conference on Machine Learning (ICML) (2019), arXiv:1905.00414
126. Williams, A.H., Kunz, E., Kornblith, S., Linderman, S.: Generalized shape metrics on neural representations. *Advances in neural information processing systems* **34**, 4738–4750 (2021)
127. Sucholutsky, I., Muttenthaler, L., Weller, A., et al.: Getting Aligned on Representational Alignment. arXiv preprint arXiv:2310.13018 (2023)
128. Klabunde, M., Schumacher, T., Strohmaier, M., Lemmerich, F.: Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys* **57**(9), 1–52 (2025)
129. Dhimoila, G., Boutin, V., Picard, A.M., et al.: A Unifying Framework for Concept-Based Representational Similarity. arXiv preprint arXiv:2606.09653 (2026)
130. Carter, S., Armstrong, Z., Schubert, L., Johnson, I., Olah, C.: Activation atlas. *Distill* (2019). <https://doi.org/10.23915/distill.00015>, <https://distill.pub/2019/activation-atlas>
131. Goh, G., Cammarata, N., Voss, C., et al.: Multimodal Neurons in Artificial Neural Networks. *Distill* (2021)
132. Deng, J., Dong, W., Socher, R., et al.: ImageNet: A Large-Scale Hierarchical Image Database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
133. Gandselman, Y., Efros, A.A., Steinhardt, J.: Interpreting CLIP’S Image Representation via Text-Based Decomposition. In: International Conference on Learning Representations (ICLR) (2024), arXiv:2310.05916
134. Darcet, T., Oquab, M., Mairal, J., et al.: Vision Transformers Need Registers. In: International Conference on Learning Representations (ICLR) (2024), arXiv:2309.16588
135. Stevens, S., Chao, W.L., Berger-Wolf, T., et al.: Sparse Autoencoders for Scientifically Rigorous Interpretation of Vision Models. arXiv preprint arXiv:2502.06755 (2025), arXiv:2502.06755
136. Surkov, V., Wendler, C., Mari, A., et al.: Unpacking SDXL Turbo: Interpreting Text-to-Image Models with Sparse Autoencoders. arXiv preprint arXiv:2410.22366 (2024), arXiv:2410.22366
137. Zimmermann, R.S., Klindt, D., Brendel, W.: Measuring Per-Unit Interpretability at Scale without Humans. In: Globerson, A., Mackey, L., Belgrave, D., et al. (eds.) *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 37, pp. 48448–48483. Curran Associates, Inc. (2024)

138. Bansal, Y., Nakkiran, P., Barak, B.: Revisiting Model Stitching to Compare Neural Representations. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021), arXiv:2106.07682
139. Raghu, M., Unterthiner, T., Kornblith, S., et al.: Do Vision Transformers See Like Convolutional Neural Networks? In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021), arXiv:2108.08810
140. Dravid, A., Gandselman, Y., Efros, A.A., et al.: Rosetta Neurons: Mining the Common Units in a Model Zoo. In: *International Conference on Computer Vision (ICCV)* (2023), arXiv:2306.09346
141. Mikriukov, G., Schwalbe, G., Hellert, C., et al.: Revealing Similar Semantics Inside CNNs: An Interpretable Concept-Based Comparison of Feature Spaces. In: *Machine Learning and Knowledge Discovery in Databases (ECML PKDD) Workshops*. pp. 3–20. Springer (2023)
142. Mustapha, A., Khreich, W., Masri, W.: Inter-model Interpretability: Self-supervised Models as a Case Study. *Array* **22**, 100350 (2024)
143. Dorszewski, T., Tětková, L., Jenssen, R., et al.: From Colors to Classes: Emergence of Concepts in Vision Transformers. In: *World Conference on Explainable Artificial Intelligence (xAI)*. pp. 28–47. Springer (2025)
144. Huh, M., Cheung, B., Wang, T., et al.: The Platonic Representation Hypothesis. In: *International Conference on Machine Learning (ICML)* (2024), arXiv:2405.07987
145. Gröger, F., Wen, S., Brbić, M.: Revisiting the Platonic Representation Hypothesis: An Aristotelian View. arXiv preprint arXiv:2602.14486 (2026)
146. Muttenthaler, L., Dippel, J., Linhardt, L., et al.: Human Alignment of Neural Network Representations. In: *International Conference on Learning Representations (ICLR)* (2023), arXiv:2211.01201
147. Narayanan, M., Chen, E., He, J., et al.: How Do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation (2018)
148. Lage, I., Chen, E., He, J., et al.: An Evaluation of the Human-Interpretability of Explanation. In: *Workshop on Correcting and Critiquing Trends in Machine Learning, Advances in Neural Information Processing Systems (NIPS)* (2019)
149. Shen, H., Huang, T.H.: How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. vol. 8, pp. 168–172 (2020)
150. Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems (NeurIPS)* (2021)
151. Casper, S., Bu, T., Li, Y., Li, J., Zhang, K., Hariharan, K., Hadfield-Menell, D.: Red teaming deep neural networks with feature synthesis tools. *Advances in Neural Information Processing Systems* **36**, 80470–80516 (2023)
152. Rong, Y., Leemann, T., Nguyen, T.T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E.: Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence* **46**(4), 2104–2122 (2023)
153. Templeton, A., Conerly, T., Marcus, J., et al.: Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024)
154. Elhage, N., Hume, T., Olsson, C., et al.: Toy Models of Superposition. *Transformer Circuits Thread* (2022)

155. Böhle, M., Fritz, M., Schiele, B.: B-cos Networks: Alignment Is All We Need for Interpretability. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), arXiv:2205.10268
156. Oikarinen, T., Das, S., Nguyen, L.M., et al.: Label-Free Concept Bottleneck Models. In: International Conference on Learning Representations (ICLR) (2023), arXiv:2304.06129
157. Rocchi-Henry, A., Fel, T., Franchi, G.: A Geometric Unification of Concept Learning with Concept Cones. arXiv preprint arXiv:2512.07355 (2025)
158. Fel, T., Rodriguez Rodriguez, I.F., Linsley, D., et al.: Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 9432–9446 (2022)
159. Sundaram, S., Fu, S., Muttenthaler, L., et al.: When does perceptual alignment benefit vision representations? *Advances in Neural Information Processing Systems (NeurIPS)* **37**, 55314–55341 (2024)
160. Muttenthaler, L., Greff, K., Born, F., et al.: Aligning machine and human visual representations across abstraction levels. *Nature* **647**(8089), 349–355 (2025)
161. Ross, A., Doshi-Velez, F.: Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 32 (2018)
162. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
163. Wang, Z., Fredrikson, M., Datta, A.: Robust models are more interpretable because attributions look normal. In: *International Conference on Machine Learning*. pp. 22625–22651. PMLR (2022)