Fixing Foundational Concepts in Machine Learning: A Methodological Primer[1]

Name: Thomas Grote (corresponding)
Email: thomas.grote@utn.de
Orcid: https://orcid.org/0000-0002-9832-6046
Affiliation: University of Technology Nuremberg; Department of Computer Science and Artificial Intelligence; Dr.-Luise-Herzberg-Str. 4, 90461 Nuremberg, Germany.

Name: Alice C.W. Huang
Email: alice.huang@uwo.ca
Orcid: 0000-0002-1719-1945
Affiliations:
Western University Department of Philosophy and Department of Computer Science
1151 Richmond Street, London, ON, N6A 3K7, Canada.
Schwartz Reisman Institute for Technology and Society
108 College St, Toronto, ON M5G 0C6, Canada.

Abstract: Many foundational concepts in machine learning have been criticized as inadequate. Philosophers have therefore taken it upon themselves to sort out the conceptual terrain—with conceptual engineering being the method of choice. This paper takes a step back to provide theoretical and methodological grounding for future work on conceptual engineering in machine learning. To this end, we consider the functional roles of concepts in machine learning, the underlying causes and types of deficiency, and map out criteria for the successful propagation of reengineered concepts within and beyond the machine learning community. Moreover, we discuss how the space of viable conceptual revisions in machine learning is constrained by the need for operationalization, and how tensions can occur between the sociopolitical desirability and the computational implementability of relevant conceptual engineering projects. Overall, our goal is to delineate how conceptual work in philosophy *ought* to be if the goal is for our contribution to permeate through the science and practice of machine learning.

Keywords: Machine Learning; Conceptual Engineering; Methods in Philosophy; Explication; Artificial Intelligence;

[1] Joint first authorship

1. Introduction

Many concepts in machine learning have been criticized as inadequate. For example, the 'bias-variance trade-off', a tenet of standard machine learning theory, says that as a model becomes more complex to fit the training data better, its ability to generalize to new data often decreases, and *vice versa*. It is, however, unable to explain why deep neural networks perform so well (Zhang et al., 2021; Belkin et al., 2019). For 'fairness', a menagerie of different statistical criteria has been proposed, many of which are mutually incompatible (Kleinberg et al., 2016). The notion of 'interpretability', which broadly refers to how easily humans can understand a model, is ill-defined (Lipton, 2018). Moreover, the advent of large language models has drastically increased the nonchalant use of human-centric concepts from cognitive psychology—such as 'general intelligence' and 'theory of mind' – to assess model behavior (Bubeck et al., 2023; Binz & Schulz, 2023).

Against this backdrop, it is hardly surprising that many philosophers have taken it upon themselves to sort out the conceptual repertoire in machine learning. One promising avenue is *conceptual engineering*: In broad strokes, it presents a *revisionary* approach to fixing concepts. Rather than looking for a definition that tracks the extensions of a concept, the starting point is to consider what function said concept ought to play in a given (social, scientific, or philosophical) practice and then modify the concept so that it fulfills this function more adequately (Cappelen, 2018; Haslanger, 2012). Among others, these functions can be combating gender injustice (Haslanger, 2012; Jenkins, 2016), achieving scientific exactness (Carnap, 1950), or resolving logical paradoxes (Sharp, 2013). For instance, in Haslanger's (2000) seminal paper, she argues that, to understand race and gender, instead of studying what the words 'gender' and 'race' refer to in ordinary discourse, we should be asking how to define these concepts so that they serve our political goal of anti-oppression.

Examples of conceptual engineering in machine learning are manifold. Consider some examples: Krishnan (2020) argues that we should reformulate talk about the 'interpretability problem' in favor of a number of more tangible ends (for which interpretability can be a means). Chollet (2019) argues that current benchmarks on narrow skills are ill-equipped to measure intelligence in artificial systems. He reframes 'intelligence' as 'skill-acquisition efficiency', on the grounds of which he develops a benchmark to measure artificial general intelligence. Beigang (2023) modifies the statistical fairness criteria of 'equalized odds' and 'predictive parity' in such a way that they can be satisfied simultaneously, while retaining their intuitive appeal. Fazelpour (2024) expands the

definition of 'trade-offs' by shifting the frame of analysis from a formal setup toward a sociotechnical perspective. Others explore how machine learning practitioners can incorporate conceptual engineering methods into the process of model development and evaluation—with a particular emphasis on the achievement of ethical goals: Köhler (2025) argues that a core part of what data scientists are doing should be understood as conceptual engineering and that this reframing allows us to better consider what is at stake when they, for example, select target variables. Similarly, Rudolph et al. (2025) emphasize how conceptual engineering can be used as a tool in attempts to de-bias large language models; in particular, it can provide normative guidance for machine learning techniques like Reinforcement Learning from Human Feedback.

Our paper takes a step back by looking at conceptual engineering in machine learning through a methodological angle. More precisely, we are interested in three interrelated questions:

1. What kinds of functions do (foundational) concepts serve in machine learning?
2. Where can we expect the need for conceptual engineering to arise?
3. How can we successfully propagate engineered concepts within and beyond the machine learning community?

Tackling these questions allows us to get a better grip on what the bounds are when trying to engineer concepts in a meaningful way. The nature of machine learning constrains the solution space in which revisions are feasible. Unlike ordinary language, which is shaped by everyday use and social conventions, machine learning is highly technical and data-driven. As a result, conceptual engineering in machine learning involves different considerations, such as the use of formal notation, the requirement for computational implementability, and the construction of benchmarks to enable operationalization and support empirical evaluation. The focus of this paper are *foundational concepts*, i.e., concepts that play a crucial role in machine learning theory and the process of model development and evaluation. However, many of the considerations discussed should also generalize to other types of concepts in machine learning.

Our aim is to provide theoretical grounding and methodological guidance for future conceptual engineering efforts. We ask: How can work in philosophy possibly feed back to machine learning research and practice? Research in machine learning is technically demanding, hyper-specialized, and fast-paced. In all likelihood, philosophers will neither solve profound theoretical puzzles

surrounding why highly complex models like deep neural networks seem to defy the bias-variance tradeoff (Belkin et al., 2019), nor invent new model architectures—-at least not while sitting in the front-row. Conceptual work remains the bread-and-butter task. Hence the hope is that our paper improves our understanding of what conceptual work in machine learning is and how it ought to be.

The contribution of our paper is methodological—we define the parameters within which conceptual engineering projects in machine learning should operate: We chart out different varieties of conceptual engineering projects, provide a taxonomy of different functions that foundational concepts in machine learning serve, along with different types and causes of deficiency, and consider the norms and boundaries of successful conceptual engineering projects.

This is how we proceed: Section II spells out a more precise account of conceptual engineering. To this end, we also analyze cases of conceptual engineering in machine learning. Section III discusses the functional role of concepts in machine learning and maps out different types of deficiency. In section IV, we turn to success criteria for the propagation of concepts. Finally, in section V, we distill some key takeaways and point to future research avenues.

2. Conceptual Engineering in Machine Learning

The section sets out to lay down the building blocks of conceptual engineering in machine learning. We discuss different approaches to conceptual engineering, along with examples of conceptual engineering done in machine learning, with the primary aim of highlighting the underlying methodology.

The moniker *conceptual engineering* subsumes a family of methods concerned with the assessment and improvement of concepts (Cappelen, 2018; Isaac et al. 2022; Köhler et al., 2025). What is distinctive about conceptual engineering is that it represents a *revisionary*, as opposed to a *descriptive* approach to sorting out the conceptual terrain. It is guided by the assumption that language is not fixed, but can be changed through intentional activity. In its barest essentials, the process of conceptual engineering starts by identifying a concept deemed to be unsatisfactory in relation to a given epistemic, ethical, or sociopolitical function. A new meaning is then proposed for the concept, one that better serves this function.

A paradigmatic example of conceptual engineering is Haslanger's (2000) work on the concept 'woman'. In contrast to a descriptive approach, which aims to capture how the term is used in ordinary language, Haslanger begins by identifying the sociopolitical goal of resisting oppression. Guided by this aim, she proposes a revised definition of 'woman' that tracks the systematically subordinate role of women in society, in lieu of the merely biological definition that obscures the social position of women. Her approach is sometimes called *ameliorative analysis*.

A different strand of conceptual engineering is *Carnapian explication*. The main idea of explication is to take an informal or vague concept (the explicandum) from everyday language or from science, and refine it, often by way of formalizations, so that we end up with a more exact concept (the explicatum) (Carnap, 1950). As an example, the concept 'fish' in ordinary language includes whales and dolphins. For the purpose of biological theory, however, there are significant anatomical and evolutionary differences between marine mammals and the rest of the aquatic animals. Carnap therefore proposes replacing the folk notion of fish with the more scientifically precise definition of 'piscis'.

Among others, ameliorative analysis and Carnapian explication are two overlapping but non-identical methodological strands of conceptual engineering. Whereas explication, especially through its emphasis on formalization, is methodologically continuous with the sciences, ameliorative analysis is often informed by theories in the humanities and the social sciences, such as work on structural injustice (Dutilh Novaes, 2020). Despite their different emphases, both approaches focus on what a concept *should* mean in order to be practically useful, rather than on how the concept is currently applied.

The methodological difference between descriptive conceptual analysis and revisionary conceptual engineering can also be found in the philosophy of machine learning. It is a well-established problem that deep neural networks, while predictively powerful, are black-boxes—i.e., their underlying decision-rule is elusive to human understanding. In high stakes domains, many have therefore called for the use of inherently interpretable models, as opposed to black-box machine learning models (Rudin, 2019). The problem, however, is that 'interpretability' is ill-defined: It lacks a formal technical definition, and bundles a loosely related set of desiderata and model properties (Lipton, 2018).

Räz (2024) is representative of a descriptive approach to understanding 'interpretability': His strategy is to analyze two textbook cases of inherently interpretable models, namely decision-trees and regression models. The goal is that, by examining properties that make these models amiable to human understanding, we can better understand what we mean by interpretability. Krishnan (2020), by contrast, is emblematic of a revisionary approach: She observes that the term interpretability lacks precise meaning when applied to machine learning models, and moreover, that interpretability is mainly valuable as a means to further ends. The revisionary strategy, therefore, is to sidestep talk of interpretability in favor of a more precise articulation of what kinds of information about the model we need and for what purpose (e.g., facilitating alignment between the machine learning model and human experts).

Another notable case of conceptual engineering in machine learning is Beigang's (2023) explication of the concept of fairness. Several impossibility theorems have shown that two prominent fairness criteria 'equalized odds' and 'predictive parity' cannot be simultaneously satisfied by a machine learning model (Kleinberg et al., 2016; Choudechova, 2017). Some have taken this to indicate that the notion of 'algorithmic fairness' is inconsistent. In response, Beigang modifies these fairness criteria so that they are made compatible. He does so by (i) clarifying the motivation of the different fairness criteria; (ii) modifying the fairness criteria while ensuring that they retain their intuitive appeal; and (iii) proving that the modified fairness criteria can be jointly satisfied.

In machine learning, concepts come in at least three varieties. First, there are *foundational concepts* such as prediction, accuracy, bias, generalization, and so on. These are concepts that play a crucial role in machine learning theory, and are closely tied to the training and evaluation process of machine learning models. Foundational concepts are not model and problem specific. Notable examples here include the aforementioned papers by Rudolph et al. (2025) on 'bias', Beigang (2023) on 'fairness', and Krishnan (2020) on 'interpretability'.

The second category are concepts relevant for specific data science tasks, such as the constructs that machine learning models are supposed to help predict—e.g., creditworthiness, suitability for a certain job, or the risk of relapsing from addiction. What makes these concepts particularly amenable for conceptual engineering projects is that the constructs often cannot be measured directly and the question of how to operationalize them adequately is encroached by moral considerations (Jacobs & Wallach, 2021; Köhler, 2025).

Unlike the first two categories, the third involves cases where the development and deployment of machine learning models disrupt existing conceptual frameworks in science and in ordinary discourse (see also Hopster & Löhr, 2023). For example, the use of large language models is beginning to reshape how we understand the role of machine learning models in scientific discovery. They arguably are no longer mere tools (like a microscope), but have become something more akin to an 'artificial muse' or an 'agent of understanding' (Krenn et al., 2022). In the same vein, it has been argued that the use of generative models in, say, archeology puts pressure on the traditional concept of 'scientific evidence' (Khosrowi & Finn, 2025), or that machine learning models force us to broaden our definition of what constitutes 'toy models' (Sullivan, 2024).

While the second and third category undoubtedly represent important fields of activity for conceptual engineering, this paper will focus on the first variety, foundational concepts in machine learning.

We have seen that conceptual engineering comprises a family of methods aimed at revising concepts to better serve specific functions. In the next section, we examine why certain concepts in machine learning may be inadequate for the roles they are intended to play.

3. Foundational Concepts in Machine Learning: Functions and Failure Modes

Standard machine learning can be characterized by a pipeline that begins with data collection, proceeds through defining a loss function, optimizing parameters to minimize that loss, and concludes with evaluation against benchmarks (Hardt, 2025) or real-world tasks.

Machine learning exhibits qualities of science in two respects.[2] First, the aim of machine learning is usually to increase the accuracy of predictions across a wider range of tasks, and one of the most effective routes to achieving this is through learning causal features of the world from data (Schölkopf et al., 2021). In this sense, machine learning models resemble scientific models, though only partially (Boge, 2022, Duede, 2023; Sullivan, 2024). Second, predictive success or failure in

[2] Conceptual changes often accompany scientific progress. We can see this from the fact that Carnapian explication, a prominent form of conceptual engineering, focuses on refining concepts so as to better serve scientific purposes. Another illustrative example is the evolving definition of 'planet' over time. However, not all scientific work requires conceptual engineering. For instance, understanding genetic dispositions to cancer does not demand reconceptualizing what cancer is, and learning how a kidney functions does not involve rethinking what a kidney is. Our paper concerns only the subset of theoretical development that involves rethinking what certain concepts should mean, and not all theoretical development in science.

machine learning often occurs without complete understanding of why due to complex model architectures. This prompts computer scientists to inquire into model behavior and architecture using theoretical and experimental methods akin to those in scientific inquiries. The science of machine learning is thus driven by the goals of understanding the statistical, computational, and information-theoretic laws governing learning systems, and designing algorithms that enable models to automatically improve by learning from data (Jordan & Mitchell, 2015).

On the other hand, machine learning is also an engineering discipline in three ways. First, while alignment with genuine causal relations is desirable, the ultimate goal is task performance, be it accurate predictions or efficient allocations, which is often achieved through many arbitrary design choices rather than based on deep understanding of causal phenomena. Much like the design choices in engineering, many model choices also have to be made to tackle specific data-science problems, such as predicting how likely a loan applicant will pay back, whether a defendant will relapse if released pre-trial, or the risk of kidney failure for patients in the Intensive Care Unit. Second, machine learning development is constrained by practical factors such as time and computational resources, much like engineering projects are constrained by budget and material availability. Third, machine learning systems have direct societal applications, and their results and applications are subject to regulatory frameworks.

While there is no agreed upon definition of what a concept is, conceptual engineers commonly deem concepts to be cognitive entities that fulfil different epistemic functions (Isaac, 2023; Machery, 2009; Brigandt, 2010). In machine learning, these functions are manifold, reflecting machine learning's position at the crossroads between science and engineering.

In the following sections, we outline the key roles that foundational concepts play in machine learning, both as a scientific discipline and as an engineering practice. While the functions discussed are not exhaustive, they represent what we consider the main functions of concepts in these domains. The first two—*unification (section 3.1)* and *explanation (section 3.2)*—are primarily scientific functions. In contrast, *action guidance (section 3.4)* and *regulatory guidance (section 3.5)* pertain more to machine learning as an engineering discipline. Although we distinguish between scientific and engineering goals in this way, the boundary is sometimes blurred. For example, the function of *guiding development (section 3.3)* is relevant to both domains. We also discuss corresponding ways in which a concept may fail to fulfill each of these functions, and therefore be in need of refinement.

3.1. Unification

One function of concepts in machine learning is *unification*. Foundational concepts reduce theoretical clutter and increase parsimony by allowing us to understand different mathematical entities (definitions, functions, theorems, equations, etc.) as instances of a single notion with shared properties and structures. For example, machine learning involves searching for a solution that minimizes a loss function. The concept of a 'loss function' helps us think more generally about the structure and process of model training across different types of learning (e.g., supervised, unsupervised, and reinforcement learning), different prediction tasks (e.g., classification and regression), and different domains (e.g., image recognition and natural language processing). The concepts of 'overfitting' and 'underfitting' provide a vocabulary for understanding different ways a model can fail – regardless of the model's functional form (e.g., decision trees, deep neural networks, generalized additive models, etc.). The concept of a 'proxy variable', which refers to quantifiable and observable features used to approximate the unobservable feature that we truly want to measure, helps us reason about ways in which a model can fail at problem-solving even when it achieves high accuracy.

Unification plays a crucial role in advancing the theoretical foundations of machine learning. While machine learning is applied to highly practical problems, its scientific core is deeply mathematical. Unified concepts allow us to develop general theories about how to learn effectively and find optimal solutions across broad classes of problems, rather than addressing each problem in isolation.

For example, having a unified notion of a loss function enables us to study algorithms that can reliably identify optimal parameter sets, regardless of the specific loss function used, as long as certain mathematical properties hold. Similarly, unified concepts of overfitting and underfitting allow us to analyze trade-offs that occur in nearly all learning tasks and to design methods to find the sweet spot of model complexity. Finally, a unified concept of proxy variables makes it possible to prove theoretical results explaining why optimizing for a proxy often degrades performance on the true objective we care about (Zhuang, S., & Hadfield-Menell, 2020).

In this way, the role of conceptual unification bears similarity to abstraction in scientific modelling, where some features of reality are deliberately ignored or distorted so that we may focus on understanding the effects of variables most relevant to a given scientific inquiry (Morrison & Morgan, 1999). However, they are different. Concepts are the building blocks of models and

theories. We need first to have unified concepts of particles, pressure, temperature, volume and gas, for instance, to even begin abstracting away from other features of the world to formulate the ideal gas law. Some of these concepts themselves play important unifying roles. For example, the concepts of particles or molecules are themselves unifying concepts that encompass a variety of different entities sharing some common properties and structures.

Unification also plays a sociological role, by demarcating different research (sub)fields. The machine learning research community is vast and hyper-specialized; and many research fields are centered around a specific concept. Take, for instance, the 'fairness'-community, which by now has dedicated conferences[3] and specialized tracks at leading generalist conferences. It is important to note that unification does not require all researchers to share a similar *semantic* definition of what fairness in machine learning is—the opposite is likely to be true. Instead, we are sometimes dealing with a *lexical* definition, where a concept subsumes a variety of interconnected research problems. Applied to 'fairness', these would be the fairness-accuracy trade-off, the problem of proxy discrimination, the de-biasing of models, causal approaches to fairness, or intersectional discrimination. You get the gist. This lexical definition entails certain normative commitments regarding the research problems worth pursuing, the publications worth reading, and the choice of methodology.

Certain concepts can fail to fulfil the goal of *unification* for research purposes when different specialized (sub)fields in machine learning work with their own narrow technical definitions of a concept, but it is unclear what these definitions have in common, if any.[4] As an example, consider the concept 'robustness'. In certain subfields, robustness is taken to mean *adversarial robustness*, the model's ability to maintain reliable performance against examples that are intentionally perturbed in ways that are imperceptible for humans, but that cause the model to make incorrect predictions. In other subfields, robustness is about whether the model can generalize when there is a natural distribution shift due to normal changes in the environment or population. Because robustness is defined in various ways, it is facilitated and evaluated using a range of different metrics and amelioration strategies (Freiesleben & Grote, 2023). Without clarifying what these technical definitions have in common—and what the concept of robustness is actually meant to capture, we

[3] For example, ACM FAccT, the largest conference on the ethics of computer science, is an acronym for fairness, accountability, and transparency.

[4] Note, however, that *unification* and *heterogeneity* are not disjunct categories, but should be thought of as a continuum. Some degree of heterogeneity might be even necessary for new ideas to be taken up by a scientific community. Moreover, a concept whose meaning is heterogeneous can be still useful for the transmission of information within a scientific community, and serves as an indicator of what the subfield considers to be important and relevant.

may end up with seemingly contradictory claims about a model that are not genuine scientific disagreements but artifacts of mismatched definitions. While this could in principle be resolved by making finer distinctions for a concept, if we are not careful, the lack of unifying meaning of a term can mislead downstream practitioners, and hinder the accumulation of coherent knowledge across subfields. Without a shared core meaning of the concept, the term 'robustness' may cease to function as a productive unifying target for inquiry.

3.2. Explanation

Explanation is another function of concepts in machine learning—they help us make sense of certain observed phenomena in machine learning and understand potential interventions. The concept of 'performative predictions' is one such example. A good machine learning model should learn general and meaningful patterns rather than merely memorizing the training data. That is, we want the model to capture stable features in the phenomenon of interest, rather than the quirks and kinks of the specific dataset that it happens to be trained on.[5] For example, we want an image recognition model to be able to distinguish huskies from wolves because it learned the difference in appearance between the animals, not because it learned that photos with huskies tend to have a different background than photos with wolves in the dataset. The difference in appearance is stable, in the sense that most huskies and wolves will exhibit this difference. On the other hand, a photo of a husky could easily be taken with a different background.

For this reason, learning the general structure increases the chance that the model will predict accurately beyond the training data. Sometimes, however, a model can fail to make correct predictions even when it has learned patterns that should be generalizable, or *vice versa,* a model that fails to learn generalizable patterns might nevertheless end up making correct predictions. The concept of performative predictions explains how these situations can occur. Performative predictions are those that change the outcomes they aim to predict by inducing a shift to the underlying probability distribution (Perdomo et al., 2020; Hardt & Mendler-Dünner, 2023). One

[5]The distinction between data and phenomena as articulated by Bogen and Woodward (1988) is helpful here. Data are observations obtained from experimentation, whereas phenomena are stable features of the world that scientists aim to understand and infer from reliable data. However, for Bogen and Woodward, phenomena tend to be unobservable, whereas in machine learning, meaningful features can be observable ones, as long as they are stable features of our target of interest and not merely idiosyncrasies of the specific dataset.

often-discussed type of performative prediction is a *self-fulfilling prophecy*. For instance, predicting that crimes are likely to occur in a neighborhood increases police deployment in that neighborhood, which, in turn, increases the likelihood that crimes in that neighborhood are recorded (Khosrowi & van Basshuysen, 2024). By influencing the data-generating process, the predictions themselves increase the probability that they are correct. This helps us make sense of why sometimes models that do not generalize well nevertheless make accurate predictions when deployed.

As a rapidly developing field, in machine learning, concepts often need to be updated to explain new observations and phenomena. In classical statistics and machine learning, 'regularization' refers to the set of techniques used to ensure that the model does not overfit. For example, in linear regression, LASSO applies penalties to models that have too many parameters to encourage sparsity. In neural networks, DropOut (Hinton et al., 2012, Srivastava et al. 2014) and DropConnect (Wan et al. 2013) set the activations or weights of a subset of neurons in the network to zero during training. These techniques are intentionally designed for this purpose and explicitly applied during training to control what kinds of models we end up with.

However, in light of observations that often no such technique is needed in deep learning to learn generalizable models, along with experiments establishing that the application of regularization techniques cannot explain the generalizability of deep neural networks (Zhang et al. 2021), the classical concept of regularization became too narrow. Thinking about regularization only in terms of the techniques explicitly applied is no longer adequate to explain generalizability. The concept of regularization was thus broadened and the term 'implicit regularization' was introduced to capture this shift. It refers to the hypothesis that certain training procedures, in particular *gradient descent*, have the byproduct of regularization because they are biased towards low-complexity solutions.

3.3. Guiding Development

Aside from the theoretical functions of unification and explanation, concepts in machine learning also *guide the development of new learning algorithms*. Often, it is *models* in science that guide future developments. For example, models of biological evolution have inspired the creation of evolutionary algorithms and genetic algorithms in computer science. Similarly, classical conditioning has influenced the development of reinforcement learning, and models of feed-forward processing in the biological brain have informed the design of artificial neural networks in machine learning

(Chirimuuta, 2024; Sinz et al., 2019). Philosophers of science have argued that one of the goals of models in science is to guide the development of future theories (Weisberg, 2012; Gelfert, 2016; Potochnik, 2017; Shech, 2023). Gelfert (2016), for instance, emphasizes the exploratory role of models that do not aim for accuracy but instead provide 'how-possibly' explanations and serve as starting points for constructing more refined models.

In the previous examples, concepts played only an indirect role as necessary components of models. However, recent developments in machine learning demonstrate that concepts alone can sometimes inspire innovation. The current artificial intelligence boom, driven by large language models, was sparked by a model architecture known as *transformers*. The key innovation in this architecture is a mechanism called *attention* (Vaswani et al., 2017). Interestingly, the development of attention in machine learning appears to have been inspired solely by the concept of attention in cognitive science—the idea of selectively focusing on relevant stimuli while ignoring others. It does not draw on detailed models of how attention operates in cognitive science, nor on its neurological underpinnings in neuroscience. This case seems to be best understood as an instance where machine learning advances were motivated by a concept in isolation, rather than by an entire scientific model.

These borrowed concepts are often used liberally. For instance, concepts like 'memory' or 'attention' in deep neural networks bear little resemblance to their actual counterparts in biological brains (Schaeffer et al., 2022; Zhao et al., 2024). Arguably, this insouciance is less of a problem when the function of concepts is to *inspire* new strands of research.

Problems can arise, however, when we use borrowed concepts to assess and interpret model behavior. For instance, it has become commonplace to assess large language models using standard human-centric reasoning tests from cognitive psychology, and describe their behavior using rich psychological notions like 'general intelligence', 'empathy', or 'theory of mind' – disregarding the fact that these models are mainly optimized for next-token prediction (Bubeck et al., 2023; Binz & Schulz, 2023). This transgression of psychological vocabulary into the domain of machine learning can be troublesome for various reasons. Most notably, we may overestimate the abilities of large language models, since the tests are insensitive to the models' idiosyncratic weaknesses (McCoy et al., 2024). This point is particularly forceful, since it can drive misconceptions about large language models' abilities. In addition, an overly liberal attribution of human-centric mental capacities to machine learning models risks licensing unwarranted inferences about their moral status (Shevlin & Halina, 2019). It has therefore been suggested that, instead of directly borrowing rich psychological

notions, we should develop tests designed specifically for large language models (McCoy et al., 2024).

In some cases, the transgression of concepts into machine learning can also involve philosophy. For example, complex theoretical considerations about justice might be compressed into a simple statistical fairness notion, like in the case of 'equality of opportunity' (Hardt et al., 2016). This is particularly problematic if, in the sense of Goodhart's Law, the measure becomes the target. When the simpler statistical notion replaces the rich philosophical notion, this can in turn erode complex justice-theoretical considerations.

3.4. Action Guidance

Turning to the engineering side of machine learning where model selection takes center stage, an important function is *action guidance*. Consider an example where a developer has to select a model for a data-science task. The concept of 'epistemic risk', the risk that a false positive or false negative error will lead to harmful consequences, helps the data scientist think about which model to select, or where to draw the decision boundary (Biddle, 2022).[6] For certain high-stakes tasks, it may also be crucial that humans understand how a model generates its predictions. In such cases, conceptual clarity about different types of interpretability can help developers choose a model that offers the kind of information required for the context.

Concepts introduced to help us reason about how to make model design choices that align with practical goals can nevertheless fail to do so. The concept of 'interpretability' is an example. One of the main reasons to distinguish between interpretable and uninterpretable models is to constrain which models are acceptable for a given purpose. The concept of 'interpretability' supposedly helps make claims like "this model should not be used in the criminal justice context because it is not interpretable."

However, as Krishnan (2020) has forcefully argued, our current concept of interpretability fails to achieve this purpose because it encompasses too diverse a range of different meanings. For many of these meanings, it is unclear how they contribute to the goals that we take interpretability to be

[6] Note that this characterization captures only one aspect of 'epistemic risk' and needs to be supplemented by representational risks, i.e., a misalignment between what the machine learning model *models* and the target phenomena. For an in-depth discussion of epistemic risk within the context of machine learning , see also Sullivan (2022) and the response by Tamir and Shech (2022). We thank an anonymous reviewer for this helpful comment.

instrumental for. It is therefore unclear how the heterogeneous array of things under the hood of 'interpretability' can guide our model choices. In this sense, the concept of 'interpretability' is idle: it is introduced for the purpose of informing our decisions about model choices and deployment, but fails to do so. Krishnan (2020) therefore proposed that we abandon talk of interpretability altogether in favor of articulating specific properties a model should possess (e.g., the model should reveal the uncertainty of a prediction) in order to satisfy each practical goal we have (e.g., synthesizing human judgment and model prediction).

3.5. Regulation

The last function to be discussed is *regulation*: Facilitating public trust is a fundamental concern for socially relevant applications of machine learning. One possible way to achieve this is to regulate the development and deployment of machine learning models at the national or international level. The EU Artificial Intelligence Act (2024) is arguably the forerunner in that respect.

Definitions of these concepts such as 'fairness', 'privacy' and 'explanation' should retain their intuitive appeal and remain grounded in philosophical understanding. However, a purely philosophical treatment is often difficult to formalize in both legal and mathematical contexts. For one, philosophical debates sometimes focus on accommodating edge cases and thought experiments, whereas legal frameworks prioritize typical cases and practical contexts. Because these regulations govern how models are developed and deployed, definitions must also be amenable to operationalization and be capable of guiding compliance and implementation in real-world machine learning systems. When thinking about these concepts, we therefore need to go beyond philosophical demands and define them in a way that is helpful for thinking about how to ensure regulatory compliance, while also remaining close enough to terminology in machine learning research (Nolte et al., 2025).

One salient problem that arises is the tension between the deliberately high-level general requirements of law and the precise mathematical operationalization required for concepts in machine learning. For example, in the EU anti-discrimination law, legislation is deliberately written with a high level of generality to allow Member States the flexibility to adapt it to their specific legal, cultural, and institutional contexts (Wachter et al. 2021). However, in machine learning, training

models to satisfy anti-discrimination requirements and testing for discrimination both require precise statistical metrics of discrimination. In these situations, the concept of discrimination that we need for machine learning research will be inadequate for the purpose of ensuring legal compliance, since it cannot account for contextual factors.

Finally, we note that the need to revise a concept can also arise not because it fails to serve a particular function, but because of *internal inconsistency*. The concept of 'fairness' in machine learning is a pertinent example. Several well-known impossibility theorems (Kleinberg et al., 2016; Chouldechova, 2017) demonstrate that seemingly reasonable formal definitions of group fairness are mutually incompatible. This presents a challenge for applying the concept of fairness in machine learning contexts and prompts a reconsideration of what kinds of definitions are not only intuitive, but also feasible within technical and normative constraints.

We have outlined a range of functions that certain foundational concepts in machine learning might be expected to fulfill, and illustrated with examples how these concepts can fall short of these roles. The range of functions is broad and each function imposes distinct requirements on how a concept must be refined or repaired. This suggests that revising a concept to better serve one function may inadvertently undermine its ability to fulfill other important functions. In such cases, conceptual engineering involves navigating trade-offs, where gains in one dimension may come at the cost of losses in another.

In the next section, we discuss the success conditions for engineering and implementing concepts in machine learning.

4. Conditions for Propagation

We have seen that the landscape of foundational concepts in machine learning offers many fertile grounds where the need for conceptual engineering can arise. But simply proposing an alternative meaning for a concept is not enough. For philosophical work to meaningfully influence practice in machine learning, we must also consider the question of uptake. As Pinder (2017) argues, if one intends for the new concept to be adopted by a specific audience, such as biologists, policy makers, or, in our case, machine learning practitioners, one should ensure that the newly engineered concept captures what the members of the community take to be the central features and key connections of

the concept. Importantly, the conditions for uptake may vary across communities, depending on their specific practices and norms.

Conceptual engineering for foundational concepts in machine learning is particularly interesting in this regard. On the one hand, it is more straightforward for a revised concept to propagate in the machine learning community, compared to the case of ordinary language. To propagate a reengineered ordinary language concept requires intervening on how people use the term in their daily discourse, which is a slow, large-scale and potentially impossible endeavor (Cappelen, 2018; Issac et al., 2022). To propagate a reengineered machine learning concept, by contrast, it can suffice to simply publish a few influential papers or change the definitions in machine learning textbooks. On the other hand, there are challenges unique to propagating revised foundational concepts in machine learning. Our aim in this section, therefore, is to outline key conditions that facilitate the successful uptake and propagation of reengineered concepts within machine learning.

4.1 Operationalization

The first, and most obvious, condition for an engineered concept to be adopted is *similarity*. A proposed definition cannot stray too far from the original; it must continue to refer, at least roughly, to the same set of entities or phenomena. This is a condition for implementing new concepts not only in machine learning, but more broadly across domains, including ordinary language discourse.

More specifically for machine learning, the proposed new meaning for a concept must be amenable to *operationalization* for uptake to be possible in machine learning. The form this takes will vary depending on the nature of the concept, but some form of implementation is essential for the concept to be actionable in practice.

For example, Babic and Johnson-King (2025) develop an account of fairness that shifts the focus from group-level outcomes common in standard definitions to the decision thresholds for different groups. However, the decision thresholds are not always explicitly encoded in the training algorithm and often need to be inferred from the model's behavior.

Fortunately, there is complementary empirical work that helps obtain the information needed to assess fairness under Babic and Johnson-King's proposal. Simoiu et al. (2017) develop a statistical test for discrimination using a Bayesian hierarchical model to jointly estimate both the decision

threshold and the underlying risk distribution learned by the model. Building on this, Gaebler and Goel (2025) propose a simple yet robust hybrid test to infer whether different decision thresholds are applied across groups. These technical advances make Babic and Johnson-King's account of fairness adoptable in principle, and therefore, possible for uptake.

By contrast, the 'intersectionality problem' in algorithmic fairness is, at least in part, a challenge of uptake. Intersectionality theory examines how occupying multiple demographic categories simultaneously can produce unique effects. For instance, it emphasizes that the disadvantages experienced by a black woman is more than just the sum of the effects of being a woman and the effects of being black (Crenshaw, 1991; Bright et al., 2016). Applied to algorithmic fairness, this insight suggests that our concept of fairness must be enriched to account for combinations of intersecting demographic subgroups, rather than merely treating categories in isolation.

In practice, however, this richer notion of fairness is difficult to implement. The number of intersectional groups (e.g., black Christian women, Asian Christian men, black Muslim women etc.) grows exponentially with the number of demographic categories (e.g., race, religion, gender). This exponential growth leads to smaller group sizes, and the available data become increasingly sparse. Without enough data, statistical tests are meaningless.

While some have proposed statistical methods to operationalize intersectional fairness (Foulds et al., 2020; Morina et al., 2019; Kearns et al., 2018), as Himmelreich et al. (2024) convincingly point out, these methods can face serious limitations. The reason is roughly this: we often need to relax statistical requirements as the group sizes get smaller. However, this effectively lowers the fairness standard for smaller groups that are typically already marginalized. To address this, Himmelreich et al. (2024) propose an alternative, hypothesis-testing-based method that ensures a minimal standard of treatment for all groups, without requiring parity in outcomes. While this approach is better suited to handling small sample sizes, it is also significantly weaker: meeting the minimal standard may still be consistent with substantial disparities between groups.

This illustrates a trade-off in conceptual engineering for machine learning. From a sociopolitical perspective, it is desirable for the concept of fairness to address intersectional forms of discrimination and for its requirements to be suitably stringent. However, the statistical reality is that implementing intersectional fairness is often infeasible due to data limitations. As a result, achieving

practical uptake may require relaxing certain aspects of the idealized notion of fairness, balancing normative adequacy with empirical tractability.

Another example of uptake conditions shaped by pragmatic feasibility concerns the evaluation of model performance. Benchmarks are the method of choice to evaluate progress in machine learning. A benchmark typically consists of a dataset and a metric, which are used to measure model performance on one or more specific tasks. Benchmarks serve as a community-wide standard that allows ranking of the performance of different models on a predefined task (Raji et al., 2021).

Developing the right benchmark is non-trivial, however. Construct validity presents widely shared concern for many benchmarks, i.e., the benchmark dataset and the associated metrics may fail to represent a task or skill of interest in a meaningful way (Raji et al., 2021). For example, the benchmark ImageNet (Deng et al., 2009) used to be the default for ranking model performance for virtually any vision-based task, but some have argued that much of the observed progress might just be a byproduct of models overfitting to that very benchmark (but see Recht et al., 2019). Many have also criticized current assessment methods for being overly human-centric (Schaeffer et al., 2023; McCoy et al., 2024). In particular, benchmarks designed to test human abilities such as the SAT or mathematics competition problem sets are frequently used to evaluate models on specialized tasks (Hendrycks et al., 2020), even though these benchmarks may not align well with the capabilities, purposes, or modes of reasoning that characterize machine learning systems.

While these are legitimate concerns with existing human-centric benchmarks, there are limits to how much we can understand intelligence in a non-human-centric way. Any operational assessment of model performance requires a benchmark. Even if we design new benchmarks that better reflect the nature of artificial rather than human intelligence, these benchmarks are still conceived and constructed by humans. They inevitably focus on tasks that are intelligible and measurable from a human standpoint. As a result, the range of tasks we can assess remains constrained, which, in turn, limits the options for a broader conception of intelligence.

Chollet (2019) explicitly embraces the human-centric nature of assessments and focuses instead on generalizability. Instead of comparing the skill exhibited by AIs and humans at specific tasks such as board games and video games, Chollet (2019) puts forth an account of general intelligence as 'skill-acquisition efficiency'. He argues that solely measuring skill at a given task falls short of

measuring intelligence, because a skill is heavily modulated by prior knowledge and experience. With access to unlimited priors or training data, one can effectively "buy" high levels of task-specific skill, obscuring the system's actual capacity for generalization. Chollet therefore introduces a benchmark designed to measure the model's ability to acquire new skills with limited prior exposure, thereby testing its ability to generalize to unknown new tasks. The benchmark explicitly relies on human priors, and consists of problems that are relatively easy to solve by humans, but difficult for machines (at the time this paper is written).

Chollet's conception of general intelligence propagated widely in the machine learning community and beyond, thanks to the corresponding benchmark that then became a million dollar prize competition. Although having a corresponding benchmark does not guarantee uptake, the lesson is more general—the possibility to operationalize a concept is crucial for its successful propagation. While freeing the concept of intelligence from human-centric limitations opens the door to more advanced and diverse forms of AI, and is no doubt desirable in many other ways, it also raises a significant challenge: how can such a definition be meaningfully evaluated without relying on human-centered criteria? Ultimately, because benchmarks must be designed and interpreted by humans, and must involve tasks intelligible to human evaluators, our operational definitions of intelligence in machine learning remain inescapably constrained by human perspectives.

4.2 Propagating Beyond Machine Learning

In the previous section, we discussed how the need for precise definitions in machine learning can be in tension with the need for flexibility in law. But this is not always the case. The formal definition developed out of necessity in machine learning may also increase its chance of uptake in the policy realm by providing the clarity that policymakers and regulators require.

Take 'differential privacy' for example. The core idea behind differential privacy is that we need mathematically provable, worst-case guarantees on how much protection a privacy mechanism can offer. This is achieved, roughly, by limiting how much estimations of aggregate statistics such as mean and median can change when a single individual's data is added or removed, thereby bounding the potential privacy loss for any one person. In other words, whether or not your data is in the dataset should only change the results by a very small amount determined by the hyperparameters ε and δ (Dwork et al., 2006).

Differential privacy is increasingly adopted by organizations across industry and governments. The fact that it provides provable guarantees expressed through two parameters ε and δ makes it a useful framework for policy decision makers in both private companies and government institutions. It provides the common vocabulary to deliberate and negotiate appropriate privacy standards, and ensures that stakeholders have a shared understanding of the risks and protections involved, based on a transparent worst-case guarantee.

Moreover, advancements in computer science have led to a more sophisticated understanding of privacy, which in turn has broadened the scope of regulatory protections. Privacy concerns in the public sphere no longer focus solely on preventing unauthorized access to databases or hacks. Increasingly, they aim to also protect individuals from inference-based attacks, where adversaries reconstruct datasets or models using publicly available or intentionally released outputs. This illustrates how evolving technical conceptions of privacy in machine learning can influence and reshape regulatory frameworks, extending the impact of technical insights into legal and policy domains.

In summary, we have seen that the space of viable conceptual revisions in machine learning is constrained by the need for operationalization. A concept that cannot be formalized is unlikely to influence practice, and a criterion that cannot be tested risks becoming idle. This does not mean any proposed conceptual change that cannot be implemented has no value. Reengineered concepts that are not implementable may nonetheless reveal the limits of what mathematics can represent and what machine learning can achieve, even without practical and technical uptake. The demand for operationalization is not merely a constraint either. Conceptual changes born within the technical practices of machine learning can migrate outward, informing ethical, legal, and social domains and producing positive changes.

5. Taking Stock and Looking Ahead

The paper's aim was to provide theoretical grounding for future work on conceptual engineering in machine learning. To this end, we considered the functional roles of concepts in machine learning, the underlying causes and types of deficiency, and mapped out criteria for the successful propagation of reengineered concepts within and beyond the machine learning community. Our goal is to offer

methodological guidance for future conceptual engineering projects by delineating how conceptual work in philosophy *ought* to be if the goal is for our contribution to permeate through the science and practice of machine learning. In the following, we summarize some of the key takeaways and chart out avenues for future work.

Because machine learning sits at the intersection of science and engineering, the functions that concepts are expected to fulfill, and the ways in which they may fall short, are diverse. As a result, trade-offs often emerge: a revision that improves a concept's performance in one function may inadvertently hinder its effectiveness in another. Therefore, any satisfactory engineering attempt will have to be made with a very specific context in mind. One particularly important trade-off involves the tension between the sociopolitical desirability of a conceptual engineering solution and its computational tractability and implementability. Concepts that are normatively appealing may be difficult to formalize or integrate into existing technical systems, and *vice versa*.

Our emphasis on computational tractability and implementability can be seen as part of a broader movement within the conceptual engineering literature to integrate empirical methods more systematically (for example, Pinder, 2017; Nado, 2021; Landes, 2025). However, compared to other branches of conceptual engineering, the methods of choice in machine learning differ. Benchmark construction is a prominent example and we expect its importance to still grow. Machine learning is beginning to break free from the paradigm of out-of-sample testing. Many of the most advanced frontier models are generalist, as opposed to being optimized for a single task; and for many of the tasks, there is no clear ground truth, or even *unseen* data available to evaluate their performance (Hardt, 2025). Instead of a mere emphasis on accuracy, the focus of assessments has to shift to (cognitive) *abilities*. The proper operationalization of these abilities and the construction of corresponding benchmarks are intricate tasks, and thus playgrounds for conceptual engineers to explore (see also Harding & Sharadin, 2024; Herrmann & Levinstein, 2025).

The successful propagation of concepts outside of the machine learning community has also been an important topic in this paper. With a view on regulatory aims, we brought up 'differential privacy' as a positive example of how its precise technical definition allowed the concept to receive uptake in regulatory frameworks. Obviously, this is not the final word on the gulf between terminology in law and machine learning. To the best of our knowledge, no systematic account exists yet that spells out

how concepts in machine learning must be structured, so that they can be successfully implemented in regulatory frameworks like the EU Artificial Intelligence Act (2024).

Much remains to be done. Refining foundational concepts in machine learning is not a purely philosophical exercise, nor is it solely a matter of technical sophistication; it is an interdisciplinary undertaking that requires collaboration, as well as negotiation, across philosophy, computer science, statistics, law, and beyond. Only through joint efforts can we ensure our conceptual foundations keep pace with the technical advancement they underpin.

Acknowledgments:

Both authors contributed equally to the conceptualization and writing of the paper. We are grateful to two reviewers, whose feedback greatly improved the paper. We also thank Hong Yu Wong, Konstantin Genin, Bojana Grujicic, Charles Rathkopf, Lily Hu, Huzeyfe Demirtas and Emily Sullivan for helpful discussions.

Thomas Grote acknowledges funding by the Deutsche Forschungsgemeinschaft, Grant/Award Number: BE5601/4-1, while being employed by the University of Tübingen.
Alice C.W. Huang acknowledges funding by the Social Science and Humanities Research Council, Grant Number: 430-2025-00038

Conflicts of Interest: Not applicable

Data availability statement: not applicable

References:

Babic, B., & Johnson King, Z. (2025). Algorithmic fairness and resentment. *Philosophical Studies*, *182*(1), 87-119.

Beigang, F. (2023). Reconciling algorithmic fairness criteria. *Philosophy & Public Affairs*, *51*(2), 166-190.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849-15854.

Biddle, J. B. (2022). On predicting recidivism: Epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, *52*(3), 321-341.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43-75.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303-352.

Brigandt, I. (2010). The epistemic goal of a concept: Accounting for the rationality of semantic change and variation. *Synthese*, *177*(1), 19-40.

Bright, L.K., Malinsky, D., & Thompson, M. (2016). Causally interpreting intersectionality theory. *Philosophy of Science*, *83*(1), 60-81.

Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint* arXiv:2303.12712

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.

Carnap, R. (1950). *Logical foundations of probability*.. University of Chicago Press.

Chirimuuta, M. (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. MIT Press.

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Crenshaw, K. (1991). Race, gender, and sexual harassment. *s. Cal. l. Rev.*, *65*, 1467.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). Ieee.

Duede, E. (2022). Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese*, *200*(6), 491.

Dutilh Novaes, C. (2020). Carnapian explication and ameliorative analysis: A systematic comparison. *Synthese*, *197*(3), 1011-1034.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3* (pp. 265-284). Springer Berlin Heidelberg.

Fazelpour, S. (2024). Disciplining deliberation: A socio-technical perspective on machine learning trade-offs. *British Journal for Philosophy of Science*..

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). Bayesian modeling of intersectional fairness: The variance of bias*. In *Proceedings of the 2020 SIAM International Conference on Data Mining* (pp. 424-432). Society for Industrial and Applied Mathematics.

Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, *202*(4), 109.

Gaebler, J. D., & Goel, S. (2025). A simple, statistically robust test of discrimination. *Proceedings of the National Academy of Sciences*, *122*(10), e2416348122.

Gelfert, A. (2016). *How to do science with models: A philosophical primer*. Cham: Springer.

Harding, J., & Sharadin, N. (2024). What is it for a machine learning model to have a capability?. *British Journal for Philosophy of Science*..

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*.

Hardt, M., & Mendler-Dünner, C. (2023). Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*.

Hardt, M. (2025). *The emerging science of machine learning benchmarks*. https://mlbenchmarks.org

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, 34(1), 31-55.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique*. Oxford University Press.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Herrmann, D. A., & Levinstein, B. A. (2025). Standards for belief representations in LLMs. *Minds and Machines*, *35*(1), 1-25.

Himmelreich, J., Hsu, A., Lum, K., & Veomett, E. (2024). The intersectionality problem for algorithmic fairness. *arXiv preprint arXiv:2411.02569*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hopster, J., & Löhr, G. (2023). Conceptual engineering and philosophy of technology: Amelioration or adaptation?. *Philosophy & Technology*, *36*(4), 70.

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*, *17*(10), e12879.

Isaac, M. G. (2023). Which concept of concept for conceptual engineering?. *Erkenntnis*, *88*(5), 2145-2169.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375-385).

Jenkins, K. (2016). Amelioration and inclusion: Gender identity and the concept of woman. *Ethics*, *126*(2), 394-421.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning* (pp. 2564-2572). PMLR

Khosrowi, D., & Finn, F. (2025). Can Generative AI Produce Novel Evidence?. *Philosophy of Science*, 1-12.

Khosrowi, D., & van Basshuysen, P. (2024). Making a murderer: How risk assessment tools may produce rather than predict criminal behavior. *American Philosophical Quarterly*, *61*(4), 309-325.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Köhler, S. (2025). Good classification matters: Conceptual engineering in data science. *Synthese*, *205*(1), 43.

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., ... & Aspuru-Guzik, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, *4*(12), 761-769.

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*(3), 487-502.

Landes, E. (2025). Conceptual engineering should be empirical. *Erkenntnis*, 1-21.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121.

Morina, G., Oliinyk, V., Waton, J., Marusic, I., & Georgatzis, K. (2019). Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468*.

Morgan, M. S., & Morrison, M. (Eds.). (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge University Press.

Nado, J. (2021). Conceptual engineering via experimental philosophy. *Inquiry*, *64*(1-2), 76-96.

Nolte, H., Rateike, M., & Finck, M. (2025). Robustness and cybersecurity in the EU artificial intelligence act. *arXiv preprint arXiv:2502.16184*.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning* (pp. 7599-7609). PMLR.

Pinder, M. (2017). Does experimental philosophy have a role to play in Carnapian explication?. *Ratio*, *30*(4), 443-461.

Potochnik, A. (2017). Idealization and the Aims of Science. In *Idealization and the Aims of Science*. University of Chicago Press.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Räz, T. (2024). ML interpretability: Simple isn't easy. *Studies in History and Philosophy of Science*, *103*, 159-167.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do imagenet classifiers generalize to imagenet?. In *International Conference on Machine Learning* (pp. 5389-5400). PMLR.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Rudolph, R. E., Shech, E., & Tamir, M. (2025). Bias, machine learning, and conceptual engineering. *Philosophical Studies*, 1-29.

Scharp, K. (2013). *Replacing truth*. Oxford University Press.

Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in Neural Information Processing Systems*, *35*, 16052-16067.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage?. *Advances in Neural Information Processing Systems*, *36*, 55565-55581.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Shech, E. (2023). *Idealizations in physics*. Cambridge University Press.

Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, *1*(4), 165-167.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3), 1193-1216.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967-979.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

Sullivan, E. (2022). How values shape the machine learning opacity problem. In *Scientific understanding and representation* (pp. 306-322). Routledge.

Sullivan, E. (2024). Do machine learning models represent their targets?. *Philosophy of Science*, *91*(5), 1445-1455.

Tamir, M., & Shech, E. (2022). Understanding from deep learning models in context. In *Scientific Understanding and Representation* (pp. 323-340). Routledge.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. (2013, May). Regularization of neural networks using dropconnect. In *International Conference on Machine Learning* (pp. 1058-1066). PMLR.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115.

Zhao, M., Xu, D., & Gao, T. (2024). From cognition to computation: A comparative review of human attention and transformer architectures. *arXiv preprint arXiv:2407.01548*.

Zhuang, S., & Hadfield-Menell, D. (2020). Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, *33*, 15763-15773.

Legal Documents:

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)