UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

How to Leverage Machine Learning Interpretability and Explainability to Generate Hypotheses in Cognitive Psychology

Permalink

https://escholarship.org/uc/item/52s0d3dn

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors

Fedyk, Mark

Ray, Monika

Publication Date

2023

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

How to Leverage Machine Learning Interpretability and Explainability to Generate Hypotheses in Cognitive Psychology

Mark Fedyk (mfedyk@udavis.edu)
Department of Internal Medicine, University of California, Davis
4150 V St, Sacramento, CA 95817

Monika Ray (mray@ucdavis.edu)
Department of Internal Medicine, University of California, Davis
4150 V St, Sacramento, CA 95817

Abstract

This paper describes the principles of a research programme for cognitive science that exploits recent developments in machine learning (ML) to generate novel hypotheses about the structure of human cognition. Current debate over the *interpretability* and *explainability* algorithms usually focuses on the properties of the algorithms themselves in virtue of which they are either interpretable or explainable. However, we argue that there is value in conceptualizing these categories as inherently psychological constructs. Given certain mathematical features of machine learning algorithms – specifically, that many useful ML algorithms are members of *Rashomon sets* – it is possible to exploit their utility to reason using a principle of parsimony about the inferential structure of certain human cognitive tasks. Algorithms that do something the human mind can do, and are both interpretable and explicable could be, we shall argue, *inferential homologues* of certain core cognitive processes. We illustrate this proposal with an example drawn from clustering models used in exploratory data analysis, and then conclude with a discussion of some of the philosophical limitations of our proposal.

Keywords: philosophy of cognitive science, interpretability, explainability, philosophy of machine learning, concept acquisition, ethology, comparative psychology, naturalistic epistemology, cognitive psychology

Scientific Background and Context

The evolutionary pressures of natural selection do not, as a matter of principle, conserve the most optimal, efficient, rational, or balanced solutions to adaptive behavioral problems (West-Eberhard, 2003). There is therefore no general reason, given the evolutionary origins of the human cognitive system, to assume that the way the human mind forms concepts, abstracts categories from perceptual evidence, identifies causes and frames effects, makes inferences about future events, forms beliefs about hidden processes, reasons about the mind of other people is, from either a computational or mathematical or philosophical perspective, anywhere near optimal.

The same reasoning applies in the other direction. There is no reason to assume that the cognitive processes which implement both those and other centrally important cognitive abilities are profoundly sub-optimal. At best, we are licensed to conclude that all these various cognitive capacities are imperfectly useful, and that the utility of cognitive functions will vary across contexts.

But this means it is not possible to make accurate predictions about the specific structure of human cognitive processes on a priori grounds – by defining a particular cognitive task, finding out what algorithms can be used to solve the task, and then identifying the most efficient of these, and concluding that *those* algorithms are (most probably) the ones which are implemented in the human cognitive system (Boyd, 2016; Fedyk, 2015).

An alternative approach finds its home in the study of behavioral processes in comparative ethology (Barnett, 1981; Hoeschele et al., 2022; Tomonaga & Kawakami, 2022) and – slightly more relevantly to our thesis – research in developmental psychology that has examined different kinds of learning by looking for similarities and differences in the cognitive abilities of primates (Csibra & Gergely, 2009; Moll & Tomasello, 2007; Southgate et al., 2007; Tomasello, 2014). In both cases, researchers construct and evaluate hypotheses about cognitive structure based on principles and logic of comparative analysis and parsimony: the same behavioral pattern or cognitive abilities observed in two species with a relatively recent "last common ancestor" probably has the same underlying basis.

Our proposal marries the logic of this basically ethological research strategy to David Marr and Tomaso Poggio's well-known levels of description for analyzing human cognitive processes (Marr & Poggio, 1976; Marr, 2010). Marr and Poggio distinguish between the *computational*, *algorithmic*, and *implementation* levels of description. The computational level of description is a high-level analysis of what the cognitive process is doing: what its function is, and how or why the process achieves (often enough) its function. Because of this, computational descriptions of cognitive processes usually refer to situational or environmental context, agent-based or social goals or ends, and pragmatic or conventional constraints. The algorithmic level of description aims to produce a step-by-step account of how information (or representations of quanta of information) is transformed by calculations to yield new information that can be used to implement the relevant function. Finally, the implementation level of description provides an analysis in

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.* ©2023 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

terms of neurophysiology of how the relevant algorithms transformations can be physically realized.

Below, we argue that advances in machine learning research can be a source of hypotheses about the human cognitive system at that offer testable candidates for experimental investigation at both the *computational level* of analysis and the *algorithmic* level of description. The basis of this proposal rests in the technical property that many useful machine learning models are member of Rashomon sets. But before turning to this idea, we need to introduce some of the basic properties of machine learning, and then also do a bit of philosophical work with the concepts of interpretability and explainability as they apply to machine learning algorithms.

Machine Learning

We prefer a functional definition of machine learning because of its simplicity. A machine learning *model* is a mathematical formula that can be used to make predictions about the content or structure of as yet-unexamined sets of data without being explicitly programmed to make *those* predictions. The formula usually does this by being trained a set of data – the training data or sample data – that is conventionally excluded from membership in the yet-unexamined data.

This definition is useful because it yields taxonomies of ML models determined by the kind of data constitutes the model's training data, the specifics of the formulas which constitute the model and are the basis of its predictions, in terms of constraints that are placed on either the data used by or predicted by the model itself, or the usual logical combinations of the preceding categories.

To illustrate – and to introduce some of the resources that we will use for our ensuring argument – Cynthia Rudin and colleagues (Rudin et al., 2021) provide a table exhibiting taxonomy of interpretable ML models (Table 1).[1] In the left column are customary names of families of ML models and algorithms, and in the right column are the kinds of data that can be modelled by the relevant formulas. Crucially, humans can reason about all the types of data in the table.

Not listed in this table are other well-known classes of ML algorithms that do not easily lend themselves to questions of interpretability or explainability: support vector machines, various configurations of neural networks, large language / transformer models, and so on. Because they are not relevant to our argument, we will (mostly) not discuss these models any further in this paper.

Rashomon Sets

In a widely discussed article, Leo Breiman describes two different fundamental "philosophies" of statistical analysis, the "Data Modelling Culture" and the "Algorithmic Modelling" culture (Breiman, 2001). His aim in introducing this distinction was to develop a critique of the first approach, which predominates in academic (theoretical) statistics, by distinguishing it from the culture of research in machine learning, which often aims to address some real-world problem. Echoing a more general argument advanced by Richard Levins four decades earlier (Levins, 1966; Weisberg, 2006), Breiman contends that the two cultures reflect different ways of resolving an inherent tradeoff in building and using statistical models: predictive accuracy and generalizability usually come at a cost of working with unrealistic or inscrutable models.

But that is not the reason we which to build off Breiman's argument. Instead, in discussing the latter culture, Breiman says:

> What I call the Rashomon Effect is that there is often a multitude of different descriptions [equations $f(\mathbf{x})$] in a class of functions giving about the same minimum error rate. The most easily understood example is subset selection in linear regression. Suppose there are 30 variables and we want to find the best five variable linear regressions. There are about 140,000 five-variable subsets in competition. Usually we pick the one with the lowest residual sum-of-squares (RSS), or, if there is a test set, the lowest test error. But there may be (and generally are) many five-variable equations that have RSS within 1.0% of the lowest RSS. The same is true if test set error is being measured. (Breiman, 2001, p. 206)

Breiman continues, after providing examples of the same phenomenon in neural networks and decision trees, to argue that researchers in the "Algorithm modelling" culture face an inherent trade-off between interpretability and accuracy: if a set S – call this set the *Rashomon set* – of very different models can be used to generate the same "quality" of

Table 1: Varieties of Interpretable Machine Learning Models.

Model	Data
Decision trees / decision lists (rule lists) / decision sets	Tabular data, usually cleaned, with interactions
Scoring systems	Tabular data, usually somewhat cleaned
Generalized additive models	Data represented by variable with at most quadratic interactions
Case-based reasoning	Any
Disentangled neural networks	Raw data, usually representing visual images

[1] We have edited the table for concision. Furthermore, Rudin et al. do not claim that all ML models in their categories are at present interpretable, as their aim is to identify "grand challenges" to making certain sub-classes of the models interpretable.

predictions given the same data, but all the models in S are either inscrutable to humans or mean different things to humans, the models in S cannot therefore be said to be interpretable. Researchers should, according to Breiman, instead explore the members of S, seeking to identify models with the most predictive accuracy.

Interpretability and Explainability

Several researchers have since shown that the Rashomon set for many interest types of data can be sorted according to the relative interpretability of ML models applicable to the data. Actually, that is not quite right: for tabular data, Semenova and colleagues (Semenova et al., 2019) demonstrate that the larger the Rashomon set, the more likely it is that some of its elements are interpretable. Then, Rudin et al. (Rudin et al., 2021) explore the various ways in which the kinds of ML models that do not take in tabular data can also be structured so that they are interpretable without any meaningful loss in predictive accuracy, exploiting the fact that, for many of these models, Rashomon sets exist.

We are relying upon (Rudin et al., 2021) as much as we are because they provide a definition of interpretability that is extremely useful logical contrast case for the argument that we want to develop. Again, for each type of model listed in table 1., Rudin et al. provide a deep analysis of what makes the model interpretable – but in all cases, it is structural features of the model that determine the model's interpretability (e.g., its sparseness). They do not consider the possibility that that the interpretability – and explainability – of some ML model is (at least partially, but still essentially a) psychological question. And the basic, underlying "intuition" of this paper is that cognitive scientists can exploit the fact that Rashomon sets exist for many ML models with real-world uses to frame both computational-level and algorithmic-level hypotheses about cognitive function.

But to cash this intuition out, we need two further resources: explicit conceptualizations of the interpretability and explainability of classes of ML models that are inherently psychological in character.

Interpretability as Semantic Legibility

Interpretability is not simply a matter of the constraints on working memory; we start with this point because a frequent suggestion about what makes a ML model interpretable is that it have sufficiently few parameters to make it possible to hold the model in working memory. But it is also obvious that people without any mathematical knowledge cannot interpret ML models: there is at least a further conceptual dimension to interpretability that must be recognized.

Given this, we propose that the *interpretability* of an ML model be treated as a special case of *semantic legibility*. Semantic legibility is of course a familiar construct in cognitive science. For example, within "East Coast" rationalist school (Gopnik, 2009), it refers to the ability to compose and infer the meaning of a complex semantic expression ("The cat is on the mat") from its atomic elements ("cat", "mat", "on", "the") and grammatical rules (Chomsky, 1965; Glanzberg, 2021; Ludlow, 2014). In this framework interpretability is a computational process: some string of input is interpretable to the extent that the system doing the interpretation can infer the meaning of the whole string from the elements or parts of the string. Length of the input matters – strings can be too short just as they can be too long – but so too does the structure of the string. More importantly, the elements of the string are objects that can be acted on by whatever rules subserve semantic legibility in the human cognitive system.

Please note that, to a good first approximation, this definition of ML interpretability corresponds to the algorithmic level of description in the Marr-Poggio hierarchy.

Explainability as Epistemic Function

"Interpretability" and "explainability" are often conflated with one another. But from the perspective of this paper's argument, they are best kept distinct. Indeed, as Breiman remarks, "Doctors can interpret logistic regression. There is no way they can interpret a black box containing fifty trees hooked together." (Breiman, 2001, p. 209) But they can in fact put both to important scientific and clinical uses: explainability is not interpretability.

To make the distinction explicit, we propose that what makes a ML model explainable is that people to whom the model's function is relevant can understand how the model can generate evidence that is epistemically useful to them. Intuitively, this is the idea that the ML model can be used *as if* it is a mechanism that produces reliable information to any of its downstream epistemic users, and these "patterns of usefulness" can be communicated to other potential users. In more detail: given data with certain properties, the model can make accurate predictions according to 'these' rules or this logic; its ability to do generate accurate predictions breaks down in 'these' ways or in 'these' contexts; and its function is subject to 'these' limitations and caveats and quirks. In possession of this information, a scientist can use a ML model even if the model itself is not semantically legible to the scientist. Here, the model would function as an epistemic object (Chang, 2011) that can be put to different uses in different contexts, and not as a set of free-floating, abstract rules the use of which is determined almost entirely by the semantic properties of the rules.

So, explanation of the model describes its various functions – potentially omitting descriptions of the model's semantic properties. Of course, normally the person generating an explanation of a ML model will be able to (semantically) interpret the model. But if the explanation is successful, the model itself can be grasped and put to various scientific uses by people who nevertheless lack the mathematical concepts and experience necessary to interpret it, and these users will not thereby be prevented by explaining the model to other users despite their inability to interpret the model.

Again, we ask that you note that, to a good first approximation, this definition of ML explainability

corresponds to the computational level of description of the Marr-Poggio hierarchy.

Principles of a Research Programme: The Method of Searching for Inferential Homologues

We can now leverage the conceptual alignments that have emerged to characterize a research programme for cognitive psychology that can pose – and hopefully answer – novel questions about human cognitive processing.

The key methodological idea motivating our proposal is that ML models that are members of Rashomon sets for some set of data and which are *both* interpretable and explainable are interesting starting points for framing hypotheses about cognitive function. Specifically, it may be possible to discover that some of these models are *inferential homologues* of cognitive processes. These inferential homologues are philosophically analogous to the intra-species patterns of behavior which are a traditional focus of ethologist.

In biology, homologues are pairs of traits or structures that share a common structure or function *and* are descended from some common ancestral trait or structure. Cases of homology are distinct from case homoplasy, which are structures or traits that are the same but are not descended from the same common ancestor. The notion of an inferential homologue, then, repurposed both the idea that a pair judgement – one human and the other produced by an ML system – could have the same function (sorting or classifying the same objects) *and* be 'descended' from – that is: derived from, inferred on the basis of, or be the output of – the same background algorithm. Put more intuitively, inferential homologues are pairs of judgments with the same function that are 'descended from a common algorithm'.

With this concept in hand, here is how the research programme can be summarized:

1. The goal of the research programme is to identify inferential processes that are plausibly sources of *inferential homologues*. This can be accomplished by

2. Exploring the Rashomon set for a ML model that can perform work that closely resembles capacities that the human cognitive system can also perform, and more specifically

3. Identifying in that set ML models that are both *interpretable* and *explainable* using the definitions we have suggested. Given this,

4. Experimental tests can be performed that should drive an iterative process of hypothesis formation and subsequent updating that allows researchers to determine if the interpretable expression of the ML model is a sufficiently accurate algorithmic level description of the human cognitive process, *and* if the explanation of the ML model is a sufficiently accurate computational level description of the human cognitive process. Then, if they are,

5. Researchers can conclude (defeasibly) that they have identified an inferential homologue at the intersection of machine learning and human cognitive processing, as this is prima facie a parsimonious explanation of otherwise distinct phenomena (Kitcher, 1981).

Of course, these are very abstract suggestions; to make them more concrete, we will now offer a working example of how our proposed methodology can be implemented.

Working Example

A common task in machine learning is categorizing large data sets by clustering elements or observations in the data together. This of course is comparable to the human ability to categorize – or conceptualize – patterns or sets or sequences of observations. A common family of ML algorithms used to do this are the k-means algorithms (Bock, 2007).[2] These algorithms categorize data to minimize the within-category (within-cluster) variance. Specifically, clusters are formed so that observations are placed in a whichever cluster is defined by a mean value closest to the observation's value. Because of this, the clusters have circular shape: data is distributed around centroid means for each cluster.

Is this how the human minds forms categories? Perhaps. When data is clustered using k-means ML algorithms, it forms Voronoi diagrams (Figure 1). It takes little imagination to come up with various experimental protocols that would investigate whether human participants organize data in such a way that they output of they categorization inferences and judgments would form a Voronoi diagram.

But how would we know that human participants used the same algorithm as a ML model that produces the same (or sufficiently similar) output given the data? It is here that we can exploit our technical definitions of interpretability and explainability. From the principles listed above, we can (defeasibly) infer that if an explanation of the ML model's categorizations is sufficiently like participants' own explanations of their inferences and judgments, then the cognitive processes used by participants and the ML algorithm to have the same epistemic function.

This does not license the conclusion that, at the algorithmic level(s) of cognitive processing, the ML model itself is implemented in a sufficiently similar. But if the ML model is interpretable – though not necessarily, as noted above, to all the people who can understand its epistemic function – then we have a further (defeasible) reason to conclude, on the grounds of parsimony, that the ML model is an algorithm that is likely a pretty good approximation of the algorithm employed by the relevant cognitive process – at least in those study participants who can both interpret the model, who

[2] For concision, we ignore the fact that many of the relevant ML algorithms are NP hard. While we understand that this leaves us open to several difficult philosophical objections to our argument, exploring this issue is both beyond the scope of this paper, and an important outstanding challenge for any research in cognitive psychology that takes seriously the computational theory of mind to overcome.

explain both the model and their own judgments in sufficiently similar terms.

If both these conditions are satisfied, we can claim to have discovered a probable *inferential homologue*.

Of course, the value of this methodology is that not that it will immediately yield evidence of inferential homologues – but rather that it is easy enough to vary the various experimental parameters to efficiently explore the space at the intersection of interpretable and explainable ML models that can be used for tasks that the human cognitive system can also undertake.

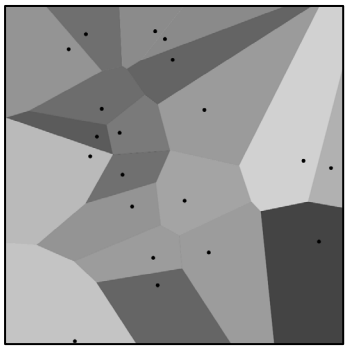

Figure 1: A greyscale Voronoi diagram (Ertl, 2015)

Put another way, when we discover that some ML model can be put to some real-world use, where that application is also something similar enough to something the human mind can do, then in the Rashomon set for the relevant ML algorithm there may be some function that could be analyzed and tested to determine if it yields both a computational-level and algorithmic-level description of some cognitive process.

Other Hypotheses

Despite the popularity of the computational theory of mind among cognitive scientists, and the widespread influence of the Marr-Poggio framework, it is relatively uncommon to find researchers proposing algorithmic accounts of "central" cognitive processing – at least compared to more "perceptual" capacities and processing. We believe that some of the scientific value of our proposed method is that it can either be a source of these accounts, or, if the methodology fails, through its failure to clarify what scientific and philosophical obstacles there are to converting the computational theory of mind from a philosophical proposal about the metaphysics of cognition [cf. (Rescorla, 2015, sec. 6.2; Turing, 1950)] into a larger set of testable hypotheses about central cognitive functions that is the ancestor of a (much) smaller set of reasonably well confirmed hypotheses about the same cognitive functions.

To that end, let us return to the classes of machine learning models listed in Table 1. As more is learned about how to write down specific ML models so that they are interpretable – again, see (Rudin et al., 2021) – it will also be possible to begin to explore whether these models are inferential homologues of human cognitive processing. To illustrate, here are suggestions about what kinds of central cognitive processing abilities might be studied using our suggested method, where the research in question is a spin-off from advances in the family of ML models listed on the left.

- Decisions trees → Voting decisions, bounded rationality/irrationality, intertemporal preference change and stability
- Scoring systems → risk, structure of cognitive boundaries and constraints
- Generalized additive models → all the above, causal inference about causes embedded in systems
- Case based reasoning → structure of concepts, reasoning about borderline / edge cases, abstraction, abductive / inference to the best explanation
- Disentangled neural networks → perception of physical properties, conceptualization of perceptual data.

This list clearly only scratches the surface. But it nevertheless makes the point that many computational and algorithmic hypotheses about central cognitive processes could be productively explored using the method we have proposed.

Conclusion

We conclude by describing three of the largest limitations facing our proposal, and along with this, offering some suggestions about how the first and last of these limitations might be overcome.

The first and largest limitation is sociological in nature. Most researchers working in machine learning and predictive analytics have little exposure to research questions and experimental paradigms in cognitive science. Furthermore, given the applied nature of most research in machine learning, it is unlikely that a large number of researchers in the field would find it interesting to turn some of their attention to basic questions about the fundamental architecture of human cognition. This is a critical obstacle of course because our proposed method of searching for inferential homologues will be fruitful only if a sufficient number of researchers in machine learning develop an interest in researching the class of ML models that perform more or less the same way the human mind performs – an unusual motivation, since the primary thrust of machine learning research is usually to design models that exceed human cognitive capabilities.

A potential solution to this is to note that the Turing test is a special case of a more general problem. That is: Let system A and system B be natural language generating systems, one of them a 25-year-old human whose first language is English, and another the latest version of ChatGPT (OpenAI, 2022) or a similar conversational ML model. A and B are Turing-equivalent if, when a sufficiently large number of randomly chosen interlocutors who themselves are English speakers, can communicate with both A and B, and their guesses about whether A or B is human are accurate only by chance. If A

and B are Turing-equivalent, then B is said to "pass" the Turing test.

But note that this is a round-about way of determining that ChatGPT implements some *computational functions* – in the Marr-Poggio framework – that are also implemented in the human cognitive system, namely those which subserve the specific cognitive abilities recruited to engaged in the specific kinds of semi-formal, semi-conversational written dialogue that are usually exhibited in performed Turing tests.[3]

When Turing tests are described this way, it is easy to see that they focus only on a subset of human cognitive processes – again, those bundles of processes which are invoked in semi-formal, semi-conversational written conversation. This insight allows us to explain why the Turing test is a special case of a more general problem: the fully generally Turing test would have similar structure as the conversational Turing test, but with all tasks that the human mind can perform.

Of course, there is a profound interest in the field of machine learning in overcoming the (special case) Turing test, and so we suggest that the philosophical motivation to spread this enthusiasm to the general form of the Turing test, and we propose that the method of searching for inferential homologues represents one proposal about how these efforts might proceed systematically.

Moving along, the next significant limitation to our approach comes in the form of the objection that it is a mistake to search for homologues between engineered computational systems and the human mind-brain system using a broadly Turing-style computational framework. While there is now ample evidence that various biologically realistic configurations of neural networks are Turing-complete (Chung & Siegelmann, 2021; Date et al., 2022; Siegelmann & Sontag, 1994), it may nevertheless be a basic mistake to assume that the human mind and computing machines compute in the same way. For example, the human mind may be a sui generis analog computation system that has computational abilities which exceed those of Turing machines (Siegelmann, 1998). If so, this undermines the idea that our proposed method is one by which a general theory of human cognition could be developed.

Finally, there is the concern that the Rashomon sets for the most cognitively relevant ML models may, upon searching them, prove to be too sparse to be the source of interesting hypotheses to feed into our proposed methodology – a concern which echoes earlier philosophical worries and warnings about drawing analogies between the human mind and computing machines (Byers, 2022; Dreyfus, 1972, 1992).

These are both important objections. And while there are philosophical replies that can be made to each of them, we believe it is better that the strength of these objections be tested by empirical investigation – perhaps the easiest way to do this would be to see what results emerge from trying our suggested method out.

With all that said, much of the value of the research program we have outlined derives from fact that there are families of algorithms and psychological theories that can be productively explored, and that there are many different methodological principles upon which these explorations can be founded. For example, it may be that parsimony is just the wrong criteria to use when inferring that a pair of judgments are inferential homologues – perhaps other statistical techniques (Bayesian models, maximum likelihood estimates, and so on) or other methodological principles (reduction, coherency, explanatory interest, novelty) would turn out to be more scientifically useful for this task. Then, it could emerge that there are different classes, families, or trees of related algorithms or psychological hypotheses that are more or less susceptible to generating inferential homologues. For example, it could be that some branches in the respective trees contain precisely zero inferential homologues, while others contain many rich examples. These and similar second-order insights could be scientifically interesting, as generating explanations of any such second-order findings could reveal deep and novel lessons about both the human cognitive system and "non-human" machine learning systems alike.

Acknowledgements

Thank you to three anonymous reviewers for their helpful feedback, and to Jim Griesemer, Alok Srivastava, and Patrick Romano for their feedback, criticism, comments, and moral encouragement.

References

Barnett, S. A. (1981). *Modern Ethology: The Science of Animal Behavior* (1st ed.). Oxford University Press.

Bock, H.-H. (2007). Clustering methods: a history of k-means algorithms. *Selected Contributions in Data Analysis and Classification*, 161–172.

Boyd, R. (2016). Boyd How Philosophers 'Learn' from Biology: Reductionist and Anti-reductionist 'Lessons.' In D. Smith (Ed.), *How Biology Shapes Philosophy: New Foundations for Naturalism*.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *16*(3), 199–215.

Byers, P. (2022). There can be no other reason for this behavior: issues in the ascription of knowledge to humans and AI. *Integrative Psychological & Behavioral Science*.

Chang, H. (2011). The Persistence of Epistemic Objects Through Scientific Change. *Erkenntnis. An International Journal of Analytic Philosophy*, *75*(3), 413–429.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. M.I.T. Press.

Chung, S., & Siegelmann, H. (2021). Turing completeness of bounded-precision recurrent neural networks. *Advances in Neural Information Processing Systems*, *34*, 28431–28441.

[3] We stress that the relevant computational functions might be about the pragmatics of conversation – e.g., picking the right kind of speech act to "perform" next – and not semantic generation and processing.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Date, P., Potok, T., Schuman, C., & Kay, B. (2022). Neuromorphic Computing is Turing-Complete. *Proceedings of the International Conference on Neuromorphic Systems 2022*, 1–10.

Dreyfus, H. L. (1972). *What Computers Can?T Do: The Limits of Artificial Intelligence*. Harper & Row.

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.

Ertl, B. (2015). *Euclidean Voronoi diagram*. https://upload.wikimedia.org/wikipedia/commons/thumb/5/54/Euclidean_Voronoi_diagram.svg/800px-Euclidean_Voronoi_diagram.svg.png

Fedyk, M. (2015). How (not) to bring psychology and biology together. *Philosophical Studies*, *172*(4), 949–967.

Glanzberg, M. (2021). Chomsky on semantics. In *A Companion to Chomsky* (pp. 416–432). Wiley. https://doi.org/10.1002/9781119598732.ch26

Gopnik, A. (2009). Rational constructivism: A new way to bridge rationalism and empiricism. *The Behavioral and Brain Sciences*, *32*(2), 208–209.

Hoeschele, M., Wagner, B., & Mann, D. C. (2022). Lessons learned in animal acoustic cognition through comparisons with humans. *Animal Cognition*. https://doi.org/10.1007/s10071-022-01735-0

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science*, *48*(4), 507–531.

Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist*, *54*(4), 421–431.

Ludlow, P. (2014). Recursion, Legibility, Use. In T. Roeper & M. Speas (Eds.), *Recursion: Complexity in Cognition* (pp. 89–112). Springer International Publishing.

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Marr, D., & Poggio, T. (1976). *From Understanding Computation to Understanding Neural Circuitry*.

Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1480), 639–648.

OpenAI. (2022). *ChatGPT: Optimizing Language Models for Dialogue*. https://openai.com/blog/chatgpt/

Rescorla, M. (2015). The Computational Theory of Mind. In *Stanford Encyclopedia of Philosophy*.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2103.11251

Semenova, L., Rudin, C., & Parr, R. (2019). On the Existence of Simpler Machine Learning Models. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1908.01755

Siegelmann, H. T. (1998). *Neural Networks and Analog Computation: Beyond the Turing Limit (Progress in Theoretical Computer Science)* (1999 ed.). Birkhäuser.

Siegelmann, H. T., & Sontag, E. D. (1994). Analog computation via neural networks. *Theoretical Computer Science*, *131*(2), 331–360.

Southgate, V., van Maanen, C., & Csibra, G. (2007). Infant pointing: communication to cooperate or communication to learn? [Review of *Infant pointing: communication to cooperate or communication to learn?*]. *Child Development*, *78*(3), 735–740.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press.

Tomonaga, M., & Kawakami, F. (2022). Do chimpanzees see a face on Mars? A search for face pareidolia in chimpanzees. *Animal Cognition*.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind; a Quarterly Review of Psychology and Philosophy*, *59*(236), 433–460.

Weisberg, M. (2006). Forty Years of 'The Strategy': Levins on Model Building and Idealization. *Biology & Philosophy*, *21*(5), 623–645.

West-Eberhard, M. J. (2003). *Developmental Plasticity and Evolution*. OUP USA.