POSITION: INTERPRETABILITY IS A BIDIRECTIONAL COMMUNICATION PROBLEM

Kola Ayonrinde*
UK AI Security Institute

ABSTRACT

Interpretability is the process of explaining neural networks in a human-understandable way. A good explanation has three core components: it is (1) faithful to the explained model, (2) understandable to the interpreter, and (3) effectively communicated. We argue that current mechanistic interpretability methods focus primarily on faithfulness and could improve by additionally considering the human interpreter and communication process. We propose and analyse two approaches to *Concept Enrichment* for the human interpreter – *Pre-Explanation Learning* and *Mechanistic Socratic Explanation* – approaches to using the AI's representations to teach the interpreter novel and useful concepts. We reframe the Interpretability Problem as a Bidirectional Communication Problem between the model and the interpreter, highlighting interpretability's pedagogical aspects. We suggest that Concept Enrichment may be a key way to aid Conceptual Alignment between AIs and humans for improved mutual understanding.

1 THE INTERPRETABILITY PROBLEM

The aim of Interpretability is to generate *explanations* of neural networks that are **human-understandable**. Two key considerations for explanations to be understandable to humans are (1) *which human* our explanation is for and (2) *how long* we might reasonably expect them to spend understanding our explanation. These considerations are concerned with the human interpreter rather than the explanation's content. As Hilton (1990) writes (emphasis original):

> *Causal explanation is first and foremost a form of social interaction. ... The verb to explain is a three-place predicate:* ***Someone*** *explains* ***something*** *to* ***someone****. Causal explanation takes the form of conversation.*

There are always two roles in an explanation: the explainer and the interpreter; a good explanation ought to consider both, as well as their interaction.

1.1 INTERPRETABILITY AND EXPLAINABILITY

Following Chalmers (2025), we make the distinction between Interpretability and Explainability as follows. *Explainability* is an explanation aimed at ordinary humans, for example, the end users of a consumer product. Conversely, *Interpretability* is an explanation aimed at an AI researcher for the purpose of understanding the behaviour and generalisation of an AI model [1] .

Explainability then, seeks explanations understandable to all humans – or at least the majority of humans – regardless of their technical background. Interpretability however, can be aimed at a small group or even a single human, and can assume as much technical background and domain knowledge as necessary. In this way, all else equal, interpretability is a significantly easier problem than explainability.

*Correspondence to: koayon@gmail.com

[1] Note that under this definition, generating explanations intended for experts in some non-ML field, for example, medical doctors, is considered explainability, not interpretability - even though terms of the explanation might look very different to those aimed at the general public.

We define *human-understandable* as "quickly understandable", say in under 2 minutes. [2]

2 BRIDGING THE CONCEPTUAL GAP

The *Mechanistic Turn* in interpretability, led by Chris Olah (Olah et al., 2020), is the Machine Learning analogue to the Cognitive Revolution in Psychology, a movement from a focus on behaviourism (input-output behaviour) to cognitivism (considering the internal representations and processes as important to understanding cognitive systems). Mechanistic Interpretability is characterised by its focus on *Model-level, Ontic and Causal-Mechanistic* explanations of neural networks (Ayonrinde, 2025). Causal-Mechanistic explanations provide a continuous causal chain from cause to effect, without any unexplained gaps. These explanations are given at a level of description that is detailed enough to be "runnable" (Cao & Yamins, 2024).

2.1 THE CONCEPTUAL GAP PROBLEM

A core problem in Mechanistic Interpretability is that the concepts an AI uses to solve a given problem might be quite different from the concepts a human might use to solve the same problem. This is the *Conceptual Gap Problem*. In fact the human may not immediately recognise or understand the AI's concepts. Widdicombe et al. (2018); Hosseini et al. (2018); Goodfellow et al. (2014); Ilyas et al. (2019), inter alia provide evidence that humans and AIs often have different priors and treat different concepts as relevant for decision-making.

One hopeful solution to the Conceptual Gap Problem is *Conceptual Alignment*: if the AI and human used the same set of concepts, then it would be easier to transfer explanations of neural networks, as all the intermediate representations would be naturally human-understandable. The Conceptual Gap between humans and machines prevents them from achieving mutual understanding of representations. Effective explanations require shared concepts.

To solve the Conceptual Gap Problem, following (Schut et al., 2023), we frame the interpretability problem as a cooperative game between two players, the AI M (Machine) and the interpreter H (Human). Each have some set of concepts which are understandable to them, C_M and C_H respectively. The players win the game if H can understand the process that M uses to generate its outputs.

We can achieve Conceptual Alignment here in three ways (see Figure 1):

1. **Coercion.** Coerce M to only use concepts that H understands. An example of Coercion would be creating and using H-interpretable architectures like linear models or Generalised Additive Models (GAMs) (Hastie & Tibshirani, 1985).
2. **Transcreation.** Approximate M's concepts with a close concept that H understands. Supervised probes with human supervision data try to understand the model on the interpreter's (rather than the model's) terms and are an example of Transcreation. Transcreation is analogous to a lossy translation of a text where the translation prioritizes the target language/culture over faithfulness to the source.
3. **Interpreter Concept Enrichment.** Enrich the concept space of H with the concepts that M uses. This can be done by creating tools to empower the interpreter, H, to better understand M, possibly through interactions with M and intuitive visualisations of M's internal workings. Concept Enrichment is analogous to carefully teaching a cross-cultural friend about a foreign word's meaning prioritising faithfulness to the source language.

Mechanistic Interpretability is primarily concerned with the Transcreation and Concept Enrichment approaches. More literal translations of model internals may be important for applications of interpretability techniques to model debugging, preventing catastrophic failures, and aiding with scientific understanding where it is important to avoid the details being "lost in translation".

[2] We note that for the purposes of *interpretability*, the interpreter may have had prior significant familiarity with the model and/or the explanation's domain. We do not count this "pre-work" as part of the time to understand the explanation. This provides another reason why interpretability is significantly more tractable than explainability: in explainability we consider the the average untrained human.

2.2 BIDIRECTIONAL ALIGNMENT

Shen et al. (2024) define Bidirectional Human-AI Alignment as a framework that encompasses two interconnected alignment processes: alignment of AIs to Humans and vice versa. *AI-to-Human Alignment* focuses on integrating human specifications to train and steer AI systems. Conversely, *Human-to-AI Alignment* focuses on human cognitive and behavioral adaptations to AI and supporting human understanding of and collaboration with AIs to preserve human agency.

In the above sense, Concept Enrichment is a form of Human-to-AI Alignment. Approaches to solving the Conceptual Gap Problem needn't focus purely on AI-to-Human Alignment (aligning AIs to humans). Directly empowering humans to better understand AI's concepts can aid Conceptual Alignment.

3 UNDERSTANDING HUMANS FOR INTERPRETABILITY

A basic fact of effective pedagogy is that explanations should be pitched at the appropriate level for the student; that is, good explanations are a function not just of the content being explained but also of the interpreter's existing knowledge, their skills, interests and goals, whether they have any unhelpful beliefs to supplant and so on. Similarly, given our setup of the communicative game in Section 2.1, we note that the best explanation is interpreter-relative. The effectiveness of an explanation depends on the interpreter's goals, background knowledge, and cognitive preferences. This implies that interpretability researchers should arguably study humans as much as we study neural networks. For example, one valuable contribution to interpretability would be understanding why humans describe their observations and reasoning processes using certain concepts and not others. Armed with such an understanding, we could also tailor our interpretability explanations to humans' preferences for explanations.

Recalling Hilton (1990)'s predicate formula of an explanation of "**someone** explaining **something** to **someone**", we focus here on the second someone, the human interpreter. The human (interpreter's) priors are a key part of the theory - as indeed it's humans that we would like to make interpretations for! In this sense, Interpretability is a fundamentally socio-technical problem which may be best addressed by a combination of understanding humans, machines and the (both current and future) interactions between the two. This brings about a key role for Human-Computer Interaction (HCI) and the Social Sciences in the field of Mechanistic Interpretability.

Analysing explanations as being *for* the human interpreters also gives us a concrete way to evaluate interpretability methods. Armed with a good explanation, the interpreter should be able to state which interventions would lead to a different (desired, say) outcome (Lindsay & Bau, 2023; Kirfel et al., 2024). Explanations communicate optimal interventions and the aim is to find the simplest explanations which afford us the most control. An interpretability method succeeds if it can provide useful knowledge to the human, where usefulness can be relative to the task of creating interventions to steer model behaviour or relative to the ability to solve some external scientific problem that the AI contains explanatory knowledge about.

Cao & Yamins (2021) frame understandability as cognitive manipulability. We can evidence this cognitive manipulability by asking the interpreter to try some manipulations: for example, "suppose this part of the explanation changed, would you still get the same answer?" This practical usefulness criteria is an antidote to the problem of Interpretability Illusions (Bolukbasi et al., 2021; Makelov et al., 2024).

4 INTERPRETABILITY AT THE INTERSECTION OF HUMANS AND MACHINES

Interpretability is a human-machine interaction problem. As such, we would like to (1) understand how machines think, (2) understand how humans think, and (3) understand how best to facilitate communication between humans and machines. We now turn our attention to the communication stage, facilitating effective information transfer from AIs to humans. Given that we are aiming at interpretability (i.e. increasing an AI researcher's understanding), we note that there are two core under-explored ways to achieve Concept Enrichment for the Interpreter and hence improve interpretability:

- *Before explanation time*, the interpreter may engage in an educational process. Such learning may be traditional in-domain training (for example, medical training for gaining concept hooks for medical concepts), general interpretability/AI training (for example, analysing the model's internals and behaviour in general domains) or in-domain interpretability training (for example, practising the same interpretability task, predicting the results of interventions and learning from the true results). We refer to this as *Pre-Explanation Learning*.
- *At explanation time*, the interpreter may engage in an interactive dialogue with an interpretability interface for the model. Human-to-human explanations are rarely one-shot; we ask questions, clarify, probe at concepts, identify borderline cases and so on, to understand new concepts. We refer to this dialogic use of explanation time in the case of interpretability as *Mechanistic Socratic Explanation*.

We might draw analogies to to the way that humans better understand each other. People often report ease understanding close friends with few words, because they have shared context and conceptual understanding (Pre-Explanation Learning). Similarly, a natural approach to understanding a technical concept is a discussion with a tutor (Mechanistic Socratic Explanation). We can also draw analogies to the ways that AIs learn - through pre-training compute and in-weights learning, or through the application of test-time compute and in-context learning.

Our approach to Concept Enrichment naturally encourages the design of interactive explanation visualisations and may allow for Contrastive Explanations (Miller, 2019; 2021) to be used for clarifications. For example, when given a causal chain "Input –> p_1 –> p_2 –> p_3 –> Behaviour", an interpreter may ask "why does p_3 follow and not q_3?" to which they may receive a clarifying contrastive explanation to aid understanding.

This approach is not currently common in the field of interpretability, but we consider one proto-example of AI-to-Human Concept Enrichment from Schut et al. (2023). In this work, the authors seek to teach chess grandmasters novel and useful chess concepts from the superhuman chess AI AlphaZero (Silver et al., 2017). Here, the interpreters had done extensive Pre-Explanation Learning (both domain training for chess and, in Schut's case, interpretability training) and were given many opportunities to practise and understand the new techniques they learned from the AI system. Through this approach, Schut et al. (2023) showed that machines were able to teach previously unknown but yet interpretable concepts to humans.

We can contrast Schut et al. (2023) with the human interpretability of Sparse Autoencoder (SAE) features Bricken et al. (2023); Huben et al. (2024). To evaluate the interpretability of SAE features, interpreters generally do not go through any Pre-Explanation Learning and have to understand features in a one-shot setting before they *are* deemed "uninterpretable". We suggest that it's likely that many features are human-understandable but are not typically understood by humans in this non-interactive setting.

5 Towards Interpretability as Bidirectional Communication

To achieve our aim of Interpretability through mutual understanding, we suggest that interactive, multi-turn learning processes like Mechanistic Socratic Explanation and Pre-Explanation Learning may be vital to increasing our ability to understand and learn from AI models. In the near future, AI systems may have superhuman behaviour in a wider range of activities. For humans to acquire the knowledge and concepts that these systems have – whether scientific knowledge which can help us with technical challenges like protein folding, conflict resolution methods, or other behaviours – we would like to expand our concepts and knowledge through learning from AI systems.

Explanations intend to induce learning in the interpreter. In this way, Interpretability is the inverse problem to Machine Learning - in interpretability, it is humans who are doing the learning. We believe that enabling the co-evolution of concepts between humans and machines (Brinkmann et al., 2023) is likely to be a key driver of future progress in both science and in understanding and controlling AI systems.

ACKNOWLEDGMENTS

Thanks to Sean Trott, Louis Jaburi, Michael Pearce, and Mel Andrews for useful conversations. We're grateful to Kwamina Orleans-Pobee for additional support. This project was supported by a Foresight Institute AI Safety Grant.

REFERENCES

Kola Ayonrinde. A mathematical philosophy of explanations in mechanistic interpretability: The strange science part i.i. *Forthcoming*, 2025. Forthcoming.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, 2023.

Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868, 2023.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, pp. 101244, 2024.

David J Chalmers. Propositional interpretability in artificial intelligence. *arXiv preprint arXiv:2501.15740*, 2025.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Trevor Hastie and Robert Tibshirani. Generalized additive models; some applications. In Robert Gilchrist, Brian Francis, and Joe Whittaker (eds.), *Generalized Linear Models*, pp. 66–81, New York, NY, 1985. Springer US. ISBN 978-1-4615-7070-7.

Denis J. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107:65–81, 1990. URL https://api.semanticscholar.org/CorpusID:17460904.

Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. Assessing shape bias property of convolutional neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2004–20048. IEEE, 2018.

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.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Lara Kirfel, Jacqueline Harding, Jeong Shin, Cindy Xin, Thomas Icard, and Tobias Gerstenberg. Do as i explain: Explanations communicate optimal interventions. 05 2024. doi:10.31234/osf.io/4kyfn.

Grace W. Lindsay and David Bau. Testing methods of neural systems understanding. *Cogn. Syst. Res.*, 82:101156, December 2023. URL https://doi.org/10.1016/j.cogsys.2023.101156.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

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. *arXiv preprint arXiv:2310.16410*, 2023.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Amy Widdicombe, Simon Julier, and Been Kim. Saliency maps contain network" fingerprints". In *ICLR 2022 Workshop on PAIR {\textasciicircum} 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2018.

A CONCEPTUAL OVERLAP

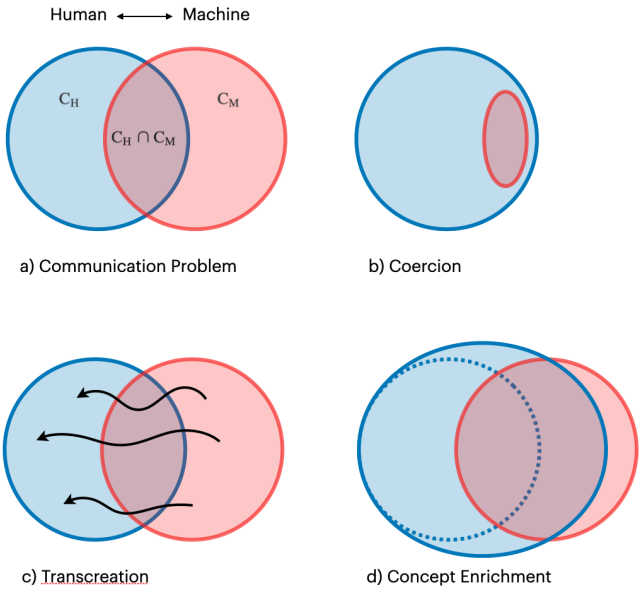

Figure 1: A series of Venn diagrams showing the relationship between the concept spaces of the machine M and human interpreter H. In (a) we see that the machine and human have some shared concepts which they can use to communicate ($C_M \cap C_H$) but there are many concepts that the machine uses which the human does not understand ($C_M \setminus C_H$). AI models that natively use language ostensibly have a larger overlap with human concepts than illiterate models, however there are many concepts that a model may use internally which are not obviously identifiable as human concepts. (b) shows the **Coercion** approach where the machine is forced to use only human-understandable concepts. (c) shows the **Transcreation** approach where the machine's concepts are lossily translated into human-understandable concepts. (d) shows the **Concept Enrichment** approach where we empower the human and increase their concept space such that it overlaps more with the machine's concept space.