Should attention be all we need? The epistemic and ethical implications of unification in machine learning

Nic Fishman*
njwfish@gmail.com
University of Oxford
Oxford, UK

Leif Hancox-Li*
leif.hancox-li@capitalone.com
Capital One
McLean, VA, USA

ABSTRACT

"Attention is all you need" has become a fundamental precept in machine learning research. Originally designed for machine translation, transformers and the attention mechanisms that underpin them now find success across many problem domains. With the apparent domain-agnostic success of transformers, many researchers are excited that similar model architectures can be successfully deployed across diverse applications in vision, language and beyond. We consider the benefits and risks of these waves of unification on both epistemic and ethical fronts. On the epistemic side, we argue that many of the arguments in favor of unification in the natural sciences fail to transfer over to the machine learning case, or transfer over only under assumptions that might not hold. Unification also introduces epistemic risks related to portability, path dependency, methodological diversity, and increased black-boxing. On the ethical side, we discuss risks emerging from epistemic concerns, further marginalizing underrepresented perspectives, the centralization of power, and having fewer models across more domains of application.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Philosophical/theoretical foundations of artificial intelligence**.

KEYWORDS

ethics, explanation, philosophy, machine learning, neural networks

ACM Reference Format:
Nic Fishman and Leif Hancox-Li. 2022. Should attention be all we need? The epistemic and ethical implications of unification in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533206

1 INTRODUCTION

In recent years, machine learning (ML) systems based on transformers have achieved new heights across multiple domains. Originally designed for translating languages [112], they have since proved effective in other Natural Language Processing (NLP) tasks and in many other domains, from mainstays of ML research like computer vision (CV) [27] to newer arenas like autonomous driving [92] and protein folding [94]. Transformers do not just work, they dominate the leaderboards that measure progress[1] in ML: they are the top models for most NLP tasks [83, 85, 87], but also in many tasks outside NLP. Transformers are the top 5 models for the largest image classification benchmark (Top 1 ImageNet accuracy) [82], hold the top positions in image segmentation and object detection [84, 86], and top the charts in speech recognition [88]. Those successes cover the most active and well researched problems in ML, at least according to the number of benchmark submissions.

The empirical success of transformers beyond NLP has excited many researchers who have lauded the "unification" of different domains and tasks in ML. Where understanding NLP papers used to require specialized training in computational linguistics, CV researchers now feel that they can easily read papers across the subdisciplinary divide, and vice versa, because of the shared underlying architecture of transformers [10, 21]. Due to the newness of this phenomenon, statements about it tend to be made in less formal settings, industry blog posts/press releases [24, 69, 96] and the popular press [81]. But we feel it is fair to claim there is a sense in the community that 1) transformers offer a unified approach to ML problems and 2) unification is beneficial for ML.

The rise of transformers is related to "foundation models" as described in a Stanford workshop report [11]. That paper defines a foundation model as "any model that is trained on broad data at scale and can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." Most "foundation models" identified in that paper are built using transformer architectures, but the two are conceptually distinct. A variational autoencoder could be a foundation model while not being a transformer; conversely AlphaFold, a transformer that predicts protein structure, is trained for a particular prediction problem rather than a general objective appropriate for fine-tuning for downstream tasks.

These differences aside, Bommasani et al. identify what we call "unification" ("homogenization" in their terms) as a key feature of the history of ML as a field [11]. They pick out two discrete waves of unification. First, the rise of deep learning around 2015, pushing non-neural methods to the periphery of ML research (at least in CV and NLP) and eliminating the need to customize feature pipelines for each problem domain. This was carried out under the auspices of end-to-end learning which could learn good models from "raw"[2] feature inputs. The more recent development and promulgation of

*Both authors contributed equally to this research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533206

[1] It is worth saying this "progress" is often fairly narrow; see [93] for a discussion of benchmarks/leaderboards and their limitations.
[2] Although "raw data" is itself an oxymoron; see [39].

transformers across ML forms a second wave of unification, further accelerating these trends. Just as deep learning marginalized other general purpose ML methods (like boosting or kernel methods), transformers have further narrowed the scope of deployed ML methods to a specific neural architecture. The success of end-to-end learning went some way toward obviating the need for domain expertise in developing ML models. Transformers continue in this direction, unifying architectures across CV and NLP, fields that had formerly been dominated by convolutional neural networks (CNNs) and recurrent neural networks (RNNs) respectively. In both developments we can see two related phenomena: the unification of general purpose ML methods and the convergence of the expertise required to deploy these methods.

In this paper, we take a critical eye to the excitement that ML researchers have expressed about unification being "the" way for ML to progress, highlighting the benefits and risks on both epistemic and ethical fronts. Our remit is less broad than what is covered in [11]. We focus on the unification aspect of this new methodology, leaving aside issues that are specifically due to models being large or trained on large amounts of data.[3]

While transformers are our motivating example for investigating unification, it is possible, or even likely, that transformers will not be the single method that unifies AI research; that may be something new that comes later. Even so, we believe that many of the arguments in this paper will apply to any unified approach that emerges out of a system like the current ML environment.[4]

We start by considering the connection between unification and "artificial general intelligence" (Section 2). These connections underpin some of the motivations underlying researchers' pursuit of unification. Then, we discuss the possible epistemic benefits (Section 3), epistemic risks (Section 4), ethical benefits (Section 5.1), and ethical risks (Section 5.2) of unification.

2 UNIFICATION AND THE PURSUIT OF "ARTIFICIAL GENERAL INTELLIGENCE"

Unification is related to artificial general intelligence (AGI)[5] in two ways. The first is a more general pragmatic connection between unification and AGI, while the second is more ontological and particular to the current neural network paradigm in ML.

It has long been a staple idea of ML research that methods which are promising for AGI should be able to be successfully deployed in a variety of applications. Early work on AI was explicitly predicated on the idea that deploying a "general purpose" learning method on multiple tasks was evidence that the method could be extended to fully realize AGI [65, 89, 99]. Although it is rarely as explicit in modern ML research, this lineage still animates a pragmatist vein of AGI optimism: if a particular general purpose learning framework is working well across all the narrow tasks the field can throw at it, then perhaps that framework is the one that will finally lead to the white whale [72, 93].

A different perspective common in the field today is that, to quote a popular deep learning textbook, "the brain provides a proof by example that intelligent behavior is possible, and a conceptually straightforward path to building intelligence is to reverse engineer the computational principles behind the brain and duplicate its functionality" [40]. As early as 2014 (before many of the most impressive deep learning advances) a survey found that almost half of ML experts believed that the route to AGI would be via cognitive science, and a further 39% believed it would be specifically achieved by artificial neural networks [73]. This view (implicitly) posits an ontological basis for intelligence, supporting the belief that the way to achieve artificial intelligence is to simulate that basis in relevant ways.[6]

This connection between unification and AGI is adjacent to the epistemic/ethical concerns we discuss below, which center around the benefits and risks of unification regardless of whether it leads to AGI. We completely bracket the ethical benefits or risks of AGI, both due to the uncertainty about whether current approaches will bring about AGI [57, 105] and in the interest of deferring to the large literature on the subject [12, 13, 29]. However, AGI is related to our discussion of the epistemic benefits of unification. On certain philosophical views, a shared ontology of learning systems (which underlies the connection between neural networks and AGI) presents real epistemic benefits. In Section 3.1 we discuss whether the ontological motivation for pursuing AGI applies to today's transformer-driven trends of unification.

Unlike the ontological case, the pragmatic case for the connection between unification and AGI is essentially empiricist if we set aside questions about whether it will actually bring about AGI: it simply states that ML practitioners should pursue what works empirically. We address the epistemology of this in Section 4.2 and 4.3. In doing so, we question how "pragmatic" this methodology actually is.

3 EPISTEMIC BENEFITS OF UNIFICATION

Unification is often considered a virtue in science. Unification is a methodological driving force in physics, as demonstrated by the importance of the Standard Model in particle physics, or by the widespread influence of a few "fundamental laws" in diverse areas of physics.[7] Similarly in biology, Darwinism and later the Modern Synthesis were heralded partly because they provided unified accounts of diverse biological phenomena [54, 91, 106]. Philosophers of science have provided several reasons for why unification might indeed be a virtue [20].

In light of this, perhaps it's natural for AI practitioners to seek unified approaches to AI. But when we consider the trends towards unification in ML that were identified in Section 1, it becomes less clear that unification in these senses is an epistemically desirable move. In what follows, we'll look at some general arguments for the desirability of unification in the sciences, and consider whether they transfer well to unification in ML.

[3] While most of the models that fall under our discussion are trained this way, we think that our points will apply even if they are trained on smaller datasets. See [7] and [11] for analyses of the risks of training on big data in particular.

[4] In fact, we believe many of the epistemic concerns we identify apply equally well to unification in any methodological field. We do not discuss these wider issues, but an interested reader can see [18] for an account of the epistemic risks entailed in the social sciences' unification behind the causal inference paradigm.

[5] Also known as strong AI or human-level machine intelligence.

[6] The metaphor of brain as computer may itself have ethical implications as well [6], but we do not examine these here.

[7] We realize that there is disagreement over how unified physics actually is; see [19] for a countervailing perspective. Our aim here is not to resolve this debate but to investigate the extent to which arguments for unification in science can apply to unification in AI.

A note before we proceed: we rely on many accounts of unification in science that are disputed by some philosophers. Resolving these disputes is beyond the scope of this paper. Our intent here is merely to pick out the few most prominent reasons for why we might want unification in science, and see if they apply to the types of unification in AI we consider. Of course, if these accounts of unification are fundamentally flawed, then that could shed doubt on whether they apply to unification in AI. Here we set aside the possibility of these fundamental flaws and merely consider if these accounts of unification can be transferred to the AI case, given the differences between AI and the natural sciences.

3.1 Ontological unification

Unification is often seen as desirable in science because in many cases, we think that the disparate phenomena we want to unify share an ontological basis. Given this basis, it may be reasonable to expect a shared model describing how that basis generates these phenomena. One representative definition of ontological unification is as follows: "Ontological unification is a matter of redescribing apparently independent and diverse phenomena as manifestations (outcomes, phases, forms, aspects) of one and the same small number of entities, powers, and processes" [63]. For example, our knowledge that all matter is composed of the same finite set of elementary particles governed by four fundamental forces leads us to expect fruitful consequences from redescribing various phenomena in terms of these particles and forces.

Could ontological unification be a reason to pursue unification in ML? One reason it could be is if the deep learning architectures we discuss have a shared ontological basis with the human mind. In that case, we could redescribe both human minds and these ML systems as manifestations of the same underlying structures of real or simulated neurons. In other words, unifying ML systems could be a good epistemic move if we expect that the unified architectures represent the structure of the brain and if one of our goals is to create human-like artificial intelligence or AGI. To understand the appeal of ontological unification in ML, we'll next interrogate the extent to which current "state-of-the-art" artificial neural networks (ANNs) resemble the human brain.

The ANNs used in deep learning were initially developed in analogy to biological brains. ML researchers have claimed that they are similar to brains [40, 64]. Many who write about the promise of deep learning do indeed lean on this idea that ever bigger and more sophisticated deep learning systems are the "right" systems to study in pursuit of AGI [73]. Whether they are actually similar to brains and, more importantly, whether improvements in ANNs are driven by pushing ANNs to better approximate the functioning of brains, is a central point in the ontological case for unification.

In 1998 O'Reilly laid out six bare-minimum principles for a network to be considered biologically inspired [80]. These were 1) biological plausibility, 2) distributed representations, 3) inhibitory competition (there should be mechanisms limiting the number of neurons that can fire at once), 4) bidirectional activation propagation, 5) error-driven learning, 6) Hebbian learning (learning is the product of local, not global, information). At the time, ANNs satisfied only two of these (distributed representations and error-driven learning). Today ANNs have not moved much further in the direction of plausibility. They still lack inhibitory competition, they are feed-forward (not bidirectional) for the most part , and they (almost by definition) use global learning algorithms driven by globally defined loss functions. Perhaps more to the point, they are fundamentally not motivated by biological plausibility. Our intention here is to elaborate on this last point by examining some of the history of the microstructure [65] of ANNs to understand at various moments whether development turned on biological or pragmatic concerns. We will argue that the clear orientation toward "what works" and the differences between artificial and biological networks (both those outlined above and additional ones noted below) render the leap from "human brains are AGI" to "ANNs can be AGI" yet unwarranted. To be clear: it is possible that ANNs can be brought closer and closer in line with biology, and even possible that this will bring us closer and closer to AGI. We only intend to point out that such an outcome is not a certainty, nor is it even particularly strongly supported by current evidence.

The ANNs we know today are the descendants of the original perceptron algorithm [99], which was based on a simplified mathematical model of a single neuron [66, 103]. These earliest models used hard connectivity restrictions to maintain biological plausibility. Each neuron could only be connected to so many inputs, mimicking the physical constraints on biological neurons. The limitations of this model were at the heart of Minsky and Papert's original impossibility result, which proved that a perceptron could not compute various quantities without being fully connected to all inputs, that is, without being non-local (alternatively, the perceptron could be multi-layered, but this was dismissed as computationally intractable) [71, 79]. Although the issues with multi-layered networks were eventually resolved, ANNs today are almost always fully connected in precisely the way that Minsky and Papert argued they would need to be, defying biological plausibility.

In the 1980s two solutions emerged for training multi-layered perceptrons: first the Boltzmann machine learning algorithm [2] and second the backpropagation approach [101]. The Boltzmann approach had the benefit of being biologically plausible, the update step depending only on inputs and outputs of single neurons, which is in line with the Hebbian plasticity found in the neural cortex [103]. The backpropagation approach, on the other hand, is inconsistent with biological brains because it is a global update step. But backpropagation allowed for much more efficient training and was widely adopted, despite its deviance from biological plausibility. The "non-locality" of ANNs was mentioned above, but it bears repeating because it is probably the most central deviation from biological plausibility; central in the sense that it is so fundamentally structured into how practitioners work with deep networks. Backpropagation is the defining method that unites the vast array of methods now living under the "neural" umbrella: modern deep learning libraries are effectively auto-differentiation engines built to facilitate backpropagation [1, 15, 90].

Furthermore, while the original perceptron models certainly worked as analogues for biological neurons, they were greatly simplified. These simplifications are exemplified by two features of biological neural networks strikingly absent from their artificial counterparts: spiking behavior and dendritic non-linearities. In biological neural networks neuron activations are event-driven: the neuron is activated by spikes in its inputs, responding only when

inputs change. On a practical level, spiking neurons are much more energy-efficient than contemporary ANNs, but this difference also means that the kind of computation that artificial and biological neural nets perform are fundamentally different [100]. Although spiking behavior has been known about and proposed as an alternative to standard neural nets since the 1990s, it has never really gained significant purchase in the ML community, in large part because it would be difficult to facilitate with standard computational hardware [100].

Artificial and biological neurons also differ in their dendritic non-linearities. In deep neural nets a neuron takes a linear combination of its inputs and then applies the non-linearity, that is $f\left(\sum_i w^{(i)} x_{in}^{(i)}\right)$, but in biological neurons dendrites impose non-linearities on the inputs before summation, that is $f\left(\sum_i w^{(i)} g^{(i)}\left(x_{in}^{(i)}\right)\right)$ [51]. This change would be very straightforward to implement in modern ANNs, in many cases requiring changes to only a single line of code. But it has not seen widespread adoption because biological plausibility is not the main force animating deep learning research.

In contrast to perceptrons, the recent move to transformers had no biological inspiration at all.[8]. Rather, they developed out of a series of insights to improve language translation, the development of the attention mechanism to improve RNNs, and the subsequent realization that "attention is all you need" [112]. Transformers now regularly outperform the biologically-inspired CNNs on precisely the kind of CV tasks that CNN-like cells perform in the visual cortex [22]. This may be an artefact of path dependency in research (see Sec. 4.2), but if that were true it would only further validate our point in this section: ML researchers are not primarily interested in biological plausibility. If they are so not-motivated by biological plausibility that they abandon biologically inspired architectures capable of performing equally well without real cause, that may be symptomatic of problems in the field, but it certainly does not detract from the argument here.

We do not intend this to be an exhaustive outline of differences between artificial and biological neural networks. Rather we have argued two things we think most ANN researchers would agree with. Firstly, we showed how ANN research is driven primarily by methodological pragmatism rather than biological plausibility. Secondly, we outlined how this pragmatism has pushed ANNs used in practice away from biology, rather than towards it, over time. Pragmatic reasons drove fully connected architectures, led to the adoption of backpropagation over Hebbian update rules, underlie the failure of spiking neurons to ever receive significant uptake, and are why transformers swept the research landscape despite a lack of any biological basis. We have not even touched on more fundamental[9] differences between deep learning and cognition, such as the non-causal, non-experimental and non-compositional nature of the most successful ML models [57],[10] the lack of ambiguity in AI [8], or that intelligence is not all in the brain [72].

To be clear, these limitations do not mean ANNs cannot be useful for investigating biological brains. Neuroscience has found some use in ANNs as a model [97, 115]. But using ANNs in that way requires careful construction and comparison to ensure meaningful inferences can be drawn precisely because of these (and other) disanalogies baked into the technology. Failure to account for this can lead to misleading conclusions and faulty science [14]. On the other hand, nothing we have said means that ANNs cannot be brought further in line with biology to fruitful ends. Post-hoc biological analogies can point to promising avenues for taking the pragmatically successful aspects of modern ANNs and "biologizing" them,[11] and if that process (or more generally, the process of bringing ANNs more in line with their biological counterparts) led to more performant models, it could form the foundation of an argument for a shared ontology. But that work has not yet been done, and so cannot (yet) be used to demonstrate a shared ontological basis for biological and artificial neural networks. As it stands, structural differences sufficiently distinguish biological and artificial neural networks to make simplistic analogies ("ANNs are simulations of biological brains") inadequate as grounds for a shared ontological basis.

In short, we do not think the current trajectory of unifying ML through transformers is well-motivated by a methodological principle of ontological unification, because the current top candidates for unification are too different from human brains and show no sign of coming closer to them.

3.2 Explanatory unification

Having addressed the ontological argument for unification, we now turn to more pragmatic reasons for unification.

One reason to pursue unification in science is that unification is itself explanatory, and explanation is one of the goals of science. Friedman uses the example of the kinetic theory of gases as having "reduced a multiplicity of unexplained, independent phenomena to one" [32]. Where before we had separate empirical laws like the Boyle-Charles Law and Graham's Law, we can now use the kinetic theory of gases to understand why those laws and many others hold. This reduces "the total number of independent phenomena that we have to accept as ultimate or given."[12] Similarly, one might hope that instead of accepting as a brute fact that many model architectures work similarly well, having one or two model architectures that perform better than others would improve our understanding of AI.

We think that this reason to pursue unification does not apply to the types of unification we're addressing in this paper (at least right now). In particular, as we outline below, researchers struggle to explain how or why ANNs work in the settings where they are successfully deployed today. Thus explainability is not currently an epistemic benefit to the unification behind transformers (or ANNs before them), because although these models *may* become explainable, that is by no means guaranteed. To make this clear, it

[8] Although CNNs, which previously dominated CV research, did stem from study of complex cells in the visual cortex [33, 97]

[9] In the sense that architectural tweaks alone cannot fix them.

[10] Reinforcement learning (RL) models can bridge this gap to some extent, but the RL models that achieve human-level performance invariably require architectures that are highly biologically implausible. For example, AlphaGo [102] uses an explicit tree search to plan, which is decidedly non-biological.

[11] See [16], which constructs such an analogy for transformers.

[12] Examples of explanatory unification may overlap with ontological unification, because a shared ontology is often thought to be explanatory. We separate ontological and explanatory unification here because we are leaving open the possibility that are types of explanatory unification that aren't identical to ontological unification.

is worth delineating the possible phenomena that this unification might explain.

3.2.1 Possible explananda 1: commonalities between AI systems and human intelligence. One group of candidate explananda is the alleged similarities between AI and human intelligence. The thought is that if the human brain has a similar structure to some ANN that's successful across a broad range of tasks and domains in AI, that could be explained by the hypothesis that the structure of that neural network is the key to AGI.

Whether these are valid explananda for the AI architectures in question depends on the issues we discuss in Section 3.1. As we argue there, we do not think that current neuroscientific evidence supports the hypothesis that the current crop of allegedly general-purpose neural network architectures resemble the brain's architecture.

It's possible that one day we will have a unifying AI architecture that better resembles what we know of the brain's structure. If models with that architecture prove to resemble human intelligence in their performance, then that architecture may explain why humans and those AI systems are similarly intelligent. But as things currently stand, the architectures being pursued in the name of unification will not be able to explain these hypothetical explananda.

3.2.2 Possible explananda 2: inductive biases. The structures of ANNs encode particular inductive biases that allow them to perform well across different tasks. This points to a second possible type of explanatory unification: a unified explanation for these inductive biases. In the ML literature, "inductive biases" refers to "preferences" for the kinds of functions a model represents [41]. For example, ridge regression has an inductive bias for the coefficients to be closer to zero, and the strength of this bias can be tuned by a regularization parameter λ. The No Free Lunch theorem in ML [114] essentially states that all ML algorithms will encode some such set of inductive biases, and that these preferences over the function space are necessary for ML models to generalize to new data. Neural networks clearly encode some kind of inductive biases, and these biases are almost by definition central to their ability to generalize to new data. If the full array of inductive biases of ANNs could be outlined, it would go a long way toward explaining how and why AI models work. Unified models could help provide a common basis for these shared inductive biases. This area has been of central importance to the literature on the theory of deep learning for a number of years [77], but deep learning is a fundamentally empiricist field and the theory tends to lag far behind practice. As it stands today, deep learning theory cannot outline the inductive biases of simple multi-layer perceptron neural networks trained with stochastic gradient descent, much less give an account of the sophisticated architectures achieving state of the art results. Ongoing work on this project is promising as a way forward for the field [41]. If an account of the inductive biases induced by different architectural choices could be constructed, that would go a long way toward arguing for a unified ANN architecture on the basis of explanatory unification.

3.2.3 Possible explananda 3: successful performance by one system across different tasks and modalities. Another possible class of

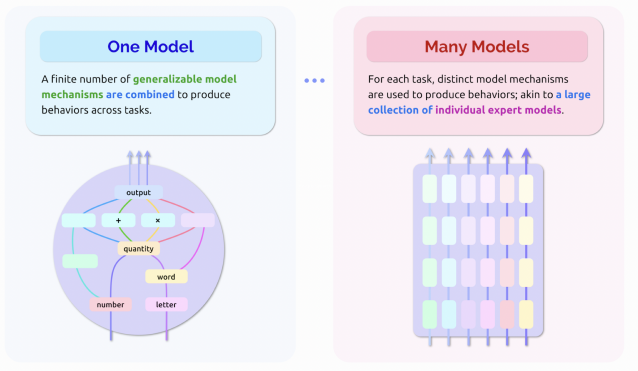

Figure 1: Unified models may fall on a spectrum from having a few mechanisms that generalize to many modalities and tasks (left hand side), or to the other extreme of having separate mechanisms for each task they can perform (right hand side). Figure by Bommasani et al. [11], licensed under CC BY 4.0.

explananda that unified models may explain is the mechanisms underlying successful prediction across different tasks and modalities. We suspect that this is part of why the expanded success of transformers across modalities is so exciting to many researchers: they hope that multi-modal, multi-task models are sharing information across different modalities and tasks. Similarly, there is a hope of shared mechanisms: if what we think transformers are doing with sequential information in language is also what they're doing with image patches engineered to be sequential, then we can understand predictive success in multiple domains using a similar mechanism.

It is as yet unclear if current trends towards unification will bring about this type of unification-through-shared-mechanisms. As Bommasani et al. point out, unified models fall on a spectrum between one model with shared mechanisms for many tasks, or many models that just happen to be "glued" together into one large model (Figure 1):

> As one model, behavior can be made interpretable by identifying and characterising the finite number of generalizable model mechanisms that the model uses to produce behaviors across tasks (e.g., mechanisms that assign meaning to words, compare quantities, and perform arithmetic). As many models, explanations of model behavior in one task are not necessarily informative about behavior in other tasks; thus requiring [sic] to study behavior independently in each task [11].

Given this, if unified models are more like the "one model" end of the spectrum, then they would be better at explaining the explananda described here. If they are more like the "many models" end of the spectrum, then these explananda would not find a unified explanation within the model. Given the newness of these unified models and the difficulties we have in understanding how they make decisions [111], it is unclear at present where on the spectrum they lie.

3.2.4 Review of explanatory unification arguments. In short, the prospects for explanatory unification through the type of architectural unification we discuss in this paper are mixed. For now, it is unclear if unification would be explanatory of intelligence as a phenomenon, of shared inductive biases, or of success across multiple tasks and domains. As the field develops and we get better insight into the inner mechanisms of the unified models, we will hopefully get a clearer picture of how far explanatory unification can be a reason to prefer a unified model landscape.

3.3 Unification as parsimony

In addition to unification being itself explanatory, philosophers of science have also put forward unification as a methodological principle that helps us achieve the epistemic value of simplicity or parsimony. The thought is that "simpler hypotheses, models or theories present a higher likelihood of truth, empirical support and accurate prediction [20]." This thought can be cashed out in different statistical frameworks [75, 107], but we will make some general points that do not assume any particular statistical framework. In this view of unification-as-parsimony, having unified models explaining diverse domains of phenomena is simpler than having many different models for each domain.

How does ML unification fare under this argument? The verdict is mixed here. On the one hand, unification does lead to fewer models covering the same phenomena, and this could be viewed as a simplification. On the other hand, this unification is typically achieved by using more complex models than the counterfactual domain-specific models. The more domains a model has to cover, the more complex it typically is. The complexity consists not just in its internal mechanisms, but in the shape of the predictive function. It's unclear how the unification-as-parsimony arguments would assess this trade-off, as the canonical examples of unification they discuss in science don't have this trade-off. This issue is also related to the one-model-many-models spectrum discussed in Section 3.2.3. To the extent that unified models are near the end of the many-models spectrum, they are likely to be less parsimonious.

A more fundamental problem with the unification-as-parsimony argument is that it's unclear how ML models of the type we're discussing relate to goals like "truth" and "empirical support". These models are selected on the basis of accurate prediction on a held-out test set, so by definition they already satisfy the goal of accurate prediction on past data. The emphasis on their utility as a tool for prediction makes it hard to understand what their relationship to "truth", for example, is. In contrast, models in the natural sciences are often aimed at revealing some true underlying mechanism of nature. It is relatively intelligible to say of Darwinism, for example, that it is true, or that it depicts the truth about some phenomena, but it's unclear what it would mean to say that a transformer-based ML model depicts the truth.[13] Given differences like this between ML and the natural sciences, the unification-as-parsimony argument may need to be modified before it can be intelligibly applied to ML.

[13] We are not endorsing the controversial thesis of scientific realism in this claim–we ourselves take no position on whether scientific theories can be said to be true. We make this point to address audiences who see truth as one of the aims of scientific model-building, since one of the arguments for unification-as-parsimony appeals to the allegedly greater likelihood of simpler models being true.

3.4 Unity of tools

While we have been somewhat skeptical of various arguments for unification presented above, we do not want to deny the real epistemic advantages that unification can offer, which ML researchers have gestured to in their commentary on the phenomenon. Researchers are excited that they can now more easily read ML papers in other domains and transfer techniques useful in one domain to another domain [10, 21, 52]. These are not trivial benefits. When we consider transformers as shared tools that are easily imported between contexts, they can be thought of as "trading zones" or "boundary objects" as defined by historians and sociologists of science [34, 108]. Both material and abstract objects can provide this kind of tool-based unity. In the "trading zones" and "boundary objects" framings of shared tools, a key idea is that shared tools provide a common ground for communication between collaborators, drawing on each collaborator's local understanding without necessitating global agreement. Insofar as we see the facilitation of such collaborations as an epistemic good, we can view transformers as providing epistemic benefits.

4 EPISTEMIC RISKS

Having previously considered the possible benefits of having unified approaches to model-building across multiple tasks and domains, we now consider some possible epistemic risks introduced by greater model unification.

4.1 Turbocharging the tendency to ignore domain experts

One epistemic disadvantage from having fewer, more unified models is the increased risk of falling into what Selbst et al. call the "Portability Trap": "Failure to understand how re-purposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context" [104].

This trap has led to some high-profile ML failures [26, 53, 78]. As Selbst et al. point out, the Portability Trap arises from the modularity taught as good computer programming practice: creating programs in a modular way so that modules can be easily applied to instances outside the originally intended domain.

Another way to characterize the trap is that it's generally a good idea to design a model with the help of domain experts, who can tell you what data to use, how to encode the data as features, how to detect when outputs are not what they should be, and so on. With the advent of unified models, the temptation to ignore domain experts, for cost reasons or otherwise, increases. After all, the model is supposed to work in multiple domains, modalities, and tasks, seemingly obviating the need to customize it to particular contexts.

Ignoring domain experts can also lead to a particular type of ethical risk in which the viewpoints and knowledge marginalized groups are unjustifiably ignored. We discuss this particular ethical risk in Section 5.2.2.

4.2 Path dependency

Enormous amounts of time have been invested into software infrastructure like Tensorflow and Pytorch. Researchers have carefully

refined skills and intuitions about ANNs and how to get them to work well. The volume of published findings about ANNs and the excitement around the area dwarfs that of any other subject in ML. All of this means that investigating other existing paradigms is structurally disincentivized: the lack of infrastructure means developing new models is more costly, the lack of intuition leads to a lack of ideas and frustrations in implementation, and the lack of excitement and new results means investments are unlikely to be rewarded. Put another way, the "switching cost" is extremely high.

Additionally, the particular path ML has gone down is in many senses arbitrary: it is the product of hardware lotteries and the ever increasing availability of compute [49] combined with norms that push for state of the art results on metrics that are riddled with problems [93]. It would be one thing if these concerns were speculative, but there is evidence as well as a sense within the ML community that these are real problems. Within ANN research, there are clear cases of path dependency where technologies that work are abandoned or pursued for arbitrary reasons. One such case discussed by historians is that of over-specialization in chess-playing [28]. Another example is where the combination of incentives to publish novel methods, expanding computational power, and a lack of incentives to fine-tune baselines leads to older, simpler methods being outcompeted by newer methods trained and fine tuned with larger compute budgets. This is clearly what happened in metric learning, where over a decade very little progress was made after accounting for increased compute [74]. In a different but related vein, transformers have found huge success in CV, but some of this is due to engineering innovations in deep architecture design that can be implemented in many kinds of ANNs; importing some of these architectural innovations into convolutional ResNets leads them to be competitive with transformers on large image datasets [61].

These two facts—the high switching cost and the arbitrariness of the current trajectory of research—mean that the ML community, like almost all social communities [60, 62], is extremely path dependent.

In social contexts, path dependency often leads to sub-optimal social outcomes [62], which are of immediate ethical concern. In ML, path dependence is an epistemic concern if the community fails to fully explore alternative ways of developing machine learners. It may be useful to think about path dependence in the context of the explore/exploit trade off [5, 25], which philosophers of science have found useful in thinking through epistemic communities [42, 109]. Path dependence compounds the dilemma posed by the explore/exploit trade off by lowering the cost of exploiting each time the decision is made to exploit. Developing a Bayesian method to compete with ANNs on image classification (for example) is risky in the classic explore/exploit sense in that it might not work, but the dice are loaded: the comparative lack of infrastructure, intuition, and community make it extremely difficult to develop a model of comparable sophistication in a non-deep learning paradigm. In the end, the only way to know whether alternative approaches to ML would be more effective would be to invest similar amounts of time and energy into those alternatives. We take this to be a central epistemic risk of any methodologically homogeneous community.

4.3 Methodological Diversity in the sciences

ML research is at least partly preoccupied with producing methods that can be applied to generate knowledge in other fields, from physics and biology to economics and history. The importance of these "epistemic spillovers" to the epistemic value of ML research creates another key risk in the unification of ML. In scientific enterprises, deriving the same answer from multiple different methodologies is an important tool for confirming hypotheses. This is known as triangulation, and it leads to more robust scientific conclusions [46]. Importantly, this result holds even if one is unsure which methods can be trusted. Such a powerful approach can undoubtedly be very helpful in the difficult non-parametric settings where ML is often required. When ML becomes more unified, that crowds out alternative methods (see Sec. 4.2) which undermines the possibility of triangulation. If there are not enough methods that work sufficiently well to ask the relevant questions, then the answers cannot be triangulated to give the added confidence in the conclusions.

Another argument for methodological plurality comes from research on cognitive diversity [48]. This line of research argues that a group of diverse problem solvers are (under a particular set of assumptions) more effective than a group of individually effective problem solvers. These arguments can be extended to epistemic communities [116], although some caution is required. In particular, Grim et al. note that the results proved in [48] show that under certain circumstances, a group of diverse problem solvers (models) *can* outperform a homogeneous group of individually effective problem solvers [43]. They also demonstrate that true "experts" (as opposed to just more-effective problem solvers) can be substantially better than a diverse group of non-experts. We agree that expert input in the form of domain knowledge is often crucial, especially in science. But that argument cuts against the premise of general purpose ML tools, which tend to push domain experts out of the process (see Sec. 4.1).

The diversity results in [48] apply equally to ML methods: a diverse ensemble of models can more effectively solve problems than an ensemble of individually highly effective models. This fact has also been noted independently within ML where diversity of base classifiers empirically improves boosting [70] and diverse deep ensembles empirically improve performance and uncertainty quantification [59]. The diversity version of the argument for methodological pluralism more directly responds to the pragmatic argument for ML unification ("do what performs well on benchmarks"): first, if we consider ensembles, diverse base models empirically perform well relative to homogenous base models; second, with respect to the field's epistemic spillovers, a diverse array of minimally viable methods is likely epistemically preferable to a single best method.

Combining the benefits of diversity with the confirmatory basis afforded by triangulation, there is a strong epistemic case against unification in ML. Both of these arguments rely on the existence of alternatives, but these alternatives must meet a bare threshold of viability. By viability we mean that alternatives must be flexible enough to accommodate the kinds of research questions practitioners will seek to answer, computationally efficient enough to return answers in a reasonable time frame, and refined enough

for those answers to be meaningful (at minimum providing better-than-random predictions). Considering the problems raised by path dependency in ML, it is clear how the issues discussed in Section 4.2 compound these concerns about depriving non-ML researchers of methodological pluralism.

4.4 Increased black-boxing

The drive to unification will, at least at first, have a tendency to increase the opaque nature of the resulting models. Referring back to the one model-many models concept introduced in Section 3.2.3, unification often occurs by making an existing black-box ANN more complicated. The following moves made in the service of unification perpetuate this black-boxing tendency:

(1) Eliminating feature engineering makes the model less understandable to humans, because feature engineering is often a stage where human-understandable features are created out of data that is less human-understandable.
(2) Many of the more unified multi-modal, multi-domain models come about by combining model architecture ideas inspired by different modalities. How the resulting model generates outputs becomes even more opaque than in the original unimodal models that inspired it, because information from different modalities is combined in ways that we don't fully understand.

Against this, it must be noted that advances in model explainability might ameliorate some of these concerns. But it is by no means guaranteed that such advances will occur or would improve explainability enough to compensate for the model's increased complexity. Explainability for ANNs remains a contested issue. Criticisms of well-known explanation methods abound, and the criteria for what counts as a successful explanation are still being debated [36–38].

Increased black-boxing also has ethical consequences distinct from the potentially higher risk of model failure. In many contexts, model explainability may be considered to be an ethical obligation, particularly when making high-stakes decisions. This is the intuition behind explainability requirements in laws like the GDPR [31]. Increased black-boxing in these contexts, then, would be not just an epistemic risk, but an ethical risk, as finding accurate explanations for ANNs is much more difficult than for, say, generalized additive models.

5 ETHICAL RISKS AND BENEFITS OF UNIFICATION

5.1 Ethical Benefits

While subsequent sections will cover the many ethical risks that further architectural unification in ML presents, it's worth considering some possible ethical benefits that might result, given the appropriate socio-economic conditions.

One benefit that has arguably resulted from the prevalence of widely used open source software (OSS) in all kinds of software applications is that many people gain access to the features of a shared software library. This claim is contestable, since one might argue that society would be better off without any of the software that rely on widely-used open source packages, or that one could have developed similarly good software without those open source packages.

We do want to indicate the possibility, though, that ANNs with unified architectures could be managed as OSS, open to anyone to use and contribute to. Hugging Face is an example of a company that claims to "democratize *good* machine learning" (emphasis theirs) by making ANN models open source [30]. In some conceptions of how society can and should work, this could lead to net social benefits. However, the details around how to manage such unified AI systems to avoid unintended negative consequences are still under-specified. For that reason, we list this as a *possible* benefit from unification, not an actual one.

Similarly, in analogy with OSS, if AI models were open sourced they could be assessed for vulnerabilities or audited for discriminatory biases. This is the "many eyeballs" model of security, and it has certain additional advantages when considering ML systems [4, 35]. The benefit comes with an attendant risk: if a vulnerability does escape the notice of many eyes, then the fact that one centralized package is used by many different people can become a security issue of staggering scale. The log4j vulnerability [68] is a recent example, but within cybersecurity this tradeoff has been debated for decades [47].

In short, we think having fewer, more unified models could have some ethical benefits with the appropriate social and institutional structures, but it's unclear if these benefits would accrue in the actual world.

5.2 Ethical Risks

The risks of large models have recently been discussed extensively [7, 11]. Some of the points we make here will overlap with those discussions, because the model types we're considering overlap with the model types discussed in those papers. However, we focus particularly on the risks emanating from the unification of models.

5.2.1 Ethical Implications of Epistemic Risks. ML systems are deployed in the real world and make decisions that affect the lives of billions of people every day, from whose social media posts get algorithmically boosted to which results appear on the first page of search engine to the automatic annotation of personal photos. This means that many of the epistemic risks we outline in Section 4 have immediate ethical implications. On a basic level, the tendency to ignore domain experts (Section 4.1) and the issues around path dependency (Section 4.2) may make ML systems less effective tools, which for systems that are so widely used has immediate social welfare implications.[14] And as we discussed in Section 4.4, the trend of increased black-boxing itself has distinct ethical implications related to subjects' rights to explanations.

5.2.2 Making it easier to ignore marginalized perspectives. As we argued previously in Section 4.1, model unification can amplify the existing tendency to ignore domain experts. This in turn adds to an ongoing pattern of building models without getting input from marginalized perspectives that could have informed key decisions about the model. We think this is not just an epistemic risk, but also an ethical risk. ML has a track record of disproportionately harming

[14] Although if those systems are designed to operate against the best interests of their decision subjects, their suboptimality may be an ethical silver lining.

marginalized groups, and many critics have pointed out that the lack of diversity among the builders of ML probably doesn't help [17, 58, 67].

When a system could harm marginalized groups, it is reasonable to think of members of these groups as domain experts on the decisions made by the system. When domain experts are left out of processes like determining what features are valid ones for predictions, these marginalized perspectives will also be left out. End-to-end learning is designed to cut humans out of the process of determining what features are good ones for the model to rely on. Having a plug-and-play model that can be easily used in different domains will put more power in the hands of ML engineers and adjacent roles, while taking power away from domain experts.

This will also abstract the systems further from their social context. If any engineer can take a CV model and plug-and-play it into an NLP setting, there is less friction in this process to encourage the engineer to consult a language expert. One ethical repercussion is that perspectives from sociolinguistics about language-related harms will be more likely to be left out.

This is related to the feminist epistemological values of considering context and multiple points of view [44, 45]. If we think of each model as giving us a certain perspective into its domains, feminist epistemologists favor having multiple such perspectives, since each perspective is influenced by its own particular background. There is no one "best" perspective. Models, like perspectives, embed particular social values, and the fewer models we have, the more a few privileged perspectives will get to determine the shape of the world. Having one or even just a few general purpose model architectures that supposedly can "understand" diverse domains is eerily close to the notion of having a "view from nowhere" [76], which feminist epistemologists have criticized as privileging the white masculine gaze over more marginalized perspectives.

5.2.3 Further centralization of power. The increasing centralization of power in AI has been discussed by various activists and researchers, especially as it pertains to large ML models [113]. There is an additional aspect of unification that lends itself to centralized power beyond the fact that these models are large. Unification expands the potential applications of a single model over many more areas of life. The vision that some researchers have is to be able to train a model on, say, a language corpus and then have it perform tasks in CV, knowledge-based reasoning, and so on. This ambition, to be general-purpose across different modalities, tasks, domains, and features, means that whoever controls a pre-trained model of this general purpose nature would wield broader power over more areas of human life.

5.2.4 The "algorithmic leviathan" argument against algorithmic monocultures. Another ethical risk of having a more unified model landscape is that *systematic* arbitrariness in decision making can be morally wrong in a way distinct from any wrongs that might be attributed to arbitrariness in individual decisions. Creel and Hellman argue that arbitrariness in individual decisions may be morally problematic sometimes, but in situations where the arbitrariness does not violate fundamental rights, there could be an instrumental justification for arbitrariness [23]. In contrast, arbitrariness at scale, where a single model is able to make arbitrary decisions that affect many people in many domains, introduces an element of wrongness over and above any wrongs wrought by individual arbitrary decisions. With arbitrariness at scale, an individual may face an arbitrary decision from an algorithm not just in one instance, but in many instances across many opportunities. This systematic exclusion is morally problematic under many moral theories: it could translate to a lack of capacities, lack of genuine freedom, or being at the bottom of a social hierarchy of esteem or domination [3, 95, 98]. Model unification increases the systematicity of arbitrariness, because it enables fewer models to make decisions over more domains.

5.2.5 Epistemic homogeneity's impact on social welfare. In Section 4.2, we put forward some epistemic considerations against narrowing the path of ML research to be more focused on unified models. In addition to these potential epistemic drawbacks, emerging research suggests that there could also be negative impacts on aggregate social welfare if the models used by organizations to make decisions are more accurate but more homogeneous. Kleinberg and Raghavan argue that introducing a more accurate algorithm that's used by many organizations can drive society into an equilibrium that's worse than having many less accurate algorithms [55]. This leads to the average quality of decisions across society being lower, even if it makes sense for individual organizations to each opt for the most accurate algorithm. Kleinberg and Raghavan make their point using the case of ranking algorithms, so the implications for other types of algorithms are unclear, but their work shows that we should be wary of assuming that having everyone use the same "state of the art" algorithm is necessarily better for society.

6 CONCLUSION

In this paper, we characterized key trends in recent AI research that unify model architectures across different modalities and tasks. We considered the potential benefits and risks of these trends on both epistemic and ethical fronts. We've focused specifically on the features of these models that are directly related to unification—that is, directly related to the aspiration to provide model architectures that can be applied across multiple domains, tasks, and modalities with little to no adjustment. This provides a complementary perspective to previous work that has discussed the implications of other properties of these models, such as their large size and their reliance on large corpuses of training data [7, 11].

We argue that on current evidence, the epistemic benefits are not as strong as analogous benefits from unification in the natural sciences. We discuss multiple epistemic risks arising from having too homogeneous a methodological community. Finally, we discuss possible ethical benefits from open-source unified models, and ethical risks like further marginalizing underrepresented perspectives and facilitating centralized, homogeneous decisions.

Moving forward, we think there are further possible consequences of unification that are worth exploring. For example, the performativity of ML models has been noted by other commentators—models have the potential to change the behavior of decision subjects, with this changed behavior in turn potentially reinforcing the models' "correctness", and so on in a positive feedback loop [9, 50, 56]. Do more unified models strengthen this dynamic where it exists, or make such feedback loops more probable? What about

the phenomenon of "Holy Grail performativity"—where the presence of an overarching goal of doing well across multiple predefined tasks, modalities and domains changes researcher behavior to be oriented around this quest, in a way that potentially forecloses other possibilities [110]? We hope that these and other possible consequences of a more unified model landscape will be explored in future work.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for Boltzmann machines. *Cognitive science* 9, 1 (1985), 147–169.

[3] Elizabeth S Anderson. 1999. What is the Point of Equality? *Ethics* 109, 2 (1999), 287–337.

[4] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[5] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2 (2002), 235–256.

[6] Alexis T Baria and Keith Cross. 2021. The brain is a computer is a brain: neuroscience's internal debate and the social significance of the Computational Metaphor. *arXiv preprint arXiv:2107.14042* (2021).

[7] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[8] Abeba Birhane. 2021. The Impossibility of Automating Ambiguity. *Artificial Life* 27, 1 (2021), 44–61.

[9] David Bloor. 1996. Idealism and the Sociology of Knowledge. *Social studies of science* 26, 4 (1996), 839–856.

[10] John Bohannon and Sam Charrington. 2022. Trends in NLP with John Bohannon. https://twimlai.com/trends-in-nlp-with-john-bohannon/

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[12] Nick Bostrom. 2017. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[13] Nick Bostrom and Eliezer Yudkowsky. 2018. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, Roman V. Yampolskiy (Ed.). Chapman and Hall/CRC, Boca Raton, Florida, 57–69.

[14] Jeffrey Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John Hummel, Rachel Flood Heaton, Benjamin Evans, Jeff Mitchell, and Ryan Blything. 2022. Deep Problems with Neural Network Models of Human Vision. https://doi.org/10.31234/osf.io/5zf4s

[15] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. *JAX: composable transformations of Python+NumPy programs*. http://github.com/google/jax Accessed May 2, 2022.

[16] Trenton Bricken and Cengiz Pehlevan. 2021. Attention Approximates Sparse Distributed Memory. *Advances in Neural Information Processing Systems* 34 (2021).

[17] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI Now 2017 Report. https://ainowinstitute.org/AI_Now_2017_Report.pdf. Accessed: 2022-01-21.

[18] Nancy Cartwright. 2021. Rigour versus the need for evidential diversity. *Synthese* 199, 5 (2021), 13095–13119.

[19] Nancy Cartwright et al. 1999. *The dappled world: A study of the boundaries of science*. Cambridge University Press.

[20] Jordi Cat. 2022. The Unity of Science. In *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[21] Sam Charrington and Georgia Gkioxari. 2022. Trends in Computer Vision with Georgia Gkioxari. https://twimlai.com/trends-in-computer-vision-with-georgia-gkioxari/

[22] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 5 (2021), 1–32.

[23] Kathleen Creel and Deborah Hellman. forthcoming. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems. *Canadian Journal of Philosophy* (forthcoming).

[24] DeepLearning.AI. 2021. The Batch. https://read.deeplearning.ai/the-batch/issue-123/ "Originally developed for natural language processing, transformers are becoming the Swiss Army Knife of deep learning.".

[25] Berna Devezer, Luis G. Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. 2019. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE* 14, 5 (05 2019), 1–23. https://doi.org/10.1371/journal.pone.0216125

[26] Ben Dickson. 2019. Human Help Wanted: Why AI Is Terrible at Content Moderation. https://www.pcmag.com/opinions/human-help-wanted-why-ai-is-terrible-at-content-moderation Accessed Jan 18, 2022.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[28] Nathan Ensmenger. 2012. Is chess the drosophila of artificial intelligence? A social history of an algorithm. *Social studies of science* 42, 1 (2012), 5–30.

[29] Tom Everitt, Gary Lea, and Marcus Hutter. 2018. AGI safety literature review. *arXiv preprint arXiv:1805.01109* (2018).

[30] Hugging Face. n.d.. huggingface (Hugging Face). https://huggingface.co/huggingface Accessed Jan 17, 2022.

[31] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2019. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society* 6, 1 (2019), 2053951719860542.

[32] Michael Friedman. 1974. Explanation and scientific understanding. *The Journal of Philosophy* 71, 1 (1974), 5–19.

[33] Kunihiko Fukushima. 1980. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36 (1980), 193–202.

[34] Peter Galison et al. 1997. *Image and logic: A material culture of microphysics*. University of Chicago Press.

[35] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[36] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[37] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.

[38] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[39] Lisa Gitelman (Ed.). 2013. *Raw data is an oxymoron*. MIT Press, Cambridge, Massachusetts.

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, Massachusetts. http://www.deeplearningbook.org.

[41] Anirudh Goyal and Yoshua Bengio. 2020. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091* (2020).

[42] Patrick Grim. 2019. Modeling epistemology: examples and analysis in computational philosophy of science. In *2019 Spring Simulation Conference (SpringSim)*. IEEE, 1–12.

[43] Patrick Grim, Daniel J Singer, Aaron Bramson, Bennett Holman, Sean McGeehan, and William J Berger. 2019. Diversity, ability, and expertise in epistemic communities. *Philosophy of Science* 86, 1 (2019), 98–123.

[44] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[45] Sandra Harding. 1995. "Strong objectivity": A response to the new objectivity question. *Synthese* 104, 3 (1995), 331–349.

[46] Remco Heesen, Liam Kofi Bright, and Andrew Zucker. 2019. Vindicating methodological triangulation. *Synthese* 196, 8 (2019), 3067–3081.

[47] Scott A. Hissam, Daniel Plakosh, and C Weinstock. 2002. Trust and vulnerability in open source software. *IEE Proceedings-Software* 149, 1 (2002), 47–51.

[48] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[49] Sara Hooker. 2021. The hardware lottery. *Commun. ACM* 64, 12 (2021), 58–65.

[50] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*. 1389–1398.

[51] Ilenna Simone Jones and Konrad Paul Kording. 2020. Can single neurons solve MNIST? The computational power of biological dendritic trees. *arXiv preprint arXiv:2009.01269* (2020).

[52] @karpathy (Andrej Karpathy). 2021. The ongoing consolidation in AI is incredible.... https://twitter.com/karpathy/status/1468370605229547522?s=20&t=go3X-8IlBf_X-ekQ07oh7g Accessed: 2022-05-01.

[53] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* 17, 1 (2019), 1–9.

[54] Philip Kitcher. 1981. Explanatory unification. *Philosophy of Science* 48, 4 (1981), 507–531.

[55] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021). https://doi.org/10.1073/pnas.2018340118

[56] Paul Kockelman. 2020. The epistemic and performative dynamics of machine learning praxis. *Signs and Society* 8, 2 (2020), 319–355.

[57] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences* 40 (2017).

[58] Sage Lazzaro. 2021. Are AI ethics teams doomed to be a facade? Women who pioneered them weigh in. https://venturebeat.com/2021/09/30/are-ai-ethics-teams-doomed-to-be-a-facade-the-women-who-pioneered-them-weigh-in/ Accessed Jan 18, 2022.

[59] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. 2015. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314* (2015).

[60] Stan J Liebowitz and Stephen E Margolis. 1995. Path dependence, lock-in, and history. *Journal of Law, Economics, & Organization* 11 (1995), 205–226. Issue 1.

[61] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. *arXiv preprint arXiv:2201.03545* (2022).

[62] James Mahoney. 2000. Path dependence in historical sociology. *Theory and Society* 29, 4 (2000), 507–548.

[63] Uskali Mäki. 2001. Explanatory unification: Double and doubtful. *Philosophy of the Social Sciences* 31, 4 (2001), 488–506.

[64] James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences* 117, 42 (2020), 25966–25974.

[65] James L McClelland, David E Rumelhart, PDP Research Group, et al. 1986. *Parallel distributed processing*. Vol. 2. MIT Press, Cambridge, Massachusetts.

[66] Warren S McCulloch and Walter Pitts. 1990. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology* 52, 1 (1990), 99–115.

[67] Stuart McLennan, Meredith M Lee, Amelia Fiske, and Leo Anthony Celi. 2020. AI ethics is not a panacea. *The American Journal of Bioethics* 20, 11 (2020), 20–22.

[68] Mike Melanson. 2021. Log4j Is One Big "I Told You So" for Open Source Communities. https://thenewstack.io/log4j-is-one-big-i-told-you-so-for-open-source-communities/ Accessed Jan 17, 2022.

[69] John. P. Mellow Jr. 2022. State of AI Report: Transformers Are Taking the AI World by Storm. https://exchange.scale.com/public/blogs/state-of-ai-report-2021-transformers-taking-ai-world-by-storm-nathan-benaich "Transformers [...] took the machine learning world by storm in 2021. Originally designed to work with natural language processing models, the technology has blasted out of NLP over the last 12 months to emerge as a general-purpose architecture for ML." Accessed May 9, 2022.

[70] Prem Melville and Raymond J Mooney. 2004. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*. 74.

[71] Marvin Minsky and Seymour A Papert. 2017. *Perceptrons: An introduction to computational geometry*. MIT Press, Cambridge, Massachusetts.

[72] Melanie Mitchell. 2021. Why AI is harder than we think. *arXiv preprint arXiv:2104.12871* (2021).

[73] Vincent C Müller and Nick Bostrom. 2016. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence*. Springer, 555–572.

[74] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *European Conference on Computer Vision*. Springer, 681–699.

[75] Wayne C Myrvold. 2003. A Bayesian account of the virtue of unification. *Philosophy of Science* 70, 2 (2003), 399–423.

[76] Thomas Nagel. 1989. *The view from nowhere*. Oxford University Press.

[77] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2014. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614* (2014).

[78] Safiya Umoja Noble. 2018. *Algorithms of oppression*. New York University Press.

[79] Mikel Olazaran. 1996. A sociological study of the official history of the perceptrons controversy. *Social Studies of Science* 26, 3 (1996), 611–659.

[80] Randall C O'Reilly. 1998. Six principles for biologically based computational models of cortical cognition. *Trends in cognitive sciences* 2, 11 (1998), 455–462.

[81] Stephen Ornes. 2022. Will Transformers Take Over Artificial Intelligence? https://www.quantamagazine.org/will-transformers-take-over-artificial-intelligence-20220310/ "Wang, for example, thinks the transformer may be a big step toward achieving a kind of convergence of neural net architectures, resulting in a universal approach to computer vision — and perhaps to other AI tasks as well." Accessed May 9, 2022.

[82] Papers With Code. 2022. Image Classification on ImageNet. https://paperswithcode.com/sota/image-classification-on-imagenet Accessed Apr 22, 2022.

[83] Papers With Code. 2022. Machine Translation on WMT2014 English-German. https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german Accessed Apr 22, 2022.

[84] Papers With Code. 2022. Object Detection on COCO test-dev. https://paperswithcode.com/sota/object-detection-on-coco Accessed Apr 22, 2022.

[85] Papers With Code. 2022. Question Answering on SQuAD1.1. https://paperswithcode.com/sota/question-answering-on-squad11 Accessed Apr 22, 2022.

[86] Papers With Code. 2022. Semantic Segmentation on ADE20K. https://paperswithcode.com/sota/semantic-segmentation-on-ade20k Accessed Apr 22, 2022.

[87] Papers With Code. 2022. Sentiment Analysis on SST-2 Binary classification. https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary Accessed Apr 22, 2022.

[88] Papers With Code. 2022. Speech Recognition on LibriSpeech test-clean. https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean Accessed Apr 22, 2022.

[89] Seymour Papert. 1988. One AI or many? *Daedalus* 117 (1988), 1–14. Issue 1.

[90] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[91] Anya Plutynski. 2005. Explanatory unification and the early synthesis. *The British journal for the philosophy of science* 56, 3 (2005), 595–609.

[92] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7077–7087.

[93] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks* (2021).

[94] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 2020. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*.

[95] Joseph Raz. 1986. *The morality of freedom*. Clarendon Press, Oxford.

[96] Microsoft Research. 2021. Five reasons to embrace Transformer in computer vision. https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/articles/five-reasons-to-embrace-transformer-in-computer-vision/ "Unity is a common goal in many disciplines. [...] In natural language processing (NLP), the current dominant modeling network is Transformer; in computer vision (CV), for a long time, the dominant architecture was convolutional neural networks (CNN); in social computing field, the dominant architecture is graph networks; and so on. However, the situation has changed since the end of last year, when Transformers began to demonstrate revolutionary performance improvements for a variety of computer vision tasks." Accessed May 9, 2022.

[97] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. 2019. A deep learning framework for neuroscience. *Nature neuroscience* 22, 11 (2019), 1761–1770.

[98] Ingrid Robeyns and Morten Fibieger Byskov. 2021. The Capability Approach. In *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[99] Frank Rosenblatt. 1961. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Technical Report. Cornell Aeronautical Lab Inc Buffalo NY.

[100] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature* 575, 7784 (2019), 607–617.

[101] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536.

[102] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588, 7839 (2020), 604–609.

[103] Terrence J Sejnowski. 2020. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30033–30038.

[104] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[105] Henry Shevlin, Karina Vold, Matthew Crosby, and Marta Halina. 2019. The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. *EMBO Reports* 20, 10 (2019), e49177.

[106] Robert A Skipper Jr. 1999. Selection and the extent of explanatory unification. *Philosophy of Science* 66 (1999), S196–S209.

[107] Elliott Sober. 2003. Two uses of unification. In *The Vienna Circle and logical empiricism*, Friedrich Stadler (Ed.). Kluwer Academic Publishers, New York, 205–216.

[108] Susan Leigh Star. 1989. The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In *Distributed artificial intelligence*. Elsevier, 37–54.

[109] Johanna Thoma. 2015. The epistemic division of labor revisited. *Philosophy of Science* 82, 3 (2015), 454–472.

[110] Lav R Varshney, Nitish Shirish Keskar, and Richard Socher. 2019. Pretrained AI models: performativity, mobility, and change. *arXiv preprint arXiv:1909.03290* (2019).

[111] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218* (2019).

[112] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.

[113] Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (2021), 50–55.

[114] David H Wolpert and William G Macready. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1, 1 (1997), 67–82.

[115] Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Yisen Wang, and Zhouchen Lin. 2021. Training Feedback Spiking Neural Networks by Implicit Differentiation on the Equilibrium State. *Advances in Neural Information Processing Systems* 34 (2021).

[116] Kevin JS Zollman. 2010. The epistemic benefit of transient diversity. *Erkenntnis* 72, 1 (2010), 17–35.