Synthese (2023) 201:51
https://doi.org/10.1007/s11229-022-03999-y

ORIGINAL RESEARCH

Machine understanding and deep learning representation

Michael Tamir[1] · **Elay Shech**[2]

Received: 7 July 2021 / Accepted: 3 December 2022 / Published online: 30 January 2023

Abstract
Practical ability manifested through robust and reliable task performance, as well as information relevance and well-structured representation, are key factors indicative of understanding in the philosophical literature. We explore these factors in the context of deep learning, identifying prominent patterns in how the results of these algorithms represent information. While the estimation applications of modern neural networks do not qualify as the mental activity of persons, we argue that coupling analyses from philosophical accounts with the empirical and theoretical basis for identifying these factors in deep learning representations provides a framework for discussing and critically evaluating potential machine understanding given the continually improving task performance enabled by such algorithms.

Keywords Machine learning · Deep learning · Artificial intelligence · Understanding · Representation · Information theory

1 Introduction

Advances in machine learning (ML), especially using deep learning (DL) techniques, have accelerated performance in numerous areas of practical application. One metric worthy of attention is the rate at which DL has enabled algorithms to compete with human benchmarks on specific tasks. Image classification, for instance, has evolved dramatically thanks to a series of specific improvements in DL, including convolutional neural network (CNN) and more recently vision transformer (ViT) architectures coupled with technical advances in the optimization of neural networks with multiple

[1] This article belongs to topical collection : Philosophy of Science in Light of Artificial Intelligence edited by Atoosa Kasirzadeh, Sarita Rosenstock, and David Danks

✉ Michael Tamir
mike.tamir@berkeley.edu

✉ Elay Shech
eshech@auburn.edu

[1] University of California, Berkeley, California, USA

[2] Auburn University, Auburn, Alabama, USA

hidden layers, leading to DL beating human performance on the ImageNet benchmark data set (He et al., 2015). AlphaGo's defeat of Lee Sedol in 2016 is another celebrated example of DL in interactive reinforcement learning contexts. Similarly, better than human deep reinforcement learning successes were achieved by OpenAI in Dota2 competitions, and CMU's Libratus and Plaribus poker algorithms. More recently, tasks in modern natural language processing (NLP) have also seen ostensible breakthroughs by becoming competitive with human performance. Hassan et al. (2018) achieved parity with human translation on the WMT17 benchmark, leveraging DL Transformer architectures. Transformer architectures have also inspired a wave of advances leading to performance increases on the general language understanding evaluation (GLUE) benchmark (Wang et al., 2018), overtaking non-expert human performance in (Nangia & Bowman, 2019). Similarly, over a dozen DL Transformer based techniques have bested human performance scores on the Stanford Question Answering Dataset 2.0 (SQUAD 2.0) (Rajpurkar et al., 2018).

Human competitive performance on such benchmarks has accompanied an increased use of terms like "understanding" in artificial contexts. Machine understanding of natural language applications is commonly discussed by researchers both in terms of task goals as well as model capabilities. The GLUE benchmarks in "Natural Language Understanding" (NLU) tasks are framed in terms of "aspir[ing] to develop models with understanding beyond the detection of superficial correspondences between inputs and outputs" (Wang et al., 2018, p. 353). The SuperGLUE benchmark lists as the first criteria that "[t]asks should test a system's ability to understand and reason about texts" (Wang et al., 2019, p. 4). Devlin et al. (2019, p. 4174) motivate specific techniques "[i]n order to train a model that understands sentence relationships," while Raffel et al. (2020) more modestly claim that techniques such as those of (Devlin et al., 2019) "can be loosely viewed as developing general-purpose knowledge that allows the model to 'understand' text." Researcher discussion of machine understanding is even bolder in areas focused on DL representation learning. Bengio et al. (2013) influentially framed conversations on machine understanding in terms of disentanglement, arguing that "the ultimate goal of AI research is to build machines that can understand the world around us, i.e., disentangle the factors and causes it involves." Chen et al. (2016) motivate using generative techniques with "the belief that the ability to synthesize, or 'create' the observed data entails some form of understanding, and it is hoped that a good generative model will automatically learn a disentangled representation," and Higgins et al. (2017) claim that representations with disentangled factors are "an important precursor for the development of artificial intelligence that understands the world in the same way that humans do." DL successes coupled with such loose (if not bold) claims about potential machine understanding have prompted responses from intersecting research in cognitive psychology (Lake et al., 2017; Marcus, 2020) and linguistics (Bender & Koller, 2020). These responses make the easy case that models trained for human competitive performance in specific tasks fail to possess what Marcus calls "deep understanding" (characterized as the kind of understanding found in humans), citing failures to perform when "circumstances deviate from training data" (Marcus, 2020). While few claim that current algorithms possess such "deep" or "human level" understanding, the more interesting question of which conceptual criteria are appropriate for evaluating (partial) machine

understanding has not received sufficient attention. Can the philosophy of science and epistemology literature on understanding shed light on which conceptual criteria are important for machine understanding? Are there trends and patterns in how DL trained algorithms process data from a representation and information compression standpoint that could partially or fully satisfy such conceptual criteria? If so, do such patterns provide insight into critically evaluating and interpreting the relevance of concepts like "understanding" in an artificial context?

In this work, we answer these questions, identifying three key factors taken from the philosophy of understanding literature which we argue have a basis for evaluation in DL trained algorithm performance and learned data representations. Our aim is twofold. First, viewing DL successes in the context of philosophy of understanding may shed light on the extent to which references to "understanding" in DL research have any grounding (or not) in traditional analyses of the concept. Second, the philosophy of understanding literature provides valuable resources for identifying the conceptual criteria that are appropriate for evaluating and comparing any partial applicability of concepts like "understanding" to machines. Using these resources, we identify methods for experimentally detecting the presence of key factors indicative of understanding, allowing for future evaluation of potentially more complete or so called "deep" machine understanding in the rapidly evolving field.

We lay out the paper as follows. Drawing from select philosophical accounts of understanding in Sect. 2, we identify reliable and robust task performance, as well as information relevance and well-structured representation, as three key factors. In Sect. 3, we provide a brief introduction to ML and DL practices. While even successful individual DL trained algorithms are not minded agents, in Sect. 4 we show how phenomena analogous to said factors can be observed and evaluated in DL applications through an information theoretic analysis. Specifically, deep neural networks use multiple layers of representations that systematically learn to extract and organize relevant information, and this process directly relates to methodologies used by DL researchers to ensure reliable and robust success. Information relevance is learned by the neural network, preserving task-relevant information in deeper hidden layer representations of the raw data. When successful, learned representations develop insensitivity to unimportant factors while optimally leveraging and organizing the relevant features, thereby disentangling raw details in deeper layers based on (task) significance. Section 5 ends the paper with a consideration of three related objections to evaluating "understanding" in the context of automated task performance. Our goal is to establish a discussion framework for understanding in ML and DL, and to encourage future investigation grounded in philosophically coherent terms and direct engagement with the technology driving these accomplishments.

2 Three key factors of understanding

The philosophical literature on understanding is large and growing,[1] but certain common factors stand out. Drawing on selective accounts, and working with guiding examples, we identify three key factors of understanding. Before doing so, let us consider three caveats: First, the key factors that we identify are considered by various philosophers to be *constitutive* of understanding, so that their presence is either necessary and/or (jointly) sufficient for understanding. We wish to remain neutral on such issues and so we speak of said factors as *indicative* of understanding instead of conditions constitutive of understanding. That is, said factors may provide (defeasible) evidence that understanding is present or that there is comparatively more understanding. Second, other key factors not discussed here may also be worthy of investigation. Third, it has become common to note that understanding admits *degrees* (Khalifa, 2017; Shech, 2022) and is *gradual* in the sense that it can "vary in breadth, depth and accuracy" (Baumberger, 2014, p. 83). This provides a basis for finer distinctions among factors (where some factors promote greater understanding) and generally fits well with how we consider potential machine understanding.

To begin, consider the prosaic example of evaluating a child's understanding of cat identification[2] If the child can remember every cat in the neighborhood, and can identify each one as a cat, we might suspect they understand how to identify cats. Successful task performance is indicative of understanding. Looking at the philosophical literature, it has been noted by various authors that practical ability in task performance is especially telling of understanding. For example, Catherine Elgin holds that "[understanding] physics is not merely or mainly a matter of knowing physical truths. It involves ... a *capacity to operate successfully*..."(Elgin, 1993, pp. 14-15). De Regt and Dieks (2005) characterize understanding in terms of what it enables one to do. This idea is encapsulated by De Regt's claim that the "quintessence of scientific understanding lies in the *ability to perform a difficult task* rather than in knowing the answer to a difficult question" (De Regt, 2015, p. 3790; our emphasis).

Of course understanding requires more than task completion in specific instances. In more recent work, De Regt and Gijsbers (2016, 55-56; our emphasis) suggest an effectiveness condition on understanding: "Understanding can only be gained from representational devices that are, for a subject in a context, *effective*... [wherein a] device is effective just in case the device is *usable* by the scientist and using it *reliably* leads to scientific success...." If a new cat were to come by and the child could not classify this cat because, for example, it is not one of the original memorized cats, then arguably they do not (yet) understand very well. In contrast, a child who can identify new cats and correctly identify what are (and are not) cats when tested has more understanding because of this *reliability*. Their ability to generally complete the task in new instances is a mark of the kind of reliability indicative of understanding. Compare, for instance, with Mark Newman's account of theoretical understanding,

[1] See, e.g., (De Regt, 2009; Grimm et al., 2017; Lawler et al., 2023) for recent contributions and surveys of the literature. See (Tamir & Shech, 2023; Shech & Tamir, 2023) specifically for understanding phenomena with ML models and Shech (2022) for the roles of idealizations in facilitating understanding.

[2] To be clear, the target of understanding in this example is the task of cat-identification and not full understanding of the cat itself.

which holds that: "[Subject] S understands scientific theory T if and only if S can *reliably* use principles, Pn, constitutive of T to make goal-conducive inferences ... that *reliably* results in solutions to qualitative problems relevant to that theory" (Newman 2017, p. 582; our emphasis).

Beyond performing a task successfully in additional instances (what we will call *reliability*), the ability to perform a related new type of task (what we will call *robustness*[3]) is especially indicative of understanding. This point has been emphasized in Alison Hills' account of understanding why p, which is constituted in part by the ability to "draw the conclusion that p' (or that probably p') from the information that q' (where p' and q' are *similar but not identical* to p and q)" and assuming that q is why p (Hills, 2016, p. 663; our emphasis). That is to say, she emphasizes that the ability to extend successful application to new similar tasks, as symbolized in p', is essential to understanding. Similarly, Wilkenfeld (2019, p. 2815) says that "[r]eal understanding requires the ability to take what is in one's cognitive possession and apply it to a new array of cases," providing examples of generating a "new consequence of a mathematical theory" and "wholly new proof" as variations of this novelty. These positions reflect that understanding requires more than reliability of repeated success for further instances of the same type of task. Extending successful performance to new (related) tasks provides further evidence of understanding. For our cat identification example, if the child can identify cats appropriately in photographs, as cartoon drawings, as species of large cats or other new (but related) task applications, then the child has (better) understanding. They can apply this skill both reliably and robustly, that is to say, they can do it repeatedly for the same type of task in question (e.g., more house cats), and for related but non-identical types or variations of the tasks (e.g., cartoon cats or apex predator cats). Using this terminology, we have the first key factor of *reliable and robust task performance*: the ability to perform a task in additional instances (reliability), and the ability to perform a similar or related novel type of task (robustness).

We consider three immediate worries. First, one may object that while reliable and robust task performance is perhaps minimally needed for understanding, it falls short of sophisticated theoretical and scientific understanding. A second (related) worry might be that while task performance may indicate "practical understanding," it isn't as clear that it is important for the central notions "explanatory understanding" and "objectual understanding" (e.g., (Baumberger, 2014; Khalifa, 2017; Stuart, 2018)). In reply to the first worry, our aim modestly focuses on specific factors of understanding that we will argue have a basis in ML contexts. Theoretical and scientific understanding plausibly have additional requirements, but, we submit, also place importance on reliable and robust application as indicated above. For the second worry, reflecting on the literature, there is a strong tradition of accounts of understanding that incorporate some aspect of practical ability, say, the ability to give an explanation, make an inference, solve a problem, manifest a skill, etc. In fact, Hills (2016), Baumberger (2014),

[3] For the purposes of this work, we use the concepts of reliability and robustness strictly as shorthand for successful repeated application of a task, viz., reliability, and application of a new related type of task, viz., robustness. We *do not* intend our usage of robustness to be understood in relation to the notion of robustness analysis in, e.g., (Schupbach, 2018; Stegenga & Menon, 2017). Further, our usage of robustness should not be confused with adversarial DL concept of "non-robust features" in (Buckner, 2020).

and (most explicitly) Stuart (2018) note that it may be possible to reduce accounts of explanatory and objectual understanding to practical understanding, wherein ability/skill takes center stage. However, since our objective does not concern questions about whether "know-that" can be reduced to "know-how" (or vice-versa), we qualify that by "understanding" we focus here on an ability that can (at least sometimes) be measured by the observable result of performing a task. We set aside the issue of how this may be related to or compatible with other notions such as understanding why some phenomenon happened (Khalifa, 2017), understanding some proposition (Hills, 2016), understanding with a scientific theory (Strevens, 2013), understanding a theory itself (Shech, 2022; Newman, 2017), or understanding as grasping (Baumberger, 2014).

Third, one may worry that the concept of a similar or related novel task type in our articulation of robustness is unclear, or at least not easily distinguished from what we describe as reliability, namely, repeated success on different application instances. Whereas reliability is intended to cover cases where one can successfully perform what may be fairly described as additional instances or tokens of the same type of task, robustness covers cases that might fairly be described as a different sort of task type. In other words, where reliability is about one's "depth" of ability in a particular sort of example (same task type), robustness is about one's ability when it comes to the "breadth" of applications they can navigate (different task types). This is clear in Hills' own account of understanding why. Namely, in her discussion of how propositional knowledge is not sufficient for understanding, she emphasizes that the type of cognitive control required for understanding depends on "the ability to draw conclusions yourself in a new case" (Hills, 2016, p. 671). She says that while "to have knowledge, you do not need to be able to judge new cases correctly," but that the application to new cases is "essential to understanding" (p. 670). Thus, as noted above, our concept of robustness is grounded in Hills' conception of application to "similar but not identical" cases and Wilkenfeld's talk of "a new array of cases."[4]

Returning to our guiding example, a child with understanding of cat identification should not be thrown off when, for example, they cannot see the entire cat, when the cat is sitting on a countertop instead of a sofa, while wearing a new collar, etc. A child thrown off by these irrelevant details understands less well than a child who can ignore what is unimportant and attend to what is important in the available information. As Wilkenfeld notes, "attributes *relevant* to determining [the] degree of understanding in some particular context are those that enable one to make the types of inferences and perform the types of manipulation that are *relevant* in that context" (Wilkenfeld, 2013, pp. 1007-1008; our emphasis). It is only the "relevant information" that matters. Thus, an "account of understanding should reward an agent for being able to produce more of the kind of information picked out as relevant by the context" (Wilkenfeld,

[4] Ultimately, whether performing a particular task will count as an additional instance of the same task performance (reliability) or a new type of task performance (robustness) may not have a precise boundary and may for general understanding be somewhat case dependent. However, in our Sect. 4 discussion of ML models, we elaborate how reliability can be substantiated by consistent task success on data ostensibly described by the same distribution, whereas robustness can be substantiated by enabling task success on data described by different distributions (e.g., data generated from different sampling processes or with different features).

2019, pp. 2809, 2810). Accordingly, we identify as a second key factor of understanding, *information relevance*: the ability to represent the relevant and only the relevant information useful to a task or tasks.

Information relevance is not entirely orthogonal to reliable and robust task performance. Mastering reliability when different irrelevant details are introduced in additional circumstances is important. Similarly, being able to represent what matters to cat identification can be essential to robustness, like understanding which details are relevant to why Garfield is a (cartoon) cat despite the many differences from a physical house cat that, in this context, are not relevant. While detecting reliability and robustness can be couched in observable criteria of responses to examples, information relevance is not necessarily directly observable from task responses themselves. What contrasts the second key factor from the first is that it places further requirements on how the understander *represents the relevant information*.

Wilkenfeld highlights that the extent that an understander's internal representation captures what is useful, while removing the irrelevant information, is indicative (or, for him, constitutive) of understanding. Specifically, Wilkenfeld takes understanding to be essentially *representation manipulability* (URM): "A statement, attributed in context C, that thinker T understands object o, is true if and only if T possesses a mental representation R of o that T could (in counterfactuals salient in C) modify in small ways to produce R', where R' is a representation of o and possession of R' enables efficacious (according to standards relevant in C) inferences pertaining to, or manipulations, of o" (Wilkenfeld, 2013, pp. 1003-1004). More recently, he has articulated a different account of understanding as *compression* (UC), which also emphasizes the comparative and incremental aspect of understanding: "A person $p1$ understands object o in context C more than another person $p2$ in C to the extent that $p1$ has a representation/process pair that can generate more information of a kind that is *useful* in C about o (including at least some higher order information about which information is relevant in C) from an accurate, more minimal representation"(Wilkenfeld, 2019, p. 2810).[5] He highlights how information minimization viewed as compression depends on how the representation organizes the (relevant) information. Particularly, if there are regularities in the relevant information, then the extent to which an understander can identify and leverage those regularities in how they represent the important information, the better their understanding. Importantly, someone who is better able to leverage structure and regularities in a task to abstract and organize how they represent what is important has more understanding; they are better able to represent the relevant information present in the range of potential instances for which the task may be completed.

For example, compare one driver who memorizes every list of directions for any pair of locations they want to travel between in a city with a second driver who instead

[5] Since Wilkenfeld is identifying what he takes to be *constitutive* of understanding, the URM and UC accounts are rival accounts. For example, (Wilkenfeld, 2019, p. 2828) notes that while "URM claims that one's understanding of (for example) the soundness proof consists in the fact that one could correct small mistakes and that one could potentially prove soundness for other logical systems... UC predicts that they are neither necessary nor sufficient." Again, we do not commit to a particular view about what constitutes understanding. Instead, we are focused on leveraging a multitude of philosophical accounts to guide our identification of key factors that are (at least) indicative of understanding in order to show that such factors can be used as (partial) criteria for appraising machine understanding.

learns a map of the city and all the important roads connecting those same locations. The latter driver has a better understanding of navigation tasks (in this city) than the former. Assuming the former driver directions are suitably spartan, e.g., there are no details in the directions about the scenery except for essential landmarks like street signs, we might argue that both drivers only leverage relevant information, but there are still (at least) two important differences. First, contingent on the complexity of the city and location pairs, it is plausible that the directions-driver has "more things to remember" than the map-driver. Several of the direction lists may have redundant overlap in sequences of steps across lists, which could be better understood given a map of said overlapping paths. Eliminating these redundancies as with the map-driver requires less aggregate encoding of (ostensibly equivalent) information. Second, if a roadway, for instance, were temporarily blocked, the direction-driver could not complete any navigation requiring it, whereas the map-driver could more easily adjust to alternative routes.[6]

The map-driver has better organized the relevant topological graph structure of the roadways. By organizing their representation of the relevant navigation information, the map-driver can remove redundancy of what needs to be remembered (and represented). The navigation information is relevant for both the map-driver and the memorization-driver, but more efficient for the map-driver in, for example, the sense that they would have less to "write down" (in total) using their method of expressing said navigation information than the other driver. More importantly, by separating out the relevant information *in a structured way*, the usefulness of their representation is less brittle with respect to changes. This example is analogous to when Wilkenfeld (2019, p. 2808) compares the capabilities of his first-order logic students who have memorized "a derivation of De Morgan's Laws in the Lemmon-Mates system, which has something on the order of 20 or 30 steps" versus his own deeper understanding: "Because I understand, I don't need to have it memorized—I have more basic facts memorized, and a bunch of rules, heuristics, and hypotheses that let me recreate the proof fairly easily." After all, if we slightly perturb the steps memorized by a student for the derivation of De Morgan's Laws (e.g., if they forget or misremember a step), they couldn't recover: "if [they] had forgotten one step-even potentially a relatively trivial step-in the middle of the proof, [they] would have been lost" (2814). Memorization does not involve learning "enough higher-order information" in order "to reconstruct the proof from simpler representations" (2812-2814). Someone with Wilkenfeld's self-described understanding of the basic facts involved with proving De Morgan's Laws is immune to such mistakes; if Wilkenfeld misremembers a particular step, in the course of deriving the proof from his representation of more basic facts and rules the memory error could be discovered and corrected.[7] Our third key factor

[6] Note that there is an ambiguity between the modal notion of "potential compressibility" and the non-modal notion of "actual compression." The (Wilkenfeld, 2019) account leverages actual compression. In our discussion of the third key factor indicative of understanding we likewise identify actual compression as what is important for understanding. In relation to our example, both the map-driver and the memorization-driver may have the same amount of potentially compressible navigation information, but the map-driver's information is actually more compressed.

[7] Of course, Wilkenfeld's representation is not immune to any error whatsoever (e.g., forgetting basic deduction rules) but according to the argument, as a matter of degree, there are fewer critical failure points from which he could not recover.

terms this importance as *well-structured representation*: the ability to structure how (relevant) information is minimally represented to make it efficacious in a task or tasks (even or especially under perturbations).

To be clear, while *information relevance* concerns an ability to represent only the relevant information for tasks of interest, *well-structured representation* concerns the ability to represent said relevant information in a manner that is minimal and less sensitive to perturbations. We have attempted to illustrate this idea with the example of a memorization-driver and map-driver: both have relevant information, but the map-driver better organizes the relevant information as a topological graph structure such that (i) the map-driver's information is more minimal (because there is no need to memorize redundant sub-lists of directions when paths overlap) and (ii) their task-completion ability is less sensitive to perturbations (say, if a roadway were temporarily blocked).

Last, it is evident from accounts of understanding that are articulated via "Subject *S* understands...," with explicit mention of a "thinker," a "person," and a "mental" representation (as in URM and UC), that there is an additional "minded agent" or what we might summarize as a "mentality" factor of (or constraint on) understanding. To be clear, we do not claim here that conscious mentality or "strong AI" exists in the contemporary ML or DL trained algorithms we consider. Since our focus is on formalizing indicative factors for the purposes of grounding the notion of machine understanding, we delay discussion of this excluded "mentality" condition for Sect. 5.

In sum, reflecting on the philosophy of understanding literature, we have three key factors identified as substantive potential indications of understanding:

1. **Reliable and robust task performance**: the ability to perform a task in additional instances (reliability), and the ability to perform a similar or related novel type of task (robustness).
2. **Information relevance**: the ability to represent the relevant and only the relevant information useful to a task or tasks.
3. **Well-structured representation**: the ability to structure how (relevant) information is minimally represented to make it efficacious in a task or tasks (even or especially under perturbations).

Given appropriate characterization, as we provide below, such factors (hereafter the key factors) can be used as (partial) criteria for appraising machine understanding. Foreshadowing what's to come, reliable and robust task performance corresponds to low generalization error on test data sampled from the same process and a reduced need of further training for novel applications (respectively). Information relevance can be evaluated through information bottleneck analysis techniques substantiated empirically by nuisance insensitivity. Well-structured representation corresponds to representation disentanglement measured through geometric clustering and measures of mutual information with respect to a representation.

3 Machine learning fundamentals

ML can be loosely construed as an algorithmic process for generating an estimator $f : x_i \mapsto \hat{y}_i$ that assigns "output elements" $\hat{y}_i \in Y$ to "input elements" in a data set $\{x_i\}_{i \in N} \subset X$.[8] While the multiplicity of different estimation procedures f is quite broad, f is typically parameterized by some set of parameters $\theta \in \Theta$, establishing a family of estimators $\{f_\theta\}_{\theta \in \Theta}$ for the given estimation procedure. What distinguishes the generation of a particular estimator as an ML process from directly designed rule-based algorithms is that in ML, the selection of f_θ is accomplished through a process of optimizing the parameter values θ so as to best "fit the data" according to a prescribed objective. In what is called *supervised* ML, the parameters defining the estimator f_θ are selected by taking a set of sample pairs $\{(x_i, y_i)\}_{i=0}^{N} \in X \times Y$, called the *training set*, and for a given loss function $L : Y \times Y \rightarrow \mathbb{R}$, finding the optimal parameterization θ^* such that:

$$\theta^* = argmin_{\theta \in \Theta} \sum_{i=0}^{N} L(f_\theta(x_i), y_i)$$

The process of finding the optimal $f_{\theta*}$ for a given training set is called *training the model*.[9]

An estimator f_θ does not need to be complex in ML. One familiar basic example that counts as supervised ML is Ordinary Least Squares (OLS). In OLS, the input data points $\mathbf{x}_i$ are vectors of dimension m and the estimation procedure f_θ and loss function L are given as follows:

$$f_{\mathbf{w},b}(\mathbf{x}_i) := \mathbf{w} \cdot \mathbf{x}_i + b = \hat{y}_i$$
$$L(\hat{y}_i, y_i) = ||\hat{y}_i - y_i||_2^2$$

where the model parameters are just $\mathbf{w}$, a vector of dimension m, and the bias term b. Often raw input data $\mathbf{x}_i$ are transformed into some new mathematical representation $\mathbf{x}_i \mapsto \mathbf{x}'_i$, through a process called *feature engineering*, prior to training. Especially for simpler estimator processes, engineering new features $\{\mathbf{x}'_i\}_i$ from the original $\{\mathbf{x}_i\}_i$ through handcrafted transformations has traditionally been an essential step in successful ML training (Domingos, 2012). In contrast, more contemporary DL techniques reduce such feature engineering, instead shifting towards methods of leveraging ML optimization to automate this process.

DL designates a suite of related ML estimation procedures constructed through the composition of linear and nonlinear transformations referred to as a *neural network*.

[8] Alternatively, f may map to a scoring function (like a distribution) over output elements.

[9] Note, for simplicity we are here glossing over several potential alternative ML paradigms to supervised learning such as unsupervised learning and reinforcement learning paradigms in addition to exceptions and modifications including details in how the loss function is optimized (e.g., variations of gradient descent), Maximum Likelihood Estimation versus Bayesian approaches, hyper-parameterization and selection of the family of parameterized functions, etc. f may even have components that are stochastically altered during the training process (e.g., dropout), and so may not strictly speaking even be a deterministic mapping (at least for the purpose of training). More comprehensive presentations of ML fundamentals can be found in, e.g., (Murphy, 2012).

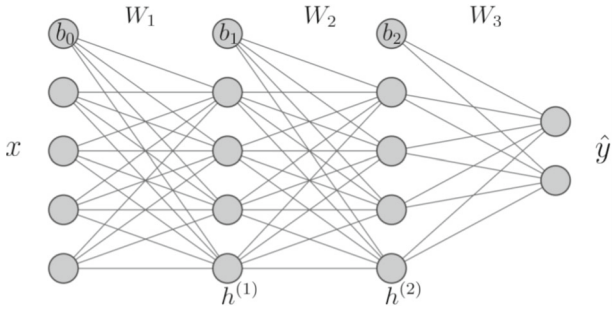

Fig. 1 Fully connected artificial neural network diagram

In the simplest style of neural network, we may set:

$$f_\theta = \sigma_L \circ z^L_{(W_L,b_L)} \circ \ldots \circ \sigma_1 \circ z^1_{(W_1,b_1)}$$

where for each $l \in 0, ..., L$, σ_l represents some non-linear transformation called the *activation function*, $z^l(h) = W_l h + b_l$ is an affine transformation parameterized by a linear transformation W_l and bias vector b_l, and the $h^{(l)}$ called the *hidden variables* are the respective outputs after each "layer" of transformations $h^{(l)} = \sigma_l(z^l(h^{(l-1)}))$ with the convention $h^{(0)} := x$ (see Fig. 1). Researchers have innovated several alternatives to this (simplest) style of neural network, called *fully connected*, that help to better manage the number of parameters needed to efficiently transfer information from one hidden layer to the next, including pruning connections, convolving hidden layers with local kernel functions (CNNs), keeping track of "cell" states whose updates are managed by sigmoidal gating functions, generating multiple pipes of hidden vectors fed through the neural network to compare relative positions using "attention" mappings or to linearly combine them at later stages, etc.

Generally, neural networks can be characterized as multiple layers of parameterized transformations of the data whereby said parameters are "learned" through ML optimization. The application of ML through layers of transformations is an example of the general concept of *representation learning* which is a central focus for the remainder of this paper. Representation learning can be defined as the process of using ML to optimize the parameters involved in transforming how the input data is mathematically represented to achieve some objective. In the case of DL, transformed data representations are the hidden layer outputs $h^{(l)}$, whose transformations are optimized to reduce the final estimator loss.[10] DL and more generally any representation learning algorithm helps to automate the feature engineering process by leveraging the

[10] In the context of representation learning, while the vector (tensor) outputt $h^{(l)}$ at each hidden layer is characterized as a "representation" of the original input data, in the sense that it is some mathematical transformation of the input data carrying (representing) the information in the input x data that is optimized for the ultimate ML estimation task, the utility of said representations (layers) is not equivocal. Deeper layers, being closer to the final output estimation layer, can be expected to be transformations to vector (tensor) representations better suited for the ML estimation task. Further, in Sect. 4.3 (below), we see how select hidden layer representations are commonly repurposed by ML researchers as part of a transfer learning process to new tasks as with text embedding methods (Devlin et al., 2019; Mikolov et al., 2013; Liu et al., 2019) or self-supervised generative autoencoding methods (Higgins et al., 2017; Xiao & Wang,

same data driven optimization techniques used to parameterize some final estimation (e.g., in neural networks the final "output" layer) in order to learn the optimal representation of the input data. *Our proposal is that investigating the nature of learned representations provides a basis for more deeply engaging in questions of machine understanding.*

4 Deep learning and the key factors

In Sect. 3 we explained how DL can be directly viewed as a framework for extensive representation learning; learning the transformations of each hidden layer of a neural network is an iterated process of learning ML optimized representations to improve the representations of the layers before. In Sect. 2 we identified three key factors prominent in philosophical analyses of understanding. In this section, leveraging an information theoretic analysis of the representation learning that occurs in DL, we argue that there are experimental and theoretical means of evaluating (1) reliable and robust task performance of the DL model, (2) the relevance of information in hidden layer representations (information relevance), and (3) the organization of information in hidden layer representations through disentanglement (well-structured representation). The presence of each of these three factors is a matter of degree (in both humans and machines). Again we neither claim that the three factors are *jointly sufficient* (in the classical conceptual analysis sense), nor that the presence of any key factor exists to a *sufficiently high degree* in any known algorithm to claim that it "understands." However, we argue that the ability to detect the presence and degree to which these key factors can be found in successful DL applications provides a basis (at least partially) for future evaluation of machine understanding to the extent that the key factors are satisfied.

4.1 Reliable and robust task performance

In Sect. 2 we identified the first key factor, reliable and robust task performance, in terms of the following: "reliability" means success completing a task in additional cases, and "robustness" means successful application to new (but related) tasks. For ML, *tasks success* can be quantified using appropriate evaluation metrics, where lower error indicates greater success. Establishing reliability of an ML application is typically done by measuring performance on out of sample data or *test data*, which was neither used to optimize the estimator parameters (training data) nor to guide hyperparameter tuning (validation data). In other words, if an algorithm can take new data, not used for training or tuning, and still "correctly" estimate the target, then, we argue, the reliability component of the key factor is satisfied (to the corresponding degree of correctness).

2019; Kingma & Welling, 2013). Note, while the hidden layer representations (typically) are not directly designed to capture human intelligible concepts, (remarkably) such techniques often detect hidden layer representations corresponding to such human intelligible concepts as illustrated in Fig. 3, such as chair-size, azimuth, etc. (see also (Iten et al., 2020) for a related physics example).

To justify this claim we must elaborate on what is meant by "correctly" estimating, and what is meant by "new" data. For "correctness," if the difference in the selected evaluation metric on test vs. train data, called *generalization error*, is low in addition to the test and train errors themselves being low, then we have good indication that the success achieved in training was reliable and not due to mere overfitting of the training data. For "newness," a new data set (viz., a test data set) is generated from the same sampling process as the training and validation data and as such is described by the same distribution. There is a subtlety here: just because a new data point is sampled from the same process, the model need not be successful on an arbitrary raw data input. In image recognition, for instance, data is represented typically as something like a bitmap. For example, if we are working with 64×64 grayscale images, then the space of all possible data points is $[0, 1]^{4096}$. Most points in this space are just white noise for which presumably no target label exists (aside from a generic null label). The data points generated by the sampling procedure approximates a low rank submanifold in this space. This low rank submanifold of "potentially" sampled data points with target labels characterizes the set of images an ML trained algorithm (or human) should be expected to classify.[11] Sampling test data from the same process (with the same distribution) captures the expectation that reliability means success on such additional examples. To the extent that a model can achieve high performance on a test set (with low generalization error) defined in this way, we can say it exhibits reliable task success.

In contrast to reliability, we have used the term "robustness" to refer to new (related) task applications (as noted in Sect. 2). Reuse of pretrained models in the context of DL applications has become commonplace in contemporary practices such as text embedding or reusing pretrained image classification or generative networks. Here, a model is trained for one task (e.g., predicting a masked word in some text) and then reused for a different task (e.g., translation, part of speech tagging, named entity recognition) using different data. This can happen with minimal or no additional training (as in zero-shot or few-shot transfer learning (Brown et al., 2020; Vinyals et al., 2016)) or, more generically, where the pretrained model is an upstream process for representing raw data that is then fed as input to additional layers trained specifically for a new task or tasks. Pretraining word representations for natural language processing (NLP) applications as in (Mikolov et al., 2013), or, in recent years, for sequences of text for natural language understanding (NLU) applications as in (Devlin et al., 2019), is the overwhelming standard for contemporary research. Similarly, it is common practice for image tasks to transfer the representations of already trained large models or self-supervised generative encoders (Higgins et al., 2017; Xiao & Wang, 2019; Chen et al., 2016; Kingma & Welling, 2013) for new applications where deeper or added layers receive further training. While limited further training, called *fine tuning*, is often used in such transfer learning of pretrained model applications, the pretraining

[11] Adversarial examples are typically characterized as data points that both have small (e.g. L_1 or L_2) deviations from labeled data points within this submanifold but are challenging for an algorithm to correctly label. Models vulnerable to adversarial attack represent a commensurate failure to meet our reliability or potentially our information relevance factors.

of a common representation core[12] often permits successful application to novel tasks with far less time and data required. The degree to which a common representational core reduces or even eliminates further fine tuning provides a measure of the robustness component of this key factor.

Recall, for the purposes of this work, in Sect. 2 we reserved the terms 'reliability' and 'robustness' as shorthand for extending to additional instances of the same task and to applications in new (related) tasks respectively. In this section, for ML applications we have proposed one way to distinguish the two can be based on the novelty of the distributions describing the respective data sample generation sources. Specifically, we propose our usage of reliability in ML can apply to performance on new data generated from the same sampling process and so ostensibly corresponding to the same underlying distribution. In contrast, we propose robustness in ML can apply not simply to using new data, but to using data corresponding to a different underlying distribution (e.g., because it is sampled through a substantively different process, or it has new variables including labels).[13] Hence, according to our proposal, in the case of ML, novelty of an underlying distribution is (at least) one mode sufficient for establishing novel applications consistent with the general account of Sect. 2.

To summarize, reliability in task performance corresponds to reduction of error on appropriately new out-of-sample test examples and robustness corresponds to the degree to which a representation core learned in hidden layers can then be reused for novel tasks reducing or eliminating the need for further training.

4.2 Relevance and the information bottleneck analysis

Our goal in this subsection is to make a connection between information relevance (the second key factor), which is the ability to represent the *relevant* and *only the relevant* information useful to tasks, and the relevance of information in hidden layer representations in DL. Specifically, we explain how the technical concepts (defined below) of *sufficiency* and *minimality* align respectively with the two directions of information relevance: Sufficiency quantifies the degree to which *relevant information* is preserved. Minimality can be framed as the degree to which *only the relevant information* is preserved, i.e., the degree to which irrelevant information is removed. The analysis that follows consists of formulating how these two (potentially conflicting) objectives can be simultaneously optimized (through what Tishby and Zaslavsky (2015) frame

[12] To be precise, such a *representation core* specifically refers to any vector (tensor) space representations resulting from hidden layer outputs of a pretrained component transferred as part of a new (larger) neural network. Such pretrained components often establish a base or "core" representation space from which downstream layers of the neural network develop deeper layer representations, hence the suggestive terminology.

[13] Of course, the underlying distributions will not be entirely unrelated (i.e., share no mutual information). After all, data used to pretrain one language model (e.g., for masked language modeling (MLM)) in a particular language presumably shares a good deal of mutual information with data used to fine tune it for a second use case (e.g., a semantic similarity task (SST)). However, the distribution of the MLM data will be substantively different from that of a labeled SST data, if only (among other potential differences) because there are relationships in the joint distribution of the latter, which contains the labeled variable(s), that did not exist in the former data. It is a virtue of our novel distribution proposal that in "different labels" examples it follows that we have a new task type (e.g., SST), consistent with the discussion of Sect. 2.

as a form of "compression"). This subsection will be devoted to understanding how these technical concepts can be used to evaluate (and potentially quantify) information relevance (the second key factor) in an ML context. The following subsection then leverages this formulation's account of "compression" to further evaluate well-structured representation (the third key factor) in an ML context.

For a neural network with at least one hidden layer, each hidden vector $\{h_i^{(l)}\}_i$ is some mathematical transformation of the raw data $\{x_i\}_i$. In a trivial sense, we can think of any mathematical transformation as a kind of representation.[14] However, clearly not any representation resulting from a mathematical transformation succeeds in capturing the relevant information. For information relevance, the $\{h_i^{(l)}\}_i$ should preserve the "relevant" and (for the most part)[15] "only the relevant" information in the data $\{x_i\}_i$ for the task. To formalize these concepts for an analysis of their presence in neural networks, following (Tishby & Zaslavsky, 2015), consider x and y as random variables with joint distribution $p(x, y)$ from which the training and test data $\{(x_i, y_i)\}_i$ are sampled. For a given layer l, the hidden vector representations $\{h_i^{(l)}\}_i$ of the $\{x_i\}_i$ can be viewed as samples from the random variable $h^{(l)}$. We can formalize $h^{(l)}$ as representing the "relevant and only the relevant information" in terms of the concepts of sufficiency and minimality:

(Sufficiency) Since the information $h^{(l)}$ shares with y comes by way of the information it shares with x, $y \to x \to h^{(l)}$ forms a Markov chain, and we have:[16]

$$\delta^l := I(x, y) - I(h^{(l)}, y) = H(y \mid h^{(l)}) - H(y \mid x) \geq 0$$

where $I(\cdot, \cdot)$ is mutual information and $H(\cdot \mid \cdot)$ is conditional entropy. When $\delta^l = 0$, the representation $h^{(l)}$ is a *sufficient* statistic of y (for the information in x). While $\delta^l = 0$ is not typical, the closer $h^{(l)}$ is to preserving information that x shares with y

[14] Mathematical transformations, mapping elements of one space to another may trivially count as such representation if only through mere ostension via said mapping. The purpose of this subsection is to investigate if such mappings go beyond such mere ostension mappings from an information perspective, preserving relevant and only the relevant information for the ML task. Of course, the term "representation" has a vast scope of application contexts widely discussed in philosophical literature, ranging from mere stipulated ostension or denotation (Callender & Cohen, 2006) to more information content rich representations used for inference in, e.g., scientific modeling, as in similarity accounts (e.g., (Weisberg, 2012)), structuralist accounts (e.g, (Da Costa & French, 2003)), and inferential accounts (e.g., (Khalifa et al., 2022)). For our present purposes, we restrict our focus to the context of how hidden layer vector (or tensor) space objects $h^{(l)}$ represent other vector (or tensor) space objects and not the (separate) question addressed in (Tamir & Shech, 2023) concerning how data may succeed or fail to represent the phenomena from which it was sampled.

[15] In the case of representations serving as pretrains for other downstream tasks, the restriction on "only the relevant" might be loosened to something like "mostly." Pretrain representations can work when they are trained on a broad enough task that the representations successful for the pretrain task overlap with further applications. There is plausibly something of a tradeoff here between the scope of minimality *for a specific task* and the generality of a representational core (robustness) that happens with pretraining.

[16] This follows from the data processing inequality.

(i.e., the smaller δ^I) the greater the *sufficiency* of $h^{(l)}$. The sufficiency of a representation $h^{(l)}$ formalizes the degree to which it represents "the relevant" information.

(Minimality) Simultaneously minimizing $I(x, h^{(l)})$ corresponds to preserving "only the relevant" information. Specifically, $h^{(l)}$ is a *minimal* representation when $I(x, h^{(l)})$ is the smallest among all sufficient representations. Couching in terms of degree, the smaller the quantity $I(x, h^{(l)})$ is (i.e., the closer to the minimal representation) the better its *minimality*.

Tishby and Zaslavsky (2015) apply a generalization of finding a minimally sufficient representation $h^{(l)}$ in DL with the Information Bottleneck Lagrangian (IBL):

$$\mathcal{L}_{IB} = H(y \mid h^{(l)}) + \beta I(x, h^{(l)})$$

where β is a Lagrange multiplier. The IBL generalizes the simultaneous optimization of finding a minimally sufficient statistic $h^{(l)}$, where β balances the trade off between sufficiency (minimizing $H(y \mid h^{(l)}) \geq H(y \mid x)$) and minimality (minimizing $I(x, h^{(l)}) \geq 0$). In this analysis they characterize the representation learning process within a neural network as removing information in x (minimizing $I(x, h^{(l)})$) "unneeded" for sufficiency (minimizing $H(y \mid h^{(l)})$) as a process of "compression." Is such compression merely a technical artifact,[17] or does simultaneous minimization of both terms in the IBL constitute the kind of compressions described particularly in Wilkenfeld's (UC)? From their theoretical constructions, the sufficiency and minimality terms in the IBL seem to track the dual requirements that the representation be "useful and relevant to context" (low $H(y \mid h^{(l)})$) but also "minimal" (low $I(x, h^{(l)})$). Do DL representations with good sufficiency and minimality exhibit the properties expected in philosophical accounts of understanding? Do they (A) ignore or abstract away unimportant information and (B) organize the important information for (reliable and robust) task success? We address (A) presently by exploring the DL phenomenon of *nuisance insensitivity*. We return to (B) in the following subsection where we investigate the concept of *disentanglement* and its relation to the third key factor.[18]

Nuisance factors are variations in the data carrying no mutual information with the target task. This can include noise factors, but also systematic influences due to data sourcing methods. For instance, differences resulting from perspective including occlusion, translation, or rotation of a pictured object exemplify nuisance factors for object recognition. Nuisance factors are incidental to particular examples, so over-

[17] Under typical circumstances if the $h^{(l)}$ is a deterministic function of x then $I(x, h^{(l)})$ is infinite or constant. This degeneracy can be avoided however either by adding noise as with the common training process practice of dropout—see (Goldfeld et al., 2019; Achille & Soatto, 2018)—by introducing noise only for evaluation, or by bucketing estimation as in (Saxe et al., 2019; Shwartz-Ziv and Tishby, 2017). Bucketing can introduce obfuscating artifacts in the estimation (Goldfeld et al., 2019; Amjad & Geiger, 2020), and as such in the results cited below we rely only on alternative estimations of $I(x, h^{(l)})$.

[18] Note, the dichotomy of these two questions does not align with the minimality and sufficiency dichotomy. (A) clearly is focused on the question of removing irrelevant information (minimality). In contrast, (B) is focused not only on preserving "the right information" (sufficiency) but more specifically with how that ("right") information is organized so as to be effective (disentanglement). Sufficiency should not be conflated with disentanglement. While both are required for task success, the former (strictly speaking) corresponds to the presence of the (relevant) information, while the latter (as we discuss in Sect. 4.3) corresponds to how effectively it is then organized for this task (if present).

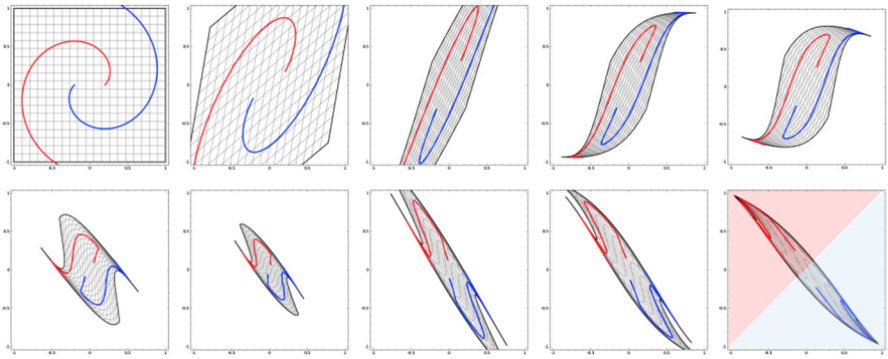

Fig. 2 Spiral label data in an original feature space undergoing a sequence of linear and nonlinear transformations into a disentangled representation space

fitting to incidental correlation with the target in training data causes generalization error. Achille and Soatto (2018a) formalize a variable n that affects x as a *nuisance* to the target y whenever $I(n, y) = 0$ even though $I(x, n) > 0$. An $h^{(l)}$ is *insensitive* to nuisance n whenever $I(h^{(l)}, n)$ is small. Achille and Soatto (2018) show experimentally that when minimality gets increased IBL weight, surrogate measures of mutual information (resulting from artificially introducing occluding nuisance factors) reduce as test performance approaches optimality.[19] Achille and Soatto (2018a) later prove more generally that as $I(x, h^{(l)})$ decreases, insensitivity to nuisances improves, and (sufficient) representations become more insensitive the deeper they are in the neural network. *This means that if a neural network is trained with enough data to ensure sufficiency of some deeper layer representation, then insensitivity to (irrelevant) nuisances is induced*. Results like these support the empirical hypothesis that the kind of compression represented by reducing the minimality term ($I(x, h^{(l)})$) is indeed more than a mere technical artifact as questioned above. Nuisance insensitivity induced by such compression directly corresponds to the intuitive expectations of the information relevance key factor, ignoring unimportant information while utilizing what's important. We hence conclude that such nuisance insensitivity is both empirically detectable and a germane manifestation of the information relevance key factor of understanding.

4.3 Well-structured representation as disentanglement

Bengio (2009, p. 6) first introduced the notion of *disentanglement*, explaining that hidden layers "can be seen as learning to transform one representation (the output of the previous stage) into another, at each step maybe disentangling better the factors of variations underlying the data." In their influential review of representation learning (Bengio, 2013, p. 1798) elaborate that an "AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify

[19] See also (Goldfeld et al., 2019; Saxe et al., 2019). Whie, Saxe et al. (2019) highlight compression in the original work (Shwartz-Ziv and Tishby, 2017) as an artifact of saturating the sigmoidal activation functions they used, they provide empirical evidence that when nuisance data is added to x in the form of manufactured irrelevant features, a compression phase specifically for nuisance features is in fact observed.

and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data." This concept is illustrated in Fig. 2 depicting the sequence of linear and non-linear neural network transformations.[20] The y labels are "tangled up" in a spiral pattern in the original feature space but in deeper representations these points are clearly mapped to separable "red" and "blue" regions respectively. Intuitively, while raw representations (like the tangled up spiral) may be sensitive to certain changes (red and blue points are close in the raw representation), by mapping the raw representation to a disentangled representation layer, such instability under perturbation is an artifact of the raw representation: small changes to blue/red data points in the final representation keep them "safely" in their respective classification neighborhoods.

The proximity of learned representations makes sense from a practical perspective. If a neural network can map diverse input data such as image bitmaps (i.e., the "milieu of low level sensory data") to more localized regions of a learned representation space, it improves the chances of a final layer estimating correctly. Early text embedding techniques such as (Mikolov et al., 2013), or more recent text-in-context embedding techniques like (Devlin et al., 2019; Liu et al., 2019) provided examples of how "low-level" information about word occurrence in text sequences can be transformed into semantically significant representation as text vectorizations in hidden layers. In particular, such embeddings can be used to find synonymous terms (or term usages in context) through proximity in the embedding vector space. *Geometric clustering*, which focuses on this kind of proximity in learned representation spaces, is one measure of disentanglement. Goldfeld et al. (2019) study how labeled data clusters in learned hidden layer representations during training. In a series of experiments,[21] they observe that as the network trains, compression measured as a reduction of the quantity $I(x, h^{(l)})$ directly corresponds to geometric clustering. Such experiments show how disentanglement claims relate to induced compression and can be empirically evaluated in terms of this relationship.

Disentanglement is central to studies of self-supervised generative neural networks used to find low rank representations of raw data (Higgins et al., 2017; Xiao & Wang, 2019; Chen et al., 2016; Kingma & Welling, 2013). This family of techniques is often used for learning low rank latent hidden layer representations of images, and are frequently optimized for loss functions closely related to the IBL (Achille & Soatto, 2018a; Higgins et al., 2017). Researchers have established a strong relationship connecting the β parameter in the IBL, which controls minimality, with both phenomenological observations and quantitative measures of disentanglement.

The images in Fig. 3 represents a typical phenomenological exploration of how dimensions in representation spaces exhibit disentanglement of latent factors. Specifically, it depicts the emergence of how high-level factors like (a) azimuth (b) chair size (c) leg style and (d) backrest vary as individual latent representation dimension values are traversed holding the other dimensions fixed. The phenomenological patterns depicted suggest that the latent representation dimensions learned by such models

[20] Figure 2 adapted from (Olah, 2014).

[21] Notably, Goldfeld et al. (2019) run a three dimensional version of the spiral disentanglement experiment (B.2) suggestively similar to the initial spirals of Fig. 2.

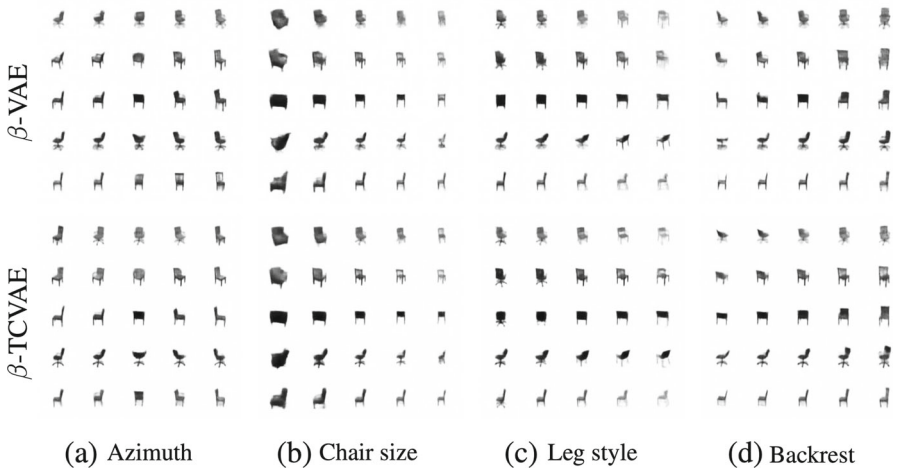

Fig. 3 Latent traversal of dimensions in representation spaces disentangled using β-VAE (top rows) (Higgins et al., 2017) and β-TCVAE (bottom rows) (Chen et al., 2018). Respective traversals along dimensions in the representation space exhibit changes in the observable factors of **a** Azimuth, **b** chair size, **c** leg style and **d** backrest

successfully factor out (disentangle) these high level properties and can be manipulated independently in the representation space. The models learn to treat each factor separately as a different dimension in the representation space. These patterns have also inspired a variety of quantitative measures of disentanglement (Achille & Soatto, 2018a; Higgins et al., 2017; Xiao & Wang, 2019; Kim & Mnih, 2018; Chen et al., 2018) that evaluate the degree to which dimensions in the representation layer vary independently. While specific measures vary, as a general pattern, increased compression/minimality induces corresponding increased disentanglement as measured by such metrics.[22]

Let us take a look now at how disentanglement relates to our third key factor that a representation be well-structured. In Sect. 2, we summarized this factor as the ability to structure how (relevant) information is minimally represented to make it efficacious in a task or tasks (even or especially under perturbations). If a hidden layer representation exhibits geometric clustering for a given task, then data points with particular labels tend to map to specific regions of the hidden layer's space. This means that not only do similarly labeled data tend to cluster in similar regions, but also that, as observed with the spiral example, the algorithm will be stable under perturbations in the representation space. This sort of stability is reminiscent of Wilkenfeld's examples of proof memorization and our example of the two drivers in Sect. 2. In both examples,

[22] For instance, Achille and Soatto (2018), measure the total correlation defined as the KL-divergence between a representation vector treated as a random variable and the factorization of its component dimensions treated as independent variables. In specific experiments, they demonstrate that as the compression/minimization penalty increases (with β), so does disentanglement as measured by the total correlation
Footnote 22 continued
of the hidden layer representations. Kim and Mnih (2018) offer an example of improvements to such total correlation based measures of disentanglement addressing tactical vulnerabilities that exist with pure total correlation measures.

the argument is that a "brute memorization" indicates less strong understanding when it comes to task completion because small perturbations to the execution of brute force representations render them less effective or ineffective. In contrast, a well organized representation of the important information (viz., what Wilkenfeld refers to as "higher order information," or, in the driver example, the topological structure of roadways) enables an understander to better adapt under similar changes. Literal perturbations of how an example is transformed into a hidden layer representation space using dropout techniques show similar stability of representations disentangled enough to ensure geometric clustering.[23] Experiments like (Achille & Soatto, 2018a; Higgins et al., 2017) show that this organization leverages more than proximity. Putting sufficient weight on the minimality term of the IBL induces representations that organize the relevant information in the data. Though disentanglement is still an active area of research, multiple techniques for evaluating different aspects of how neural networks learn to organize relevant information are clearly available. As with our other two key factors, the extent to which researchers can detect such aspects of how information transformed to hidden layer representations is well-structured allows us to evaluate the extent to which the third key factor is evident in these algorithms.

5 Deep learning and machine understanding

We noted in Sect. 1 that advances in ML and DL have enabled algorithms to perform on specific tasks in many cases at levels competitive with humans, and such improvements have been accompanied by the increased use of terms like "understanding" in artificial contexts. This led us to ask if the philosophy of understanding literature can help identify the conceptual criteria for evaluating potential machine understanding, and whether trends and patterns in how DL algorithms process data from a representation and information compression standpoint could partially or fully satisfy such criteria. We answer the former question in Sect. 2 and the latter question in Sects. 3 and 4. Specifically, in Sect. 4 we reviewed the three key factors of understanding from Sect. 2, identifying a basis for evaluating the presence of each factor either in direct task performance of DL models or in analyzing representations learned in neural net hidden layers. We argued for the following: (1) reliability and robustness can be respectively evaluated in terms of generalization error and potential reduction of further training requirements, particularly in transfer learning. (2) Information relevance can be evaluated through the sufficiency and minimality of a hidden layer representation (respectively corresponding to representing the relevant and only the relevant information) with the information bottleneck analysis, and we saw (concretely) that nuisance insensitivity to irrelevant information is induced directly by minimality. (3) Well-structured representation can be understood in terms of the DL concept of disentanglement, and specific techniques for measuring aspects of disentanglement correspond to measuring how factored dimensions in the representation

[23] Cf. Information dropout (Achille & Soatto, 2018) which generalizes this method and can be used to induce increased disentanglement.

layer organize (relevant) information and provide stability under perturbations. We conclude by considering three objections.

The first objection observes that the success of DL trained algorithms ostensibly achieving human competitive performance is often limited to narrow tasks. As Marcus (2020) argues, while incredibly large language models similar to that of (Brown et al., 2020) show remarkable success in transfer learning on a diversity of tasks, such models still have notable challenges in other language tasks. Bender and Koller (2020) further argue that even if language models trained only through word context successfully generate appropriate text responses, the meaning is "ungrounded" by external reference and vulnerable to leveraging purely syntactic language patterns instead of detecting genuine semantic relations.[24] How can we attribute understanding to machines given such current limitations? In reply, first we argue (only) that a technical basis for evaluating the presence of the key factors exists and can be conducted especially through appropriate task evaluation and information theoretic analysis of hidden layer representations (as elaborated in Sect. 4). We do not claim that this means such evaluation must conclude that any algorithm possesses a specific degree of understanding or meets the key factors associated with said understanding. Machines with understanding of the kind sought by Marcus (2020) and Bender and Koller (2020) may well consist of multiple coordinated algorithms and broader goals and reward structures perhaps more in line with advancing reinforcement learning research focused on developing internal (hidden) representations of the observed environment ("world models"), its dynamics, and self-critical action planning (Ha & Schmidhuber, 2018; Hafner et al., 2019b, a; Okada & Taniguchi, 2021). Coordination of multiple DL components for (world) model based reinforcement learning, or grounding tasks like image captioning or text to image generation (Ramesh et al., 2022; Saharia et al., 2022), leave room for the future possibility of machine understanding to a degree and with the breadth of abilities expected of such critics. It is our goal in Sect. 4 to provide a framework for evaluating such claims by more directly engaging with the abilities and representations of any such ML/DL models or machines that positively manifest potential understanding (to some degree).

As a second objection, one may be concerned that we have identified the three key factors as (merely) *indicative* of understanding instead of insisting that they are constitutive of understanding. In response, we frame the contributions of this work as identifying multiple concepts in existing philosophical literature on (human) understanding as independent touchpoints for what might arguably deserve focus when considering understanding in machines. Having identified three such objects of consideration, we show that the presence of such factors can be evaluated and even quantified to various degrees in machines, which we argue may serve as a useful critical tool. We

[24] McCoy et al. (2019), notably present evidence that transformer architectures leverage syntactic heuristics to achieve human competitive performance rather than detecting genuine semantic implication. See also (Liu et al., 2020; Nie et al., 2019) which similarly highlight foils in the most popular benchmarks, and then introduce tactics for (modestly) addressing these serious vulnerabilities. Despite such notable information

Footnote 24 continued

leakage failures, recent work training models using unstructured text simultaneously with images (Radford et al., 2021) in order to develop text to image generation models (Ramesh et al., 2022; Saharia et al., 2022) frame the more principled arguments of (Bender & Koller, 2020) against the possibility of grounding such language models as rather hollow.

claim that observations measuring the degree to which these factors (or other potential factors) are present in machines may constitute (partial) evidence of understanding, even if they do not play the deductive role of analytic conceptual conditions. Evidence comes in degrees, and we do not make claims on how much evidence warrants a decisive conclusion that a machine (or human) possesses understanding. However, we do claim that having tools for quantifying and comparing evidence constitutes a substantive step in considering the (relative) warrant(s) for said putative judgments. The arguments of Sect. 4 demonstrate how the key factors we have identified can concretely support this role.

Further, while our efforts here take no aim at settling the larger philosophical debate over the proper conceptual analysis of understanding (human or otherwise), our position is more closely aligned with positions claiming that understanding likewise comes in gradations. Not only does evidence (for any understanding) come in gradations, the degree and extent of said understanding also comes in gradations. Again, we view the contributions of Sect. 4 as an asset in comparative evaluation (as illustrated with our example below) rather than delineating some crisp conceptual border (presently) separating human understanders from machines. Indeed, we propose leveraging the methodology for evaluating the three factors to compare important aspects of understanding indicated in different machines, even when they may not qualify as having human understanding. We take this work to be the first step in a larger enterprise, which may ultimately provide more critical insight into discussions of machines that might (one day) merit full consideration as exhibiting a quality like understanding.

The last objection we consider hinges on the claim that the "mentality" of minded persons is a necessary condition for understanding. How can there be value to evaluating (partial) machine understanding in terms of the three key factors when a machine falls short of the mentality necessary for understanding? Again, we are not claiming that any current AI technique or trained algorithm successfully satisfies a mentality condition or counts as a minded agent, but we note three reasons why independent evaluation of the key factors may prove beneficial. First, as highlighted in Sect. 1, active researchers familiar with the field make use of understanding terminology, especially when discussing DL. Establishing a technical basis for the extent to which the key factors of understanding do (and do not) apply provides philosophical context for considering whether such claims merely succumb to human tendencies to anthropomorphise (Zlotowski et al., 2015) or if such usage has a technical and philosophical basis. Second, by grounding clear criteria for key factors of significance in philosophical accounts of understanding in (human) persons, we open the possibility of an account that is more sensitive to machines exhibiting *some aspects of understanding* while falling short of strong AI mentality or the kind of "deep" understanding sought in works like (Lake et al., 2017; Marcus, 2020). By highlighting the key factors and their potential for evaluation, we have (at least three) concrete methods of detecting these aspects.

As a toy example of such nuanced evaluation of understanding (or lack thereof) in artificial cases, imagine we need a gate to open for cats and only cats. To operate this gate we first have a human operator, second a passive infrared sensor (PIR) calibrated to detect heat levels of cats (but not, say, mice), and third a DL trained cat recognition model, each operating the gate by opening only for cats. The human, we stipulate,

understands this task sufficiently. What about the other two? While neither the PIR nor the neural network have the kind of understanding the human has, there is a salient question of what (if any) similarities can be found. Is the neural network "closer" to understanding how to recognize cats than the PIR? The framework laid out in Sect. 4 provides guidance for answering this question in a more nuanced manner than the coarse grained distinction of human vs. machine. For the PIR there is binary input data, which maps to a binary control (`open if correct_heat == 1, else do nothing`). Even if we say that the PIR's mapping is a representation, there isn't much basis for suggesting this representation extracts or organizes the right (and only the right) information through more than a heuristic control. In contrast, analysis of the neural network hidden layer representations could reveal which raw image features it ignores through nuisance insensitivity and which it transforms and disentangles in the deeper hidden layers. The way the neural network extracts and organizes the relevant features plays out in counterfactual reliability and robustness: For example, if a cat-sized dog were to attempt to go through the gate, the neural network might correctly respond by keeping the gate closed where the PIR would not. For the neural network this is not just observable in the performance success, but in the internal representations of how it disentangles relevant raw features and ignores other nuisances to distinguish cats from dogs and other non-cats. Such representation provides a basis for characterizing how the neural network is "closer" to exhibiting the human's understanding than the PIR.

For our third and final response, while a full account of what qualifies as mentality (particularly in non-human persons) is out of scope for this work, the speculation that possessing the potential to understand plays a role is not unreasonable. Insisting that mentality is required for understanding runs the risk of begging the question when it comes to artificial cases. Given how challenging questions of artificial personhood are, having an aspect sensitive (if partial) account of machine understanding independent of a mentality presupposition could inform such questions. By introducing this framework through engagement with the techniques, algorithms, and theory underlying DL successes while also leveraging philosophical analyses of understanding in human persons, we better position future exploration of possible artificial mentality without rejecting it out of hand.

Acknowledgements Earlier versions of this paper were presented at the 2nd Scientific Understanding and Representation (SURe) Workshop at Emory University, Atlanta, GA, and at the 2nd Machine Wisdom Workshop at University of Pittsburgh, Pittsburgh, PA. We thank the audiences for helpful suggestions and discussion. Special thanks to Balázs Gyenis for insightful conversations and many constructive comments.

Declarations

Conflict of interest The author asserts that there are no conflicts of interest.

References

Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research, 19*(1), 1947–1980.

Achille, A., & Soatto, S. (2018). Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 2897–2905.

Amjad, R. A., & Geiger, B. C. (2020). Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(9), 2225–2239.

Baumberger, C. (2014). Types of understanding: Their nature and their relation to knowledge. *Conceptus, 40*(98), 67–88.

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Bengio, Y., Mesnil, G., Dauphin, Y., & Rifai, S. (2013). Better mixing via deep representations. In *International Conference on Machine Learning* (pp. 552–560).

Bengio, Y., et al. (2009). Learning deep architectures for ai. *Foundations and Trends in Machine Learning, 2*(1), 1–127.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Buckner, C. (2020). Adversarial examples and the deeper riddle of induction: The need for a theory of artifacts in deep learning. arXiv preprint arXiv:2003.11917.

Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria. Revista de teorí a, historia y fundamentos de la ciencia, 21*(1), 67–85.

Chen, R. T. Q., Li, X., Grosse, R. B., & Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc.

Da Costa, N. C., & French, S. (2003). *Science and partial truth: A unitary approach to models and scientific reasoning*. Oxford University Press on Demand.

De Regt, H. W. (2009). Understanding and scientific explanation. *Scientific understanding: Philosophical perspectives* (pp. 21–42). University of Pittsburgh Press.

De Regt, H. W. (2015). Scientific understanding: Truth or dare? *Synthese, 192*(12), 3781–3797.

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese, 144*(1), 137–170.

De Regt, H. W., & Gijsbers, V. (2016). How false theories can yield genuine understanding. In G. Sandu, I. Parvu, & I. Toader (Eds.), *Explaining understanding* (pp. 66–91). Routledge.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87.

Elgin, C. Z. (1993). Understanding: Art and science. *Synthese, 95*(1), 13–28.

Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., & Polyanskiy, Y. (2019). Estimating information flow in deep neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2299–2308). PMLR.

Grimm, S. R., Baumberger, C., & Ammon, S. (2017). *Explaining understanding*. New York: New Perspectives from Epistemology and Philosophy of Science.

Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems.* (Vol. 31). Curran Associates Inc.

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). *Dream to control: Learning behaviors by latent imagination*. arXiv preprint arXiv:1912.01603.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). Learning latent dynamics for planning from pixels. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2555–2565). PMLR.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., et al. (2018). *Achieving human parity on automatic Chinese to English news translation*. arXiv preprint arXiv:1803.05567.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing humanlevel performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1026–1034).

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Hills, A. (2016). Understanding why. *No?us, 50*(4), 661–688.

Iten, R., Metger, T., Wilming, H., Del Rio, L., & Renner, R. (2020). Discovering physical concepts with neural networks. *Physical Review Letters, 124*(1), 010508.

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.

Khalifa, K., Millson, J., & Risjord, M. (2022). Scientific representation: An inferentialistexpressivist manifesto. *Philosophical Topics, 50*(1), 263–292.

Kim, H., & Mnih, A. (2018). Disentangling by factorising. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2649–2658). PMLR.

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational bayes*. arXiv preprint arXiv:1312.6114.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*.

Lawler, I., Khalifa, K., & Shech, E. (Eds.). (2023). *Scientific understanding and representation modeling in the physical sciences*. NewYork and London: Routledge.

Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J. (2020). *Adversarial training for large neural language models*. arXiv preprint arXiv:2004.08994.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv preprint arXiv:1907.11692.

Marcus, G. (2020). *The next decade in ai: Four steps towards robust artificial intelligence*. arXiv preprint arXiv:2002.06177.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448). Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Nangia, N., & Bowman, S. R. (2019). *Human vs. muppet: A conservative estimate of human performance on the glue benchmark*. arXiv preprint arXiv:1905.10425.

Newman, M. P. (2017). Theoretical understanding in science. *The British Journal for the Philosophy of Science, 68*(2), 571–595.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). *Adversarial nli: A new benchmark for natural language understanding*. arXiv preprint arXiv:1910.14599.

Okada, M., & Taniguchi, T. (2021). Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4209–4215).

Olah, C. (2014). *Neural networks, manifolds, and topology*. http://colah.github.io/posts/2014-03-NN-Manifolds-Topology .

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don?t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 2: Short papers) (pp. 784–789). Association for Computational Linguistics.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical textconditional image generation with clip latents*. arXiv preprint arXiv:2204.06125.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., et al. (2022). *Photorealistic text-to-image diffusion models with deep language understanding*. arXiv preprint arXiv:2205.11487.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., & Cox, D. D. (2019). On the information bottleneck theory of deep learning*. *Journal of Statistical Mechanics: Theory and Experiment, 2019*(12), 124020.

Schupbach, J. N. (2018). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science, 69*, 275–300.

Shech, E. (2022). Scientific understanding in the Aharonov-Bohm effect. *Theoria, 88*(5), 943–971.

Shech, E., & Tamir, M. (2023). Expecting too much from our machine learning models: A reply to sullivan. In I. Lawler, K. Khalifa, & E. Shech (Eds.), *Scientific understanding and representation: Modeling in the physical sciences.* (pp. 346–350). New York and London: Routledge.

Shwartz-Ziv, R., & Tishby, N. (2017). *Opening the black box of deep neural networks via information*. arXiv preprint arXiv:1703.00810.

Stegenga, J., & Menon, T. (2017). Robustness and independent evidence. *Philosophy of Science, 84*(3), 414–435.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A, 44*(3), 510–515.

Stuart, M. T. (2018). How thought experiments increase understanding. In M. T. Stuart, Y. J. H. Fehige, & J. R. Brown (Eds.), *Routledge companion to thought experiments* (pp. 526–544). Routledge.

Tamir, M., & Shech, E. (2023). Understanding from deep learning models in context. In I. Lawler, K. Khalifa, & E. Shech (Eds.), *Scientific understanding and representation: Modeling in the physical sciences.* Routledge.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)* (pp. 1–5).

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems.* (Vol. 29). Curran Associates Inc.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., et al. Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems, 32*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353–355). Association for Computational Linguistics.

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

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese, 190*(6), 997–1016.

Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies, 176*(10), 2807–2831.

Xiao, Y., & Wang, W. Y. (2019). *Disentangled representation learning with wasserstein total correlation*. arXiv preprint arXiv:1912.12818.

Zlotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human-robot interaction. *International Journal of Social Robotics, 7*(3), 347–360.