Assessing Deep Learning: A Work Program for the Humanities in the Age of Artificial Intelligence

Jan Segessenmann[1*], Thilo Stadelmann[2,3], Andrew Davison[5] and Oliver Dürr[1,4]

[1*]Center for Faith and Society, University of Fribourg, Fribourg, Switzerland.
[2]Centre for Artificial Intelligence, Zurich University of Applied Sciences, Winterthur, Switzerland.
[3]European Centre for Living Technology, Venice, Italy.
[4]Institute of Hermeneutics and Philosophy of Religion, University of Zurich, Zürich, Switzerland.
[5]Faculty of Divinity, University of Cambridge, Cambridge, United Kingdom.

*Corresponding author(s). E-mail(s): jan.segessenmann@unifr.ch;
Contributing authors: stdm@zhaw.ch; apd31@cam.ac.uk;
oliver.duerr@unifr.ch;

Abstract

Following the success of deep learning (DL) in research, we are now witnessing the fast and widespread adoption of artificial intelligence (AI) in daily life, influencing the way we act, think, and organize our lives. However, much still remains a mystery when it comes to how these systems achieve such high performance and why they reach the outputs they do. This presents us with an unusual combination: of technical mastery on the one hand, and a striking degree of mystery on the other. This conjunction is not only fascinating, but it also poses considerable risks, which urgently require our attention. Awareness of the need to analyze ethical implications, such as fairness, equality, and sustainability, is growing. However, other dimensions of inquiry receive less attention, including the subtle but pervasive ways in which our dealings with AI shape our

way of living and thinking, transforming our culture and human self-understanding. If we want to deploy AI positively in the long term, a broader and more holistic assessment of the technology is vital, involving not only scientific and technical perspectives but also those from the humanities. To this end, we present outlines of a *work program* for the humanities that aim to contribute to assessing and guiding the potential, opportunities, and risks of further developing and deploying DL systems.

How to read this paper: It is structured in four modular parts: a general introduction (section 1), an introduction to the workings of DL for uninitiated non-technical readers (section 2), a more mathematical introduction to DL (appendix A), and a main part, containing the outlines of a work program for the humanities (section 3). Readers familiar with mathematical notions might want to skip 2 and instead read A. Readers familiar with DL in general might want to ignore 2 and A altogether and instead directly read 3 after 1.

Keywords: Deep Learning, Anthropology, Humanities, Artificial Intelligence, Ethics, Philosophy

1 Introduction

With the introduction of deep learning (DL) in around 2006 (Hinton et al. 2006; Bengio et al. 2006; Ranzato et al. 2006), the field of artificial intelligence (AI) entered into what has proven to be by far its most impressive period of advancement. The methods introduced with DL perform remarkably well in identifying complex patterns in large data sets in order to make predictions. Today, DL has found its way from research into our daily lives in a multitude of applications (Stadelmann et al. 2018; Yan et al. 2023), such as internet searches, translation apps, face recognition and augmentation on social media, speech interfaces, digital art generation, and chatbots. It can achieve enormous good, e.g., by preventing secondary cancer through improved medical imaging (Amirian et al. 2023). Other recent advances have further demonstrated the astonishing capacities of DL: generative AI models caught public attention by producing striking images from text prompts (e.g., 'DALL-E 2' and its open-access brother 'Stable Diffusion', as well as 'Midjourney', Ramesh et al. 2022; Rombach et al. 2022; Borji 2023), while generalist models (e.g., 'GATO', Reed et al. 2022), and the unprecedented utility of multimodal 'large language models' (LLMs), create the impression that we are getting closer to building so-called 'artificial general intelligence' (AGI): an engineered human-like or even superhuman intelligence (Bubeck et al. 2023; Agüera y Arcas 2022). Language models respond so persuasively to prompts and questions by human inquirers that some already think they exhibit some kind of sentience, and others believe that they will in the near future (Tiku 2022; Kaplan 2022; Schmidhuber 2022). Shortly after release, language models, such as 'ChatGPT' and

'GPT-4' have quickly become an integral part of the work and everyday life for many people. They have already passed bar examinations (e.g., the US Uniform Bar Examination for lawyers, see Katz et al. 2023).

Despite these successes, our theoretical insight into why DL performs so well is still shallow, and some of its success remains a mystery (Plebe and Grasso 2019; Hodas and Stinis 2018; Poggio et al. 2019; Berner et al. 2022; Zhang et al. 2017, 2021; Sejnowski 2020). As a consequence, engineering DL models involves a substantial amount of trial and error. From a theoretical perspective, in many ways, it is guesswork: while the end product often works rather seamlessly (although there are glitches, and these systems have the significant problem of not being able to recognize where they are wildly wrong), getting to a working system can involve substantial and creative experimentation on the part of the engineers. Some have even labeled the process as 'alchemy' or 'magic' (Hutson 2018; Ford 2018; Edwards and Edwarts 2018; Domingos 2012; Martini 2019; Flessner 2018; von der Malsburg et al. 2022). Moreover, the complexity of the problems solved with DL requires use of highly complex models that are incomprehensible to humans. This confluence of technical mastery and mystery in DL applications – of remarkable capacities that defy our capacity to understand them – has been observed to lead to what we might call an 'enchanted perception' of the technology in segments of the scientific community and the broader public (Campolo and Crawford 2020). Trans- and posthumanist accounts further radicalize expectations of what such technologies can achieve (or become) by describing future visions of 'uploaded' minds, an artificial "super intelligence" (Bostrom 2014; Tegmark 2018) or a "technological singularity" (see, e.g., Kurzweil 2005; Chalmers 2010; Eden et al. 2012; Barrat 2015). Not surprisingly, the astonishing performance of DL applications has given rise to anthropomorphisms and even a longing for – or fear of (Yudkowski 2023) – *superhuman* technology. The speed, scope, and intensity with which DL is influencing our societies press for a closer inspection and assessment involving a plurality of perspectives.

1.1 A Call to Assessment From a Humanities Perspective

As DL is increasingly implemented in critical fields such as healthcare, insurance, criminal justice, employment, and hiring, as well as financial markets, the problem that we often lack an explanation for how automated decisions are made in such situations is rendered more urgent. Recent legislation in the European Union (e.g., European Parliament and Council of the European Union 2016) states that individuals have the right to an 'explanation' if they are affected by an automated decision-making process. This is a critical step in the collective regulation of such technologies in light of their societal impact (Grunwald 2019b; Pflanzer et al. 2023; Salmi 2023). Next to such concerns, engineers also have technical reasons for wanting to understand input-output relations with greater clarity for the sake of increasing efficacy and robustness. This has led to a growing body of research on model interpretability in the emerging field of 'explainable artificial intelligence' (XAI) (Došilović et al.

2018; Adadi and Berrada 2018; Confalonieri et al. 2021; Joshi et al. 2021; Madsen et al. 2022; Notovich et al. 2023; Besold and Uckelman 2018) – which is sometimes also referred to as 'intelligible' (Weld and Bansal 2019; Caruana et al. 2020) or 'reviewable' AI (Cobbe et al. 2021). (On this, see also the vital contributions of the 'National Institute of Standards and Technology', www.nist.gov). Knowing *why* a system performs the way it does helps both to counter biases and to understand malfunctions, thus enabling us to improve the technology. However, bold claims about 'explaining' DL models often fail to do justice to the gap between the kind of explanation provided and the kind needed (Lipton 2018; Besold and Uckelman 2018). Overpromising what can be explained might prove to be a bad strategy, risking a loss of confidence and support for AI research if the technology does not deliver on the promises immediately. Not long ago, such a pattern – with disappointment over lack of trustworthiness, robustness, and comprehensibility in particular – led to talk about another 'AI winter' (Floridi 2020; Yasnitsky 2020), i.e., a period of low funding and thus low resources invested in AI research. While this has largely passed out of sight with the recent success of generative AI (Dwivedi et al. 2023), societal, political, and ecological concerns remain essential (Crawford 2021) and have, for example, led to bans on facial recognition, and a consequent slowing of research in that area (Wehrli et al. 2021). We are currently seeing initiatives for banning some generative AI applications (successful in some cases) worldwide and in many institutions.

The mystery that surrounds DL involves a yet more fundamental and more subtle danger, namely the premature confusion of human intelligence with purely computational and probabilistic processes and vice versa (Tallis 2004, 2020). The danger here is conceptual and methodological confusion, with sociopolitical consequences. As well as the risk of confused thinking, it also renders difficult the practical task for distinguishing between human beings, AIs, and robots, and thus conflicts with the democratic organization of our societies around the unique worth and dignity of human beings. If *we* are but machines, then why grant us special status amongst other machines (see, e.g., Gunkel 2018; Gordon and Pasvenskiene 2021; Munn and Weijers 2023; Novelli 2023)? Although the confusion of human beings with machines, and especially computers, has a long history (Boden 2008; Black 2014; Dürr 2021; Cave et al. 2020), notable recent achievements in DL have greatly contributed to the myth of the 'electronic person' – as seen, for instance, in work by the European Commission to address the status of sophisticated robots in terms of 'persons' (European Parliament 2017). Much of this cross-talk between registers – the computer and the human – is in danger of spawning jingle-jangle errors. Historically it stems from the fact that it was an analogy drawn from biological learning, in the form of neural networks, that inspired the original core principles underlying DL (McCulloch and Pitts 1943; Hebb 1949). Thus, the perceived comparability of human and computational forms of intelligence has propelled the anthropomorphization of DL language (Lipton and Steinhardt

2019; Kostopoulos 2021; The Royal Society 2018). Now running in the opposite direction, definitions of intelligence in purely technical terms (Legg and Hutter 2007; Chollet 2019) are often projected back onto humans, and perceived as the norm of intelligence *tout court* (Dennett 1991; Churchland and Sejnowski 1992; Chalmers 2011; Boden 1988). Evidence that 'intelligence' and other characteristics of the mind can indeed be modeled as computational processes seem to be increasing (von der Malsburg 2023), as DL models continue to deliver impressive results (notable, for instance, in the tendency to ascribe previously unknown 'creativity' to AI-generated 'art', Mazzone and Elgammal 2019).

If we want to harness the promise of DL and create a fruitful and humane future with these technologies, it is crucial and urgent that we think through the implications of DL not only from the technical perspective of science and engineering but also from a more encompassing humanities perspective. The reason for this is that our understanding of, and interactions with, technology is always inextricably linked with negotiating human self-understanding (Liggieri and Müller 2019). Much care and thought must be given to making sure that our technologies do not ultimately hollow out human values, forms of sense-making, and resources that motivate action from under us – Bernard Stiegler analyzes how digital technologies tend to undermine and even eliminate reflection and questioning of their development. Having this in mind, one of the key tasks for the humanities is to deliberately and carefully think about the conditions under which we can relate to technology in a more fruitful, livable, and humane way (Stiegler 2017). Thus, the future we will create with DL ultimately depends on our understanding of the technology, our view of human beings and the values which guide us in the assessment, design and deployment of technology.

1.2 How to Read This Paper

This paper sketches some important points of a work program for the humanities on how to assess and guide the potential, opportunities, and risks of further developing DL. In section 2, we provide a brief and up-to-date introduction of the known and unknown aspects of DL, written with uninitiated, non-technical readers in mind. This should provide them with realistic technical bearings without requiring any understanding of the mathematics involved. Sections 2.1 to 2.4 provide the basic theory of DL, its workings and inevitable limits, and potential errors, also with respect to recent transformer models behind systems like ChatGPT, while section 2.5 refers to some gaps in this theory. Readers familiar with basic mathematical notions who want to gain a deeper understanding of DL can skip this introduction and read the more in-depth introduction provided in appendix A. Readers already familiar with these concepts might want to skip both introductions altogether. In section 3, we identify some pressing issues that require attention from a humanities perspective. This includes differentiating between the 'human' and the 'technological' factors in ethical AI assessments (section 3.1), efforts to contextualize DL more broadly

(section 3.2), and exemplary resources, provided by the humanities in dealing with questions arising from DL deployment (section 3.3). We want to underline here that in pointing to certain weaknesses, inherent theoretical limits, and societal challenges associated with DL, we are not advocating a universally pessimistic stance toward digital technologies, AI, and DL in particular (Marcus 2018). We are rather suggesting that a realistic picture is necessary if we want to harvest the benefits, avoid the perils and prevent a disillusioning halt for AI research.

2 Deep Learning: An Introduction for the Uninitiated

DL is a form of machine learning (Mitchell 1997), which itself is a form of AI (Russell and Norvig 2021). Machine learning is usually categorized into supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, a model is trained for a specific task based on labeled data (i.e., it is given input examples and corresponding desired outputs). For instance, if a model is to predict whether an image of human skin contains a malignant melanoma, it is trained on many example images with known 'ground truth', i.e., labeled correctly as 'contains melanoma' or 'does not contain melanoma'. In unsupervised learning, patterns are determined in unlabeled data, with data clustered and grouped by the DL system without reliance on predefined labels. Strictly speaking, using parts of the data itself as labels (e.g., predicting the upper half of an image from its lower half or the next word in a given text), which is the predominant learning paradigm for large-scale models, would also fall under this definition, but is called 'self-supervised learning' instead because methodically it uses methods from supervised learning. In reinforcement learning, a DL 'agent' is trained to interact with its environment in order to achieve a certain goal based on a punishment-reward mechanism. Reinforcement learning is mostly used in robotics, games, or wherever interaction is required of the agent, so recently also in chatbots. In this paper, we only consider supervised learning, since this type of machine learning method is the most widely used and, by a large margin, responsible for the current successes of DL. A basic understanding of supervised DL carries far in assessing the potential of the other learning paradigms.

Outlook on this section:

In what follows, we outline the fundamental workings of DL by introducing artificial neural networks (ANNs), which comprise the core building block of any DL system (section 2.1). We then elaborate how they work by means of 'universal approximation' (section 2.2). Next, we analyze a set of inevitable errors that apply to every such system, based on their architecture and training algorithm (section 2.3). Section 2.4 aims, more specifically, to familiarize readers with the core concepts of current generative language models, like

ChatGPT. The last section (2.5) introduces some open questions in the theory of DL.

2.1 Artificial Neural Networks

ANNs are the fundamental building blocks of DL (for papers written at the origin of ANNs, see Rosenblatt 1958; Minsky and Papert 1969; Rumelhart et al. 1986; for a historical summary, see Schmidhuber 2015, for a contemporary introduction Prince 2023). To understand the basic principles of DL, one has to grasp how a basic ANN works: it consists of input units, hidden units, and output units, connected in a sequence of layers (see Figure 1) that between them encode a mathematical function. In more technical terms, we have a layered network of computationally simple units, which is trained to approximate a complex function that maps any desired input to any desired form of output (called the 'target space'). As an example, an ANN could classify images showing handwritten digits into the represented digits 0 to 9 (this is a classic problem in, for instance, the task of processing bank cheques automatically, see Lecun et al. 1998). In this case, the input would contain the gray scale pixel values of an image (each unit representing the shade of a single pixel), whereas the output would consist of ten values (units) representing the probabilities that the image shows the respective digits. As one can see, input and output layers are chosen to represent something meaningful (in this case, images and respective digits), while what is going on in the hidden layers remains hidden (as the name suggests) and is usually highly complex. When properly trained, the ANN, upon receiving an input, will send a much stronger output signal to the correct output channel than it will to the other incorrect outputs, thus indicating the correct digit or 'class').

Disassembled into its basic building blocks, all that an ANN does is a string of simple calculations: no mystery, no magic, no alchemy. In the next step, we want to assess these workings on a higher level of abstraction, where things begin to be more complex.

2.2 Universal Approximation

Through a process of sequentially altering the parameters of an ANN (which is to say, how the elements of one level feed into and trigger activations in the following level), it can potentially approximate *any* input-output relation. That may be, for instance, a very simple one, like, e.g., the relation between the distance traveled and money spent on gasoline, or more complex ones, like, e.g., the relation between images showing handwritten digits and the represented digits, or even – at the upper limit of what has thus far been attempted – between a protein sequence and the three-dimensional structure to which it folds (Jumper et al. 2021). How do such approximations work?

Approached in terms of the universal approximation theorem, an ANN encodes a function that can theoretically approximate any relation between two variables with arbitrary precision (Hornik et al. 1989; Cybenko 1989; Zhou

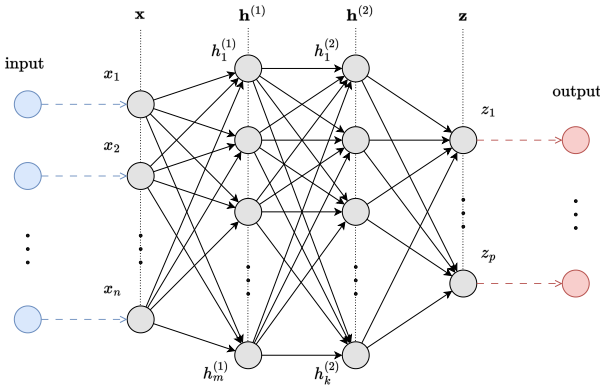


Fig. 1: Illustration of an ANN with two hidden layers $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$. The mapping from input to output is shown from left to right. Input and output consist of an array of numbers respectively, e.g., corresponding with the pixel-values of an image. Each hidden unit $h_j^{(i)}$ computes a weighted summation of values of preceding units (shown with solid arrows) and then passes it through a non-linear step function (this can be thought of as a threshold that the sum of inputs either passes or not, inspired by a biological neuron either firing or not). The weights are called the parameters of an ANN.

2020). The function encoded by an ANN is defined solely by the values of its parameters (i.e., the weights between units in any layer and those in the next layer). Before training, the input-output relation of an ANN is random, based on the randomness of the newly initialized parameters. After training, that once random constellation is trained to yield astonishing results. To understand this mapping from input to output as a single function, let us consider the example of the handwritten digits again. First, all pixel values of an input image are lined up (one row of the image after another to form one long sequence) such that they correspond to the form of the input layer in Figure 1. To better understand the workings of an ANN, every unit in the input layer (every pixel) can be thought of as an axis in multidimensional space. The value of a unit (pixel value) then defines a position on this axis. As the input consists of multiple units, the input image can be thought of as one point in this multi-dimensional data space (see Figure 2). Shifting this point along one axis corresponds with altering the value of one pixel. Picking a random spot in data space corresponds with an image consisting of random pixel values. The dimensionality of the data space is usually very high. If an image has, say, 28×28 pixels, then all pixels together define a single point in $28^2 = 784$-dimensional data space. The same is true for the units at any layer in an ANN, i.e., they all describe a single point in a multi-dimensional data space. Going from a layer with fewer units to a layer with more units thus corresponds to expanding the data space. Going from a layer with more units to a layer with

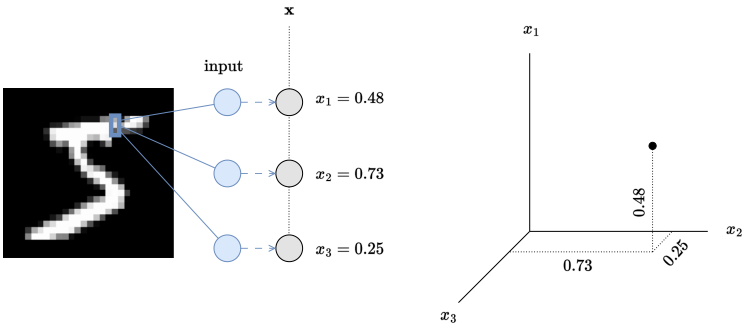


Fig. 2: Representation of an input image by a point in space. On the left is an image showing the handwritten digit '5'. Imagine that only three pixels are fed into the input layer of an ANN (the following layers are not shown), represented by their grey-scale value between 0−1. On the right, the input units are shown as axes and the unit values as positions on these axes. Thus, one point in three-dimensional space represents the three input pixels. Similarly, all the 784 pixels can be represented (although not visually illustrated) in 784-dimensional space.

fewer units corresponds to collapsing the data space. Based on the insight that meaningful inputs, such as images, can also be represented by points in space, it becomes easier to see that the relationship between input and output is a mathematical function. As every ANN encodes a function, it defines how the data space transforms, expands, and collapses from input to output, such that every input example transforms into an output example. This is also true for language models, as elaborated in section 2.4.

In practice, stacking many hidden layers (the number of layers between the input and output layers), i.e., increasing the depth of an ANN, has been shown to massively increase the capacity to approximate complex input-output relations (Bengio and LeCun 2007; Eldan and Shamir 2016; Raghu et al. 2017; Berner et al. 2022; Lin et al. 2017). This finding lies at the heart of DL: as the name suggests, the 'deep' in DL stands for the use of ANNs with many hidden layers. Since every hidden layer encodes a function itself, the function encoded by the deep ANN consists of a succession of functions (data space transformations), each cascading into the next. Although theory confirms that a single hidden layer would be sufficient to achieve the necessary transformation or linkage (if arbitrarily wide), the stacking of many hidden layers has shown to be far more efficient (Bengio and LeCun 2007; Bengio et al. 2013).

Although the benefit of deep models over shallow ones is still not really explained satisfactorily, a widely supported theory suggests that the benefits lie in their compositional structure: data representation gradually progresses through the layers from rudimentary to more complex aspects, sequentially converging on the salient features in the data (as observed initially by Lee

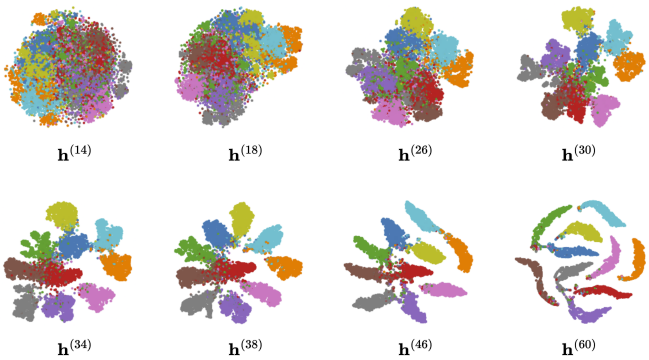


Fig. 3: Effect of the data-space transformations within an ANN that classifies images into ten classes, such as 'plane', 'car', 'bird', 'cat' etc. Every point in the plots corresponds to one image. Colours represent the respective class. Looking at the data representation at different hidden layers $\mathbf{h}^{(14)}$ to $\mathbf{h}^{(60)}$, one can see that the data is transformed in a manner that allows for easier separation of classes. The plots are taken from Hoyt and Owen (2021) (permission requested) and are obtained from real data. Note that to visualize an image in two dimensions, an algorithm was used that produces a low-dimensional representation such that distances between 2D points are reflective of the distances between the original (i.e., high-dimensional) images.

et al. (2009); Zeiler and Fergus (2014), with explanatory approaches proposed by, e.g., Mhaskar et al. (2017); Frankle and Carbin (2019); Hodas and Stinis (2018); Shwartz-Ziv and Tishby (2017) and summarized by Prince (2023)). Figure 3 shows the effect that this feature abstraction has on classification capabilities, while 4 visualizes the respective features themselves. A more low-level visualization is provided in the appendix (Figure 9).

Thus, the power of deep ANNs lies exactly in their capacity to approximate high-dimensional and complex (highly non-linear) functions by means of successive data-space transformations, combined with the property that these functions can be fitted to data, i.e., trained for specific tasks. Crucially, the true input-output relation underlying the task does not need to be known; it suffices to provide enough examples of input-output pairs. (Indeed, given the complexity of relations between elements in one hidden layer and the layer below it, such knowledge seems more or less impossible to obtain anyway). ANNs, thus, can extract complex patterns and provide human-accessible outputs that represent the underlying patterns in some meaningful way. For example, the complex patterns underlying images that show dogs or cats are transformed into two values only, representing the probabilities of the image to show a cat and a dog, respectively.

We have now seen that ANNs, on a more abstract level, can exhibit very complex functions, whose meanings, however, remain opaque to human insight

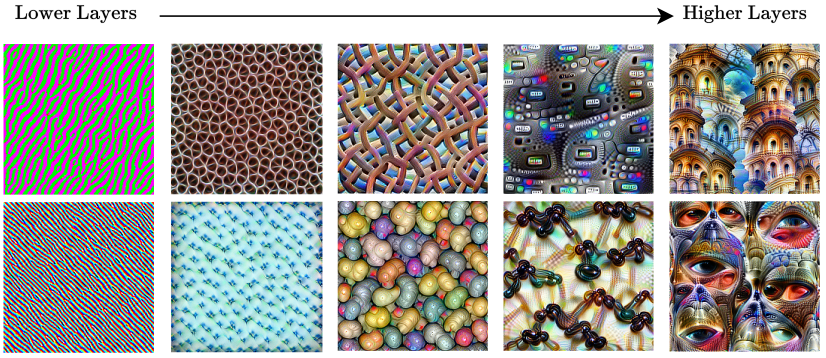


Fig. 4: Progression of data representation in the DL network 'GoogLeNet' (Szegedy et al. 2015). GoogLeNet is an instance of a convolutional neural network (CNN), which achieves state-of-the-art performance in image analysis tasks (for an accessible introduction, see Stadelmann et al. 2019b). The shown images were achieved by fixing the trained model parameters and instead optimizing the pixel values of the input image in a manner that maximizes the response of certain hidden channels (in CNNs, convolutional layers usually consist of channels, which consist of units). Thus, the obtained images show what the respective channels are detecting, i.e., how the respective channels represent input images. Going from lower to higher layers, we see that channels represent edges, then textures, then patterns, then parts, then objects, such as archways and eyes. We can see that the image representations in higher layers serve to simplify classification. Detecting cars based on raw pixel values or edges, for example, is hard, but detecting cars based on channels that represent objects, such as tires, lights, and streets (and ultimately cars themselves) is much easier. These images are taken from Olah et al. (2017) with permission.

due to their complexity. We can understand them on the lowest possible level, e.g., mathematically, but then miss the semantics of the operations that connect to meaningful concepts of human experience. Or, we can understand them on the highest possible level, e.g., mapping images of animals to the categories 'cats' and 'dogs'. But we cannot understand it in any way comparable to how something like this is achieved by a human. Thus, we can either achieve a superficial or a purely numerical understanding. But there is no explanation, on a meaningful intermediate level, of the 'reasoning' behind the level-by-level data space transformations, the performed abstractions, and the salient features identified. The complexity of the ANN itself, which enables it to automatically extract highly complex relationships from highly complex data and thus is its biggest strength, is also its weakness, as it causes the opacity as to how that works. To keep this introduction concise, we will skip the details of how an ANN can be fitted to data, i.e., how it can be trained to approximate a useful function and refer to appendix A.3 instead. Here, we rather want to give a general idea (see section 2.3 below) and then show how a fully trained

ANN inevitably deviates from its theoretical optimum of universal approximation. In other words, we present to the reader errors of trained DL models, which are unavoidable given their current architecture. When assessing DL from a humanities perspective, it is critical to keep these errors in mind, as they might bear on every DL model and application.

2.3 Inevitable Errors of Trained Deep Learning Models

Training an ANN requires the definition of a penalty (commonly referred to as the 'loss metric' or 'cost') that indicates the difference between the output produced by the model in response to a certain input and its corresponding target value (the known ground truth). If the output resembles the target, the penalty is small. The more the output deviates from the target, the higher the penalty. During training, the penalty is minimized by altering the underlying model parameters. This is an iterative process, which arrives at an increasingly better model through a mathematical method called 'gradient descent' (which ultimately is an implementation of the 'chain rule of differentiation' taught in school to 11th-graders). This is necessary because the optimal model parameters cannot be known, or directly calculated, in advance. Thus, every example 'shown' to the ANN (each input-output pair) provides some small degree of additional information about the direction in which each parameter should be nudged in order to slightly decrease the penalty, which is to say, to increase the performance of the model. After many iterations, this optimization process arrives at smaller and smaller penalties with the model parameters found so far, such that further iterations become negligible. However, since the lowest possible penalty and the optimal model parameters are not known, there is no certainty that training has reached the optimum or is as close to the optimum as it could achieve. Even if this optimum were reached for a particular ANN architecture, it is, in fact, *theoretically impossible* to arrive at an ANN perfectly solving a given task of sufficient complexity (i.e., without the slightest error, see $i - iii$ below), although the current state of DL theory suggests that any found optimum should be 'good enough' in practice (Prince 2023). The empiricist process of iterating through examples and nudging model parameters thus involves inevitable errors. These can be categorized as follows.

i. Bayes Error

The success of predicting an output based on a certain input is grounded in a sufficient correlation between the input variable and the output variable. If, therefore, we wanted to predict the duration of stay at the ICU based on a patient's shoe size, even the best model would fail, since the two variables are not correlated in any meaningful way. The Bayes error is inevitable since no practical machine learning task is based on perfect correlation.

ii. Approximation Error

If an ANN model is to fit data, it needs to exhibit a level of complexity that allows for a sufficiently close fitting, i.e., it must suffice to represent the underlying distribution in a meaningful way. In practice, no ANN model is arbitrarily complex, and thus no model can map arbitrarily complex relations. This error is also called the 'bias' of a model. An under-complex, i.e., biased, model will be 'off' in a systematic way (think of a straight line that will be systematically wrong in predicting any periodical function), being unable to fit the more complex training data entirely accurately. A model with high bias is therefore said to be 'underfitting' the training data. Models with a small number of parameters are particularly prone to this error when used with large-sized training sets with high-dimensional data.

iii. Estimation Error

A further inevitable error is due to training data not representing the underlying data distribution (input-output relation) adequately. That is to say; one can only train an ANN on some, often very small, subset of all possible examples. This error is inevitable because there exist no relevant machine learning tasks where all possible data pairs are accessible (such that the underlying distribution is fully known). The estimation error is called the 'variance' of a model, since with varying training data, models with different blind spots would be produced, corresponding to different weaknesses. A model with high variance is said to be 'overfitting' the training data, as it follows the training data so closely as to fail to generalize accurately to new ('unseen') data examples. This leads to the notorious difficulty in machine learning that the training data needs to be sampled in such a way that it is representative of the underlying data distribution, although the underlying data distribution remains unknown. Usually, a dense and hence representative sampling is simply assumed to be the case. (For more details see appendix A.3 and Figure 11). Models with high numbers of parameters are especially prone to this error, if they rely on small-sized training sets.

In sum: To minimize the overall error, the model complexity should increase with the complexity of the true underlying input-output relation, and training examples must be representative of it. Since this underlying relation remains unknown, however, there cannot be any guarantee that the trained DL model will not be wildly wrong with new examples (Delétang et al. 2023). Almost the opposite is true: for any statistical classifier, including complex DL models, examples can be generated where it will fail dramatically – these are referred to as 'adversarial examples' (Szegedy et al. 2014; Goodfellow et al. 2015). This is an inevitable characteristic of DL models (Shafahi et al. 2020; Papernot et al. 2017) and poses problems in various applications, such as self-driving cars (Brown et al. 2018; Tu et al. 2022), making the presence of additional processes to detect such out-of-distribution samples necessary (Amirian et al. 2018). While theoretical guarantees are thus absent, however, the success of

DL models is built on the empirical finding that, in practice, reasonable generalization to unseen examples usually works quite well if it can be achieved through interpolation between seen training examples (see section 2.5).

2.4 DL in Generative Pretrained Transformer Models

So far, we have outlined what ANN models are, wherein their power lies, and where difficulties arise in training them. They are valid for any type of DL model and application, among which we have looked at examples where ANNs are used for classification, namely the image-to-class example of recognizing handwritten digits, since image classification lies at the heart of the DL revolution since 2012 (Krizhevsky et al. 2012; LeCun et al. 2015). Now that LLMs have gained a great deal of public attention, this section provides a conceptual introduction to the workings of generative language models, such as ChatGPT or GPT-4 (OpenAI 2023) – see Prince (2023) for a more detailed introduction.

Generative Pretrained Transfomer (GPT) (Radford et al. 2018) models represent what is called an autoregressive transformer model. An *autoregressive* model forecasts a variable using its past values. Consider the sentence "He sits on a bench". The probability of this sentence equals the probability of starting with "He", times the probability for "sits" given "He", times the probability for "on" given "He sits", and so on. An autoregressive model sequentially predicts the next word by maximizing the joint probability between any next word given the words that precede it. Thus, every new word in a sequence is a function of the preceding words – and the model is a powerful next-word-predictor. The specific characteristic of the *transformer* architecture, originally published by Vaswani et al. (2017), bears the great advantage that it can model relations between words independently of the distance between them in a text and that it allows for what is called efficient 'parallelization', such that training on large amounts of data is feasible.

To gain a more substantial technical understanding of what a 'transformer' does, we must first turn to how words are 'embedded', i.e., numerically represented, in an ANN. The previous example is not entirely accurate since most language models predict not words but 'sub-word tokens'. The term 'token' here refers to any statistically relevant part of a word (this could be, e.g., the parts of a compound word, a punctuation mark in a sentence, a short word itself, or simply any sequence of letters appearing often enough in text). The use of tokens allows the model also to represent words that are not contained in the vocabulary as such (e.g., names), to deal properly with punctuation, and more effectively to relate words and their different suffixes (e.g., learn, learns, learned, learning). Every token is then mapped to a point in a multidimensional data space (e.g., 1024-dimensional), such that there exists a correspondence between points in data space and tokens. Note that this mapping is not deliberately fixed but learned from data during training. To simplify matters, we will, nevertheless, continue to talk about words instead of token embeddings.

A central part of the transformer model is the concept of 'self-attention'. Since language can be ambiguous, it is often not possible to infer how words

relate to each other from syntax alone. Consider, e.g., the sentence "The book does not fit into the suitcase, because it is too big". The fact that "it" refers to the book follows not from syntax but from the meaning of the words themselves. Depending on the context, the model should thus pay more 'attention' to certain words to incorporate their relation to others, hence the description as 'self-attention'. Finally, a 'score' contains the mutual connection strengths between words, depending on the structure of any sequence of words – this score serves to direct the attention toward certain preceding words when predicting a given next word. Note that it is common to have multiple self-attention modules that run in parallel. This is called 'multi-head attention' and achieves a more robust self-attention mechanism (or so it has been speculated, see, Vaswani et al. 2017). In practice, several dozens of such multi-head attention layers are stacked to build a deep model.

The final module in a transformer model is an ANN that takes the representations of the transformed, embedded words and their mutual connection weights as input and maps them to output probabilities for possible next words. A word corresponding with a high probability for the next word is then displayed in textual form (from its numerical representation). However, the same prompts do not always yield identical results because chatbots sample new words from the joint probability density instead of just going for the most probable word. In other words, they choose an option with some randomness, but with a weighting depending on probability densities. (On a side note: the 'creativity' of a transformer model corresponds with such random sampling from a few of the most probable words, which is very different from what we mean by 'creativity' as a human characteristic.) In the above example, the transformer would then be able, building on the self-attention mechanism, to refer "it" correctly to "book", and, e.g., follow that sentence with "So I carry the book by hand". As written above, this potentially holds true, even if the related words are far apart from each other in the text.

It is a matter of interpretation whether a transformer 'learning' and 'generating' text (i.e., predicting with high precision what the next word in a long sequence of text should be while drawing on almost all humanly authored text digitally stored on the web) constitutes 'understanding' of the structure of language and the workings of the world (Bender and Koller 2020; Bisk et al. 2020; Durt et al. 2023; Marcus et al. 2023; Dürr et al. 2023), and what 'understanding' would mean in that case. As we have seen, the outputs are generated upon suggestions by the statistics of words and their relative positions in a text. In human beings, the same result could have been achieved through their semantic knowledge of the terms involved as well as their embodied, lived 'experience'. Human understanding is thus grounded in all sorts of (implicit and explicit) rational, emotional states, such as thoughts, feelings, and bodily sensations, which a human being goes through while, for example, chatting. In contrast, a transformer-based chatbot strings together words that are 'likely' and statistically determined from analyzing a vast amount of text. 'Understanding' in transformers thus refers to such a statistical mechanism. It is an *interpretative*

move to say that with transformers, "statistics do amount to understanding" of semantics (Agüera y Arcas 2022) or that something like this mechanism is what we are referring to when we speak of understanding in humans (we will return to these questions in section 3.1 below). What is striking, though, is that the performance of state-of-the-art LLMs seems to reveal just how much real-world grounding is sedimented in the humanly authored texts on which those systems are trained (Durt et al. 2023), and it raises the question of how much of that is then instilled into the DL models themselves.

We have now outlined the basic workings of transformer-based LLMs. Any qualitative advance in their performance is still based on an architecture with inherent limits – just precisely where the limits of achievable results lie must be researched empirically (Pavlick 2023). Acknowledging such limitations of transformer models, prominent AI researchers, like Yann LeCun, have proposed 'embodied' model architectures that bring us closer to machines with a human-like understanding of words (e.g., "autonomous machine intelligence", see, LeCun 2022; Matsuo et al. 2022, critically discussed in Dürr et al. 2023).

So far, we have introduced the known aspects of DL concerning how it works and what its limits are. As stated above, some of the success of DL is, however, still a mystery and subject to current research. In the next section, we will turn to two example questions that still perplex researchers in DL, to give an intuition about the unknown aspects of DL success.

2.5 Our Shallow Understanding of Why DL Works

Although advances in hardware and the increasing availability of data explain the success of DL to a large extent (Lenzen 2020) and gave rise to numerous algorithmic advances, which account for another large part (Stadelmann et al. 2019b), a unified theory that fully justifies the remarkable performance of DL models is still missing (Plebe and Grasso 2019; Sejnowski 2020; Zhang et al. 2021), although progress seems to be made (Ma et al. 2022; Liu et al. 2022). Two 'unknowns' remain particularly significant, which we will discuss here – albeit only briefly (we describe two further 'unknowns' in section A.4). For a more complete and detailed overview, we refer to Plebe and Grasso (2019) or Hodas and Stinis (2018), for a more mathematical approach to Poggio et al. (2019), and to for an in-depth mathematical investigation to Berner et al. (2022). What *is* known theoretically about DL workings is summarized, e.g., in Roberts et al. (2022); Prince (2023).

i. DL Generalizes Surprisingly Well

As elaborated in section 2.3, DL models with large numbers of parameters are, in theory, prone to overfit the training set. In practice, however, models with a great many parameters generalize surprisingly well to new data examples (Zhang et al. 2017; Soltanolkotabi et al. 2019; Zhang et al. 2021; Martinetz and Martinetz 2023). A good example is the model 'Noisy Student' with 480 million parameters, trained on only 1.2 million images, which might

be expected to overfit drastically, but instead generalizes well (Xie et al. 2020). Current research into generalization focuses on the learning algorithms, suggesting that they exhibit properties of implicit regularization, i.e., a bias that prefers encoded functions of low complexity (Soudry et al. 2022; Arora et al. 2019; Poggio et al. 2019). Furthermore, it was observed that the correlation (more precisely, the mutual information) between neighboring layers in ANNs is high, which is to say that although the system would allow for the difference between levels to be higher, the functions encoded by neighboring layers are in fact not so different from each other (Hodas and Stinis 2018; Tishby and Zaslavsky 2015). In other words, the observed function complexity of an ANN is typically much lower than theory shows it could be. This is what seems to prevent large models from overfitting to the training set and thus from failing to generalize to new data examples. Overfitting was theoretically expected to stop such DL models in their tracks. In light of this, their performance in generalization is surprisingly high (Liu et al. 2022) – nevertheless, overfitting remains an issue when training deep models, and there exist several methods to counter this by penalizing complexity during training.

ii. DL Overcomes the 'Curse of Dimensionality'

Many tasks in computer science become extremely difficult when the number of dimensions of the data space is very high. The data provided for learning LLMs like GPTs could easily run to tens of thousands of dimensions. High dimensional data space is problematic because the sheer number of possible data examples with only small differences increases exponentially with its dimensions, and the number of examples required to cover all relevant configurations consequently increases exponentially as well. Consider a small 5×5 image with a pixel value range $0-9$ (from black to white). To cover all possible configurations, we would need 10^{25} image examples. Extending the image by one single pixel, we would need to cover 10^{26} configurations. That is an extension by 90 trillion trillion configurations (90×10^{24}). In computer science, this problem is referred to as the 'curse of dimensionality' (Bellman 2015; Novak and Woźniakowski 2009; Goodfellow et al. 2016). The data space in most DL applications is very high. Surprisingly, tasks involving high-dimensional data can, and have been, solved successfully for many applications using deep learning. One hypothetical explanation for this corresponds to an important idea underlying machine learning, namely that all meaningful data lies on a lower-dimensional sub-space (usually referred to as 'manifold') embedded in higher-dimensional space (Bengio et al. 2013, illustrated in Figure 12b). What does this mean? Goodfellow et al. (2016) provides a helpful illustration: Although we live in three-dimensional space, we essentially move on a two-dimensional manifold, i.e., the surface of the world, embedded in three-dimensional space. Thus, standing at a random location, we can usually ignore being above or below ground (for all relevant purposes of a given task). Likewise, the set of all possible images that show a face, for instance, is far smaller than the set of all possible images. Machine learning seems to be able to latch

onto this, which simplifies its tasks drastically. Although the 'manifold hypothesis' is not apt for all problems, and much remains unknown, there is a good deal of evidence that supports it (Goodfellow et al. 2016; Brahma et al. 2016).

These are just two examples of why our understanding of DL systems is somewhat shallow. Much more research needs to be conducted in this area if we are to reach transparency or 'explainability' in DL.

3 Work Program: Reconfiguring a DL Assessment From a Humanities Perspective

Having outlined the basic principles of DL, we can now ask how DL, and the applications to which it is put, can be engaged from a humanities perspective, and under which conditions such an engagement benefits society and culture. We want to make it clear from the outset that this work program is necessarily limited in scope and that we have primarily identified issues that we deem urgent – we invite others to chime in, further elaborate, and extend the issues addressed here.

The revolutionary potential of recent DL innovations makes it pertinent to reflect explicitly on the anthropological contexts within which we venture any constructive interpretation and critical assessments of DL: firstly, because these interpretations differ greatly today in their outlook and are in little constructive dialogue which each other; and, secondly, because such anthropological views and values necessarily shape how we organize our societies, and therefore form standards against which any technological innovation is measured. For these reasons, we aim, in this section, to provide some resources for addressing fundamental questions around DL raised from a broadly humanistic perspective.

We begin with a note on *humanism*. In what follows, we work within a broad humanistic tradition, conceived as a field in which different approaches, traditions, and streams may align with regard to shared interests and goals. While not necessarily religious in outlook, this view is more inclusive than secular*istic* and exclusively non-religious accounts of humanism (Flynn 2002, see also the website of humanists.international). Acknowledging the limits of each scientific approach to explaining and making sense of the 'human', such an inclusive humanism is open to religious and spiritual outlooks, alongside those who ashew, or do not stress, such a perspective. More strongly, we would argue that the frame of a 'religion-secular' distinction itself is not helpful and that, particularly in Western countries, arguments about 'what really matters' are conducted on the conceptual territory of the *human* – not necessarily the 'religious', but neither the absence thereof (Grey and Dürr 2023). (We stress this not least because we have ourselves experienced the fruitfulness of dialogues which include a wide range of perspectives on the human – religious and secular – in debates around the future of humanity in a digital world.) Furthermore, this inclusive humanism sees the value of the human person not in competition with those entities with which humanity shares its rationality, animality, and

life itself. Rather, it is the valuation of the human that leads inclusive humanists to value the world of which they are a part – thus, we agree with the line of questioning of existing approaches to 'inclusive humanism' (Antweiler 2012, 2013) as well as with some concerns of (critical) 'post-humanism' (Foucault 1990; Herbrechter 2009; Wolfe 2010; Braidotti 2013), without agreeing with their conclusions.

But how are we to assess DL technology from such a humanistic perspective, or within the framework of the humanities as disciplines addressing 'the human' (broadly conceived)? The point of departure, for us, is minding the *use of language* with regard to DL. How we talk about technology – most notably in marketing campaigns but also in research, journalism, and popular culture – has practical consequences. Language both opens and limits the world we can inhabit (Wittgenstein 2013 [1921], 5.6). In the age of DL, computational theories of mind and the language of human-like computational models (present or anticipated) heavily influence our cultural imagination, self-experience, and interpretation of reality: they impact our everyday lives. Such conceptions turn into public narratives, into socio-cultural notions that transform our ideas and ideals of 'intelligence', 'life', and what it is to be 'human' (Leung 2019). AI research, correspondingly, has long ceased only to concern the use of computers to get useful things done. Instead, some researchers make – implicit or explicit – claims about reality as such and aspire to answer 'big questions' about the nature of human beings, mind, behavior, and life itself (Boden 2016, see, e.g., Tegmark 2018). Not least in journalistic settings, AI researchers and engineers are increasingly asked more about such human questions than about the technical details of their research and actual competency. Ultimately, conceptions of AI feature not only as elements in explicitly articulated theories and world views but are also always elements of broader socio-cultural imaginaries, implicit world views, and quasi-metaphysical basic assumptions about reality. We believe that the elucidation and assessment of such fundamental *questions about technology and the human* are of the utmost importance today.

Outlook on this section:

In what follows, we will first argue that an engagement from a humanities perspective (i.e., having humans in view) must begin with differentiating 'the human' and 'technology', while considering that the two are always also enmeshed, such that they should neither be confused nor separated too neatly (section 3.1). These considerations will further show that any assessment of DL is inevitably grounded in anthropological, epistemological, and ontological presuppositions (traditionally addressed and reflected by the humanities) and that such statements are always interpretative and thus also questionable from various other perspectives. We then argue that current assessments often lack explicit reflection of their anthropological presuppositions and that the humanities can help clarify and navigate the debate by thinking about such assumptions and bringing them into the discussion, not least to foster constructive dialogue between rivaling viewpoints (section 3.2). Finally, we focus on

some fields of inquiry that require attention from the humanities for a holistic assessment of DL, and offer resources of ongoing work in corresponding fields (section 3.3) – in this last section, we provide practical follow-up questions pertaining to the issues addressed.

3.1 Philosophical Foundations: The 'Human' and 'Technological' Factors in Human-Technology Relations

Any holistic approach to the assessment of DL from a humanities perspective must begin with and address the 'human'. Ultimately, it is *human* actors who create and deploy technologies at a scale that has a lasting effect on the world we inhabit – the notion of the 'Anthropocene' (Crutzen and Stoermer 2013) refers precisely to this fact. From this perspective, it is vital also to note that *we* are ultimately responsible for what we do with our technologies. Therefore, a clearer view of the complex ways in which human behavior and technological innovation jointly transform our world is part of charting the way with responsible use of DL systems. This requires that we have some grasp of the qualitative difference between human beings and their technologies. However, it is precisely this that is called into question in the age of AI.

The confusion of human beings with technology can be analyzed as the result of two related tendencies: (*i*) a tendency to anthropomorphize AI, and (*ii*) the corresponding tendency to technomorphize human beings – both of which have a long pedigree. While we are convinced that we should not confuse the human and technology, neither should we separate them all too neatly. Therefore, we will consider the fact that (*iii*) technology is part of and shapes human life (nature and culture), such that human beings must be understood as inherently related to them. Further practical and applied questions and suggestions stemming from these observations are provided in section 3.3.

i. Human-Like AI? On Anthropomorphizing Technologies

'Anthropomorphism' is the act of attributing distinctively human-like emotions, mental states, behavior, and even subjectivity, to non-living objects, animals, and, more broadly, to both natural or supernatural phenomena (Epley et al. 2007). With the increasing performance of AI systems – and, especially brain-inspired AI like DL – and their embedding in the real world (e.g., as robots with human features, as virtual assistants with human voices, and the like), possibilities for confusing human beings with AI systems steadily increase. The result, among our concerns, is the attribution of distinctively human qualities to DL systems, sometimes also referred to as 'mind perception' (Waytz et al. 2010). This is a notable propensity not only in the public sphere but also in AI research (Proudfoot 2011; Salles et al. 2020; Watson 2019; Cave et al. 2019; Lipton and Steinhardt 2019). As a caveat, it must be noted that, historically, there have been different kinds of definitions of AI. Russell and Norvig (2021) suggest two main strands: one explicitly pursuing

"human performance" and the other preferring an "abstract, formal definition of intelligence called rationality" (which is cast in terms of goal-directed behavior, problem-solving, thinking and acting rationally, etc.). On this map, human-centric approaches are particularly prone to anthropomorphism, and explicitly so. But, at least since Aristotle's seminal definition of human beings as 'rational animals', the second definition can lend itself to it as well.

Blake Lemoine, a former engineer of Google for 'Responsible AI', has made the news with the claim that their "Language Model for Dialogue Application" (LaMDA) supposedly has awareness of its rights and needs, is afraid of death and thus sentient (Lemoine 2022; Tiku 2022). Others argue that robotic AI systems are candidates for personal rights (Gunkel 2018), which is true, particularly for people under thirty who believe that future robots will develop cognition and affect (de Graaf et al. 2021). Anthropomorphization is also observable in the phenomenon of bonding with chatbots, social bots, and care bots, which in many cases leads (positively or negatively) to the 'personification' of bots and AI systems (Dosovitsky and Bunge 2021; Skjuve et al. 2023; Crolic et al. 2022). This is not least due to the fact that these are engineered to engage the emotional needs of specific users and to create the illusion of mutual care (Darling 2017). Cultural variations of these phenomena can be observed, for instance, in a comparison between Europe and Japan (Haring et al. 2014; Robertson 2014). This seems to have something to do with the religious background of the Shinto religion, or 'way of life', in Japan, which ascribes spirit and personality to both organic and inorganic things (Robertson 2018). Thomas Fuchs even diagnoses a novel form of "digital animism" in Western societies (Fuchs 2022). There is a notable tendency of people to trust computers more than human beings in decision-making processes (Bogert et al. 2021), leading to an 'overtrust' (Hardré 2016; Aroyo et al. 2021). AI systems are being perceived as human-like *but* more 'objective', 'reliable', and 'trustworthy' compared to rather 'erratic', 'biased', and 'unreliable'. Neglecting that such systems lack any form of emphatic understanding of the human life-form (Fuchs 2022; Dürr et al. 2023) is most consequential if it leads to deployment in decision-making processes that existentially affect human lives, e.g., in jurisprudence (Ryberg and Roberts 2022), policing (McDaniel and Pease 2021), banking (Donepudi 2017), and insurance (Lamberton et al. 2017). In light of these dynamics, the clarification of terminology is pertinent, alongside anthropological, philosophical, and religious background assumptions.

In both AI research and among the broader public, the language deployed to speak about DL models overlaps substantially with everyday language about human beings (Salles et al. 2020). Kostopoulos (2021) has argued that in the attempt to communicate the capabilities of AI, spokespersons in research, industry, and journalism reach for parallels with human capabilities using vocabulary that is characteristic of human behavior. Although there is broad consensus in research that today's DL models are nothing other than a complex mathematical function, they are characterized as having the ability to "read and comprehend", to "compose music", to exhibit "curiosity" or "creativity",

to be "afraid", and so on (e.g., in Hermann et al. 2015; Mozer 1994; Reizinger and Szemenyei 2020; Nguyen et al. 2015; Lipton et al. 2018). Of course, humanizing technology for the sake of communicative or pedagogical simplification is a frequently encountered phenomenon. It has been common, e.g., in control theory, to speak of a controller 'seeking' a target value, although no one would think the controller is consciously doing so. However, with today's AI, that is not quite so clear anymore. Although some use anthropomorphic language metaphorically, as in control theory, others would say that, in principle, the terms used are equally appropriate or inappropriate for humans and technology, as the difference between the two is, ultimately, a matter of degree (on different ways to characterize this relation see Davison 2021). However, such interpretations and their philosophical premises remain largely implicit and are often not given enough attention (more on that in the following section). Lipton and Steinhardt (2019) argue that we should not take such anthropomorphizations lightly. They speculate that the transfer of qualities from the human to the machine is partly due to performance and funding incentives, i.e., using anthropomorphic language with regards to algorithms increases attention from media, donors, institutions, and colleagues in the field (Stadelmann et al. 2019a).

The main problem in using anthropomorphizations with AI systems is that it obscures the nuances, intricacies, and workings of the actual technology – which makes it difficult to adequately assess it. This can go both ways. It can strengthen *unwarranted confidence* in the technology's capabilities, e.g., by speaking of 'learning', which in humans refers to an adaptive ability to cope with new environments, where there is, in fact, function approximation, which does not generalize well (Brooks 2017). Negatively, it can also give rise to *fears*, e.g., by speaking of algorithmic 'bias', which in humans usually goes hand in hand with bad intentions, that cannot, in the same way, be attributed to algorithms, or, more extremely, in doomsday prophecies of a superintelligence purposefully eradicating humanity (see, e.g., Yudkowski 2023). Note that both terms, 'learning' and 'bias', refer to real technical issues with social implications, but we propose that they are best addressed without anthropomorphic distortions.

In sum: Using anthropomorphic language with reference to DL systems makes it increasingly difficult to distinguish between human actors and their technological counterparts. While the advancement of DL application blurs the line between them, we deem it urgent to think more deeply about this difference, and ask what makes human beings unique *vis-a-vis* machines.

ii. Machine-Like Humans? On Technomorphizing Human Beings

The flip side of confusing technology with human beings is the tendency to 'technomorphize' human beings. This has gained traction with the growing mutual relationship between neuroscience and AI, and the rise of DL as a 'brain-inspired technology' (Salles et al. 2020; Hassabis et al. 2017). One initial aim of creating correspondences between the workings of the brain and AI

systems was to better understand the *human* brain, self, and behavior (see, e.g., Huerta et al. 1993; Waldrop 2012; Prescott and Camilleri 2019). Indeed, AI can be very helpful in researching human beings, but its architectural similarity with the human brain should not be overstated, as the majority of what we know, e.g., about the learning process in the brain, has not been integrated in DL – or only in an immensely simplified manner (Schmidgall et al. 2023; Lillicrap et al. 2020; Ullman 2019).

DL anthropomorphism, however, and the dynamics of seeing ourselves in the image of our technology, has a pedigree reaching back to antiquity. (Müller and Liggieri 2019; Cave et al. 2020). It gained modern plausibility with the scientific and industrial revolutions, and the ascent of an all-encompassing mechanistic world picture since at least the 17th century (see, e.g., Boden 2008; Black 2014; Jank 2014; Dürr 2021, 2023; Sarasin 2001). Surveying these developments allows one to identify several leading metaphors which have impacted the conceptualization of human beings – especially as to how their bodies 'function'. Such metaphors usually mirror the most advanced technology of a certain era: in Descartes' time, these were organ pipes or the automata in the Garden of Versailles; later came cameras, radio, and the electrical systems of the early 20th century. It is not surprising, then, that computer science now informs many of the current models and conceptualizations of the human: i.e., human beings as 'biological computers', 'informational patterns' or 'processes', 'algorithms', 'software' or 'mindware' instantiated on the 'hardware' or 'wetware' of the body (see, e.g., Bray 2011; Tegmark 2018; Clark 2008; and, for a critical perspective Weizenbaum 1976). Such metaphors are often deployed without explicit philosophical intent, but they nevertheless convey an anthropology that we tentatively characterize as a 'computer-anthropology'.

Such metaphors have gained particular traction in cognitive science and the analytic philosophy of mind insofar as those have been rooted in behaviorist and functionalist frameworks of the mind (Rescorla 2020). Behaviorism deliberately brackets the deeper questions of *what* intelligence, understanding, curiosity, etc. are, and instead ascribes these characteristics to everything that passes in behaving *as if* it exhibits them (see, e.g., the famous 'Turing Test' in AI, Turing 1950). However, in order actually to understand and engineer intelligence, the question of how intelligence (or at least how some form of intelligence) works must be answered on a practical level. Thus, the field of cognitive science and the AI project – purposefully framed as the engineering quest to simulate human intelligence (McCarthy et al. 1955) – had to overcome the purely behaviorist approach to intelligence. This was achieved on the basis of functionalism (e.g. Putnam 1960) with the central concept of mental representations (e.g. Fodor 1975; Heil 2020; Pitt 2022; von der Malsburg 2023). According to representationalism, mental states and processes are constituted by their functional role in a system of symbolic structures. In our context, the system is the mind materialized in the brain, and its symbolic structures are representations of some sort, e.g., inner representations of things in the external world. As such, the mind is perceived as a machine that follows strict

syntactic rules to manipulate symbols and sequences of symbols in a meaningful way, i.e., it processes information toward certain goals. Notably, such an 'informational' account tends to focus – almost exclusively – on the human *brain* as a 'computational engine' (Churchland and Sejnowski 1992; Churchland 2013; Clark 2013). This brief outline of core assumptions that add up – implicitly or explicitly – to a 'computer-anthropology' would deserve a much fuller treatment here. We must confine ourselves here to three critical concerns that indicate the significance of deeper reflection on these issues:

Firstly, with regard to the exclusive focus on the brain, the claim that the workings of the human brain – and mind! – are essentially comparable to AI is based on the strong and highly contestable philosophical assumption that both are essentially mechanistic processes (Boden 2008). Kenny (1984) has provided helpful clarifications in addressing what he termed the "homunculus fallacy" (pp. 125–136) – also addressed as "mereological fallacy" (Bennett and Hacker 2022, pp. 79–93), and more broadly as 'cerebrocentrism' (Hagner 1997; Fuchs 2021; Dreyfus and Taylor 2015, pp. 107–123). This consists of taking "predicates whose normal application is to complete human beings or complete animals and apply[ing] them to parts of animals, such as brains, or to electrical systems." (Kenny 1984, p. 125). As if the brain itself were like 'a little human being' (*homunculus*), doing the perceiving, thinking, etc., that we usually ascribe to the whole human being. Ultimately, this would result in an infinite regress of trying to explain the capabilities of the *homunculus* with yet another little man inside it, etc. In Kenny's view, the fallacy is still "commonly defended as a harmless pedagogical device", against which he argues "that it is a dangerous practice which may lead to conceptual and methodological confusion." (Kenny 1984, p. 125). Parts of human beings (e.g., the brain) or technical devices (e.g., DL systems) can be in certain "states", which can be described by their internal (physical) properties, but that is categorically different from a "capacity", which usually can be specified with a description of "what would count as the exercise of the capacity" (Kenny 1984, p. 129). This holds against critics, who say that knowing something *is* to be in a neural state (e.g., Dennett 2007; Searle 2007), because "to know something is ability-like, and hence more akin to a potentiality than to an actuality (a state)" (Smit and Hacker 2014, 1084). Thus, confusing mental capacities (like knowing or understanding information) with physical states and processes (like containing information or performing operations on information states) results in attributing capacities – which properly are those of whole human beings, persons, or to some degree animals – to the brain, or, for that matter DL systems. The result of this can be both the anthropomorphization of DL and the technomorphization of human beings.

Secondly, purely formal approaches to cognition or intelligence – regarding them as encoded functions – fail to include our subjective everyday experience (Fuchs 2018). Janich (2009) illustrates this problem by considering an anatomist investigating the human skeleton. Her findings are valid, independently of her having a skeleton of her own, because her *explanandum*

is independent of her own constitution in that matter. With regard to her research object, she is a third-person-perspective observer. However, the same does not hold true for a physiologist investigating 'seeing' in the visual system, for he can see and knows what seeing is from everyday experience, long before entering the laboratory. Without his pre-scientific practice of seeing, he has no *explanandum* at all, which means that physiology does not define the word 'seeing' as an *explanandum*; rather, it stems from everyday language. In contrast to the anatomist investigating the skeleton, the physiologist investigating the visual system has no other option than to take a perspective of participation concerning his research object. The search for a formal description for the human mind, or 'intelligence', thus faces the serious issue that a substantial part of what constitutes everyday human cognition – as with 'seeing' – must be presumed and can only lie at the basis of a formal account, not at its conclusion. Following Janich's argument, cognition defies formal definition because the formal method has no language for any form of participatory perspective. Thus, it can only ignore the fundamental problem that here *explanandum* and *explanans* overlap, i.e., to explain the thing we want to explain, we must use that same thing which is then involved in the explanation of itself, leading to an infinite regress. If this is true, every attempt to 'explain' cognition or intelligence in purely formal terms illegitimately reduces the larger reality underlying these words and must ultimately fail. This has been argued at length with regard to 'consciousness' and pertains to AI: There are attempts to explain consciousness as what results from increasing the complexity of a system as well as what is called the 'principle of recursivity' (i.e., a feedback loop of the state of a system into its further processing). The idea is then to explain consciousness by "piling up" such systems on top of each other so that higher levels (consciously) monitor the lower (yet unconscious) mental states of the system (see, e.g., Dennett 2013, p. 325). However, any effort to elucidate consciousness using higher-order concepts and modes of formalization like recursiveness or even self-modeling ultimately results in an endless cycle of regression (Frank 2002, 2007; Zahavi 2006; Fuchs 2018).

The third concern is more grave still. Modeling humans on computers can have dehumanizing effects (Hoff 2021; Fuchs 2021; Tallis 2004, 2020; Dürr 2021). This is sometimes referred to as 'mechanistic dehumanization' (Haslam 2006; Li et al. 2014; Kuljian and Hohman 2023) The historical record of those who saw and treated people as machines, programmable at will, is sinister (Haslam 2006; Todorov 2016). At the very least, it produces a low perception of human worth with potential long-term consequences, fostering a modern form of fatalism (see, e.g., Courchamp et al. 2018). Ultimately, it is incompatible with core assumptions about human beings, which are consequential for our liberal democracies: core values, such as human dignity, liberty, and autonomy, cannot, in such a take on human beings, be meaningfully maintained because they presuppose something in individual human beings that lifts them out of the realm of disposable things. It seems difficult to argue for the unique and incalculable dignity of a human person from the assumption that they are

'nothing but' computational processes and, as such, completely replaceable with computational processes, say in machines. The same goes for the kind of freedom, rights, and duties we attribute to such dignified human beings to engage in the politics of our democratic societies – attributes we do not grant to algorithms, computers, and robots (at least for now, see Gordon and Pasvenskiene 2021). Thus, even if one tends to believe that a human being could, in principle, be exhaustively modeled by a computer, it would still be prudent not to *assume* that this is the case until the evidence is overwhelming. In the long run, computer-anthropology will have direct consequences, not just for our ethical assessment of DL, but for the principles and values guiding design processes, as well as for political and juridical decisions, and thus for the future of our societies as they grapple with the digital transformation.

These are just three prominent reasons that illustrate why we believe it is vital to reflect deeply and critically on the difference between 'the human' and 'machines' – particularly in light of DL achieving things that were hitherto considered impossible for machines, clarifying what is distinctively human is one of the great tasks of the humanities.

iii. Technological Mediation: Why We Cannot Separate the Human From Technology

It is vital to note that the emphasis on the 'human' here must not be understood within the framework of a naive instrumental conception of human-machine relations: as if neatly isolated 'human beings' were using neatly isolated 'DL tools' for their purposes, by means of their sheer will. Such a view has been labeled the 'value neutrality thesis' of technology: denoting the idea that technology is a morally and politically neutral medium and that the only relevant factor with regard to outcomes is what humans do with it (see, e.g., Pitt 2014). This view is increasingly questioned and challenged by approaches that recognize that values are embedded in technology and that technological artifacts have a kind of agency that needs to be reckoned with, not least because they lastingly affect their 'users' and culture and society more broadly (Brey 2005; Miller 2021; Kroes and Verbeek 2014; Jenkins et al. 2023). Technologies do something to us as we do something with them (Ihde 1990) and thus make vital an encompassing analysis of the structure of human-technology systems as well as their 'co-evolution' (Hughes 1987; Murphie and Potts 2017; Hoff 2021).

Several strands of research in the philosophy of technology (broadly conceived) provide us with helpful resources to conceive in a more nuanced way of human-technology *relations*: technology assessment (Grunwald 2009, 2019a; Winner 2020), media philosophy and media ecology (McLuhan 1994 [1964]; Postman 2006; Strate 2017; Cali 2017), phenomenology and postphenomenology (Ihde 1990, 1995; Verbeek 2005; Rosenberger and Verbeek 2015), and the interdisciplinary field of 'science and technology studies' (Latour 2012; Sharon 2013, see also Sismondo 2010; Felt et al. 2017). The concept of 'mediation' has proven to be valuable: "rather than seeing technologies as functional,

we need to understand how they play a mediating role in human practices and experiences. Technologies-in-use help shape relations between users and their environment" (Verbeek 2015, p. 31). In transforming our environments, DL applications are not merely neutral or passive instruments but have their own kind of agency (Latour 2012; Verbeek 2005). They transform our experiential, cultural, and social environments with lasting effect (Stiegler 2013; Hoff 2021; Kitchin and Dodge 2011; Heidenreich and Weber-Stein 2022). The importance of such considerations becomes more obvious when considering the fact that DL-based systems are not only making suggestions but also making decisions for us, and in a way that no human being has deliberately or strategically planned (Karanasiou and Pinotsis 2017). This practically forces us to revise our notion of human 'autonomy' (Prunkl 2022) (on this see section 3.3.*iii* below). What this amounts to is the need to reconceive the relationship between humans and technology in what we would term a *relational anthropology of technology*. Such an anthropology must account for the fact that human nature, technology, and culture constitute each other and continuously evolve together without either nature, technology, or culture fully determining the others (Leroi-Gourhan 1993; Hoff 2021; Noë 2023). This goes against the grain of both 'technological determinism', for which technology is the only decisive factor (Ellul 2021 [1954]) or 'socio-cultural determinism', for which it is only social and economic factors and human action which determine outcomes (Pitt 2014). Empirically, both sides seem to have a point but are lopsided in their exclusivity of other factors (Grunwald 2007).

The case for a more holistic and relational anthropology of technology sets out phenomenologically from the experience of lived embodiment (*Leiblichkeit*, see Fuchs 2018; Hoff 2021). We are capable of relating to technology in such a way that we relate to the world *through* it ('mediation'). A classic example of this is a blind person's cane, which is integrated into the sensory field so that things are felt with the tip of the cane (Merleau-Ponty and Smith 1962; Polanyi 1967). Another example is prostheses, which has led philosophers of technology to speak of the 'prostheticity' of technology more broadly (Stiegler 1998). Technologies transform our world because we, in many ways, live in and through them. Thus, we are enmeshed with the values embedded in them and the influences they exert on us as we 'use' them (Spiekermann 2023). (This has immediate implications for how we conceive of ourselves as 'free', 'responsible', and 'dignified' persons in democratic societies, but also for how we think about designing, legislating, and deploying technology, which we will discuss in sections 3.3 and 4.)

A promising anthropological starting point for such a project seems to be a line of thought under the notion "embodied cognition", which has recently attracted significant attention within and outside cognitive science, and which is most distinctly represented by theories of "enactivism" (Varela et al. 1992; Thompson 2010; Di Paolo et al. 2017; Hutto and Myin 2012; Stewart et al. 2010; Gallagher 2011, 2017; Fuchs 2018, for an introduction to the varieties of enactivism, see Ward et al. 2017, for an overview over the very dispersed field

of cognitive science in general, see Núñez et al. 2019; Andler 2009; Wilson and Golonka 2013; Margolis et al. 2012).

The main idea of enactivism is that organisms and their environments are interrelated and mutually shape one another. A living organism is an *autopoietic* system (from the greek *auto* = self; *poiesis* = creation or production), i.e., it produces and maintains itself by creating its own parts through constant metabolism, exchange, and interaction with its environment. The lived body plays a mediating role between the living being and its environment, hence 'embodied' cognition. Importantly, this is understood as a 'vital' embodiment, not just any kind of embodiment (Thompson and Stapleton 2009; Fuchs 2022) as enactivism does not sit too well with the idea of 'extended' or 'substrate independent minds' (see Rowlands 2009; Cappuccio 2017; Gallagher 2018, against, e.g., Kurzweil 2005; Clark 2008). Being embodied, a living organism perceives its environment not in a mere passive manner, as does the mind in the functionalist paradigm of mental representation, but it co-constitutes it by its actions. This means that what a living being perceives influences its actions, which in turn constitute what it perceives. The main idea of enactivism is taken up in neuroscience and philosophy under the term 'predictive processing' (Hohwy 2013; Clark 2016; Butlin et al. 2023), even if 'predictive processing' is still framed within the bounds of what we would term computer-anthropology, namely focusing on the brain as a processing machine that constantly updates a 'mental model' of its environment. On the enactivist view, a cat and a mouse have different environments and live in different worlds. They – to follow up on Janich's illustration in the previous section – 'see' the world differently. In this light, cognition is not solely explained from an observer's perspective in terms of information processing. In other words, there is no neutral 'view from nowhere' (Nagel 1989). Instead, the complex and ever-changing patterns of interaction with the environment require a more holistic approach to cognition, which understands this as a value-saturated, intentional, and goal-driven phenomenon (Varela et al. 1992, pp. 205-206; Thompson 2010; Turner 2017; Noble and Noble 2023): one that involves the whole organism-environment-system and, not least, considers that every explanatory perspective is subject to this co-constitutive interrelations as well. Instead of mental representations, enactivism works with the concept of 'flexible neuronal dispositions' which apply in different situations – 'open' behavioral 'loops' that are formed through experience and reactivated in specific situations to 'close' an organism-environment-interaction (this would deserve a more detailed treatment we cannot give here; instead, we refer to Fuchs 2020, 2018). This organic and phenomenological 'mechanism' also applies in technological environments, where it explains the 'mediating' or 'prosthetic' function of technology. This lies in marked contrast to the 'mental representation paradigm' of computer anthropologies (see section 3.1.*ii*), which presupposes a clean divide between subject and world. Enactivism cuts across this divide and thus helps ground a more holistic relational anthropology of technology. This holistic entanglement of

the human being with technology and culture makes clear that 'the human' is constantly negotiated and precarious.

In sum: Our notion of the human is invariably the frame of reference for any assessment of DL technology. Yet, this 'human factor' is co-dependent and co-constitutive with 'technological factors' and 'cultural embeddings'. Together, those factors shape our anthropology and, thus, the socio-culturally malleable frame of reference for how we shape our common life. Bracketing out either the 'human factor', the 'socio-cultural frame', or the 'technological factor' does not do justice to the complexity of the situation we are facing with the digital transformation. We are convinced that only by holding the tension of all three factors (nature, technology, culture) the delicate balance between the humanities, natural sciences, and engineering could be productively struck. Keeping this in mind thus orients the way we ethically and practically engage DL technologies.

3.2 Contextualizing Ethical Assessments of DL

From a humanities standpoint, one vital task is to analyze technology, its impact, and its interpretations against a wider anthropological background. Such broadening and contextualizing of ethical DL assessments is vital if we want to reap the benefits of novel AI technologies while managing their perils. Important research is already being conducted in the areas of 'technology assessment' and 'responsible research and innovation' (Grunwald 2019a; Coenen and Grunwald 2017), 'value-sensitive design' (Friedman and Hendry 2019), 'value-based engineering' (Spiekermann 2023), and privacy and security assessment (European Parliament and Council of the European Union 2016; Liu et al. 2021; Véliz 2021; Curzon et al. 2021), as well as research and the standardization of 'trustworthy AI', which deals with issues of reliability, safety, security, resiliency, accountability, transparency, explainability, interpretability, reviewability, and fairness with mitigation of harmful bias in AI (Kaur et al. 2022; Wing 2021; Chatila et al. 2021; Durán and Formanek 2018; Floridi 2019; Krüger and Wilson 2023; Yazdanpanah et al. 2023; Johansen et al. 2023; Li et al. 2023).

Here we see part of a notable broader 'ethical turn' in thinking about DL, or at least increasing interest in the ethical conditions and ramifications of DL applications, which resulted in an expansion of literature (for an overview of current debates and developments in the field, see, Coeckelbergh 2020; Spiekermann 2019; Dubber et al. 2020; Véliz 2023). One strong emphasis has fallen on our inability to understand the outputs and decisions of DL models (on this, see section 2.5), drawing attention to questions around harmful bias and discrimination in data-based assessments or decision-making support systems (Glüge et al. 2020; Loi et al. 2019; Baumann and Heitz 2022). Other areas of ethical attention include privacy of personal information, free speech, information flows and misinformation, the working conditions of humans training and optimizing models and data sets, military applications (Brundage et al. 2018), and ecological considerations (positively in as much as DL can help

to work toward ecological sustainability (Rolnick et al. 2022), and negatively, given the ecological impact of training DL systems themselves (Strubell et al. 2019; Bender et al. 2021; Crawford 2021)). Such work sees technology as not the solution to our societal and planetary challenges on its own. Just as important is *how* technology is designed, regulated, implemented, and used in our societies (Floridi et al. 2018; Russell 2019). The challenges of the digital transformation require more than a 'technical fix' (Weinberg 1966; Morozov 2014) because, ultimately, it is always *human beings* who deploy, use or abuse novel technical potentials. This, in turn, brings into focus the conditions under which human beings are even *capable* of living with technology in a way that allows for human flourishing.

Having the co-constitution of technology and human self-understanding in view, the increasing deployment of 'human-like' DL systems propels a human self-understanding that might end up losing sight of 'the human' (as discussed above in section 3.1). Therefore, as we have already suggested, the assessment of DL will benefit from a holistic approach, placing technological considerations and ethical evaluations in the broader context of a relational anthropology, epistemology, and ontology of technology. An example of this is the already mentioned nexus between DL systems and ecology (we will address further implications of such an approach in section 3.3): When inquiring into the impact of DL on ecology, who are 'we' that care about ecology in the first place? Why is it that we care about the environment? What motivates us to act for it, and how is this mixed up with technologies? Which values are guiding us in this? How can we achieve our goals in this regard by deploying DL without being driven off course by those technologies or other external factors? Which technologies are really adequate for such tasks, and which are not? In grappling with such questions, it is vital to acknowledge that technology is never neutral but always embodies and implements certain goods, values, and aims – the question is "which ones"? Such questions need to be asked, empirically researched and debated with a broad participation of societal stakeholders (Bélisle-Pipon et al. 2022). The point here is not to 'solve' age-old questions but to navigate consensus or democratic decisions about the future we truly *want*, not just one that technically imposes itself on us. We want to argue, however, that it takes the humanities to help navigate such a process if we truly wish to aim, for instance, at a more just, equitable, and humane society.

Such an aim, it should be noted, relies heavily on the particular anthropology one has as a basis for engaging the questions. If one operates, for example, on the basis of the abovementioned 'value neutrality hypothesis' of technology and a notion of human beings as completely free, autonomous subjects, a different set of ethical issues emerges than if one operates (as we do here) on the grounds of an enactivist and relational account of human beings. Another example is the timeline of ethical issues to be addressed with AI: Baum (2018) differentiates between "presentists" and "futurists" as factions stressing that attention needs to be given to either "near-term" or "long-term" issues with

AI. These debates were intensified with powerful LLMs and public speculations about "emergent properties" and "sparks" of AGI (Bubeck et al. 2023, for a critical perspective on such claims see, e.g., Schaeffer et al. 2023; Durt et al. 2023) and the subsequent open letter to pause "giant AI experiments", signed by leading AI-researchers and CEOs (Bengio et al. 2023). While 'presentists' – in Baum's terminology – argue for the need to mitigate current societal and ecological harm (see, e.g., Bender et al. 2021; Prabhakaran et al. 2022; Gill 2023), 'futurists' urge concentrating all resources on mitigating 'existential risks' (see, e.g., Bostrom 2013; Greaves and MacAskill 2021) not least from an out-of-control and misaligned superhuman intelligence (Bostrom 2014; Russell 2019; Tegmark 2018) or even a so-called "singularity" (Kurzweil 2005; Eden et al. 2012). It is worth noting that such debates are mainly conducted on social media, podcasts, and in the press – economic and political stakes are high. Both sides argue in an all-or-nothing manner, and there is not much communication between factions. Anticipated threats, probabilities, and timescales and thus ethical opinions differ greatly.

The interpretation and associated predictions of DL technologies rest on speculative (philosophical) grounds. The basis for these attributions is often not technical arguments but competing theoretical accounts, conceptions of the human, and even fundamental worldview assumptions. Such background assumptions (pre-)determine any ethical judgment we can arrive at because they set the values, goods, and aims implicit in any ethical evaluation of DL. These often implicit background assumptions are what the recent approach of 'hermeneutic technology assessment' (Grunwald et al. 2023) wants to help elucidate in analyzing technological future visions. Ultimately, DL applications present our societies with challenges that are *more than* technical or even ethical. While classical AI ethics efforts have 'the human' in view and as a reference point for technology assessments, they often take a high value of humans for granted, while it is, in fact, highly contested. Technical developments, like DL, thus put our value systems to the test and bring to the fore fundamental views about human beings, technology, culture, and the world we inhabit. Such views invite conceptual and theoretical analysis as well as empirical research. This is what the humanities, more broadly, can help facilitate and why they are vital for the assessment, development, and deployment of DL technologies. (A good example of such an encompassing approach is Rovetto 2023.)

In the following section, we outline how the humanities may help to navigate the engagement with and assessment of DL systems as we move into a future increasingly impacted by such technologies.

3.3 How to Navigate the Digital Future – Resources the Humanities Provide for the Assessment of DL

A realistic assessment of DL requires us to draw all of the above-mentioned threads together. In this section, we outline some of the questions and issues we deem important with regard to DL technologies – this list is in no way

exhaustive, and we want to invite others to add to, develop, and challenge our ideas.

We have selected three exemplary aspects, which are all classically associated with 'human beings' but are now challenged by DL. Those aspects deepen some of the philosophical issues addressed in section 3.1 above. They are interrelated, elucidate each other, and should therefore be viewed in parallel. (*i*) Firstly, we argue that humans are always embodied and embedded in natural, technological, and cultural environments and that this has significant implications for assessing DL. (*ii*) Secondly, we will consider the challenges we, as rational and responsible beings, face as we try to understand a world shaped by technologies we cannot comprehend. (*iii*) Thirdly, we turn to humans as morally responsible agents and explore how an assessment by the humanities can foster the use of DL systems for good.

i. Human Beings as Embodied and Embedded in an Environment

We have already introduced the basic concept of embodied cognition and enactivism in section 3.1.*iii*. What might this branch of research in cognitive science, drawing from phenomenological insights, indicate for an assessment of DL systems? We approach an answer in dialogue with Fuchs (2022) asking how a DL system (e.g., GPT-4, see section 2.4) 'understands' words.

Understanding in the human sense, argues Fuchs, requires subjectivity: to understand, there must be *someone* who understands, i.e., someone to whom a word means something (see also Spaemann 2006, for a characterization of a person as a 'someone' over against 'something'). According to behaviorist and functionalist accounts, we ascribe subjectivity to things based on solipsism and inference, i.e., we take the 'intentional stance' towards an object by deducing that it is a subject (Turing 1950; Dennett 1989). However, research on embodied cognition indicates that this is not true (Gallagher 2011). Rather, we presuppose selfhood from the outset as we engage embodied participants in a common form of living (Merleau-Ponty 1964; Moyal-Sharrock 2021). Understanding 'hunger', for example, presupposes a sharing of life of our kind in the broadest sense, one within which hunger can be felt. Thus, understanding hunger requires one to have a biological body for which nourishment and the lack thereof really mean something (Jonas 1966; Thompson and Stapleton 2009) – which is why enactivism places a strong emphasis on the biological grounds of distinctively *human* cognition (bracketing out for a moment, whether AI could develop an entirely different form of cognition). Fuchs (2022) terms this sharing of a form of living 'conviviality' (Fuchs 2022). According to this view, even today's most advanced language models, with their surprisingly human-like outputs, do not 'understand' *us* any more than a pocket calculator or a stone can. In this view, substrate does matter (as explained above), and a simulated body in a virtual space – which some label 'embodiment' (see, e.g., Xiang et al. 2023) – still does not feel 'hunger' any more than a simulation of rain is wet. There might be different forms of understanding, as there might be different forms of intelligence (e.g., human, animal, etc.), but

human understanding and the statistical 'understanding' of LLMs differ in at least this characteristic: the lived experience of vital embodiment. To the best of our knowledge, this is a fundamental difference, even though this is highly contested, e.g., by what we have outlined as computer-anthropologies above. In the same manner, consciousness, as exhibited by living beings, cannot arise in an isolated brain (and certainly not in a computer simulation) because it requires constant vital regulatory processes that involve the whole organism and its environment (Damasio 2010; Fuchs 2018; Man and Damasio 2019). What does that indicate for an assessment of DL systems?

Regarding human beings as fundamentally embodied and embedded in an environment shifts the attention away from the fear of having sentient AI anytime soon, toward the more realistic concern that DL systems could catastrophically impact our societies and ecology as powerful (but mindless) tools (see section 3.2 above). The main point here, coming from enactivism, is that even if AI has a distinct form of 'intelligence' that allows it to 'solve problems', only a biological life form (from metabolism all the way up to higher forms of cognition, consciousness, and self-awareness) actually *has* problems it intentionally and existentially wants to solve because it pertains to its self-preservation as a living being. This goes to show that anthropological considerations, far from being distractions, actually set the course for further inquiry and action. If one is convinced that today's AI systems are a step toward sentient AI or even a superintelligence that could come up with the intention to eradicate humanity, one will take the prevention of such an 'existential risk' to be "strictly more important than any other global public good" (Bostrom 2013, 2014; Russell 2019; Greaves and MacAskill 2021). On this view, when push comes to shove, resources should be directed toward AI alignment efforts and not ecological or social challenges associated with DL. Thus, anthropology (broadly conceived) is a pressing issue because it weighs rather heavily on the values, criteria, and priorities we set in the development and deployment of DL systems.

Furthermore, the embodied embeddedness of human beings underpins and reinforces efforts toward ecological sustainability because it regards concerns about our planet and other life forms as deeply human concerns. An embodied view of human beings gives weight and urgency to those efforts since it makes clear the existential connection between human beings and the rest of living things in nature, of which we are part. This realization clarifies that a human-centered perspective in AI ethics need not be in conflict with ecological concerns. All of these indications suggest that the following question should also be addressed from an encompassing humanities perspective.

Follow-Up Questions:

- How should we interpret the success of DL systems in solving tasks that were long thought to be solvable only by human beings? What does this tell us about ourselves? What can the history of thinking about human beings contribute to this question?

- What is 'the human', what is 'technology'? How can we elucidate the difference between human beings as living things and technology, and how do we assess the multiple frontiers on which this difference is challenged?
- How can we bring to light, challenge, and – where necessary – replace the anthropologies implied in DL applications and their deployment? How, particularly, can we leave behind purely behaviorist or functionalist models of human beings?
- How should we conceive of human-technology-relations? How should we deal with the fact that human beings are capable of existentially relating and bonding with non-living technological artifacts? And that these artifacts have a form of agency, that cannot be completely predicted or controlled? Are there systemic effects or risks through the interaction of human beings and such technologies that are unwanted for? What does it mean anthropologically that DL technologies are now an active and formative part of the human lifeworld?
- Who decides on what we want such systems to do? Are there things we do not want them to 'learn' or be capable of? Around which values, standards, and future visions are we creating, designing and deploying novel technologies? Who sets those markers, and with which legitimacy?
- How can we deploy DL systems to foster shared embodied experiences, community, and societal unity in the lifeworld toward human flourishing?
- How can we deploy DL systems to foster ecological sustainability? How, particularly, can we foster a renewed focus and valuation of the 'human' in a way that does not imply a devaluation of the rest of living things in nature – with whom we share much precisely of what makes us human in the first place, i.e., our animality, our biology, life, and ultimately being itself? What can we learn here from the humanistic tradition as well as its critics, such as critical-posthumanism?

ii. Human Beings as Rational Animals Who Inquire Into Reality by Way of Theory and Knowledge

At least since Aristotle, human beings have considered their 'rationality' – closely linked with their linguistic capacity – the defining feature of what makes them 'human'. The original Greek definition provided by the philosopher is *zoon logikon*, which usually is translated as 'rational animal' but might also, as Charles Taylor correctly suggests, be rendered as 'animal possessing language' (Taylor 2016, pp. 338). We are, in this classic view, animals with the capacity for linguistically mediated reason. Reason and language, furthermore, are closely linked with 'intelligence' (in Greek *nous*, and in the Latin rendering *intellectus*), i.e., the capacity to understand, to judge, and to will things. Given the relational way in which humans constitute themselves together with technology and culture (section 3.1.*iii*) it is clear, that the human capacity to understand, judge, and will must be thought enmeshed with techno-cultural settings. Which requires us to clarify how we speak and what we mean, when we

speak about 'rationality', 'intelligence', and 'language' with regard to human beings or technology (Davison 2021; Dürr et al. 2023; Durt et al. 2023). The history of these terms, it must be noted, is highly complex, and at certain times, their meaning has been stretched to the brink of referring to the contrary to the original sense of the word – especially with regard to the capacities of AI (see Hoff 2021, for a constructive overview of these developments). Today, the capabilities of current generative DL applications urge clarification.

Prominent LLMs, like ChatGPT and GPT-4 (at the moment of writing), generate content that human beings perceive as meaningful, helpful, and creative. They exhibit behavior (particularly linguistic output) that we commonly associate exclusively with human beings and that with human beings involve understanding, knowledge, and meaning. We have already mentioned that this is interpreted in very different (partly contradictory) ways. Some think this amounts to understanding, knowledge and intelligence in a human-like sense (see, e.g., Piantadosi and Hill 2022; Agüera y Arcas 2022) while others are more skeptical (Marcus et al. 2023), believe that there are other ways to explain these capabilities (Durt et al. 2023) or think we are dealing here with 'stochastic parrots' (Bender and Koller 2020; Bender et al. 2021). From an enactivist perspective, LLMs are seen as technical systems that contain information and perform operations on information, but they do not 'know' that information, much like a bus schedule contains information about bus departures but does not *know* the time of departures (Fuchs 2022, see also Tallis 2020). From a technical, engineering perspective, however, such models fundamentally encode a high-dimensional joint probability density function (PDF) of the next word (token); given a large context of previous words, one can explicitly state what its 'understanding' constitutes: To the degree that (a) language (i.e., the sequence of words) is (or, can be modeled as) a random process and (b) all variables influencing the token sequence are part of the modeling, this PDF statistically constitutes *everything there is to know* about the next word. In human beings, however, speaking meaningfully involves intentionality and extralinguistic context – as we are embodied and embedded beings (section 3.3.*i*). What the next word in a sentence of ours is, can be statistically guessed (and in many instances adequately so), but it is not confined to or determined by technical processes, and our variations are not due to randomization. Thus, the technical grasp on 'understanding' in DL helps clarify what such statistical 'understanding' is lacking from a more encompassing view within the humanities.

Everything hinges on the question if we can truly say that a DL application 'has' goals and thus exhibits meaningful 'behavior' in any sense with which we attribute those terms to human beings. One may press this further: where does the process begin? Usually, with somebody who prompts the system. And where does it come to a meaningful conclusion? Usually, with somebody to whom the output of the system means something. In human-technology relations, it is human beings, holistically, which are the relevant subjects when it comes to knowing, willing, understanding, behaving, and the like. Still, the

difference between a DL application and a human person, as mentioned above (section 3.3.*i*), is that between something and someone: while 'something' can be explained by a description of the state it is in at a given moment in time – *someone* cannot (Kenny 1984; Spaemann 2006, see also Bennett and Hacker 2022). This distinction does not 'solve' the question of what makes a person or what constitutes mind in a sense that only complete human beings, holistically conceived, can be said to possess. But it holds a tension that is foundational to our societies and that needs further clarification.

Explicating this tension (i.e., holding a clear separation between human beings and technological artifacts, but without falling into the trap of overlooking the agency as well as the formative role such artifacts play in our human constitution) is perhaps one of the greatest tasks for the humanities today. Such an account would reckon with the relational and co-evolutionary way of conceiving human-technology systems we have briefly sketched above (section 3.1.*iii*). The depth and speed with which such DL technologies transform our environments and thus influence the ways our human minds constitute themselves require us to rethink our notions of 'knowledge', 'understanding', and 'explanation':

The combination of technical mastery and explanatory mystery in DL marks a significant step in the history of human engineering and scientific inquiry into reality. Although we have some knowledge about why DL systems arrive at high performances, the workings of trained DL systems remain opaque to our understanding (section 2). We can now engineer and deploy working systems whose inner workings we do not understand and they successfully solve problems of such complexity that we cannot possibly comprehend corresponding solutions. This marks a shift from causal explanation toward statistical *correlation* (Brodie 2019). This corresponds with debates in the philosophy of science, which increasingly question the dominance of causal explanations (Reutlinger and Saatsi 2018) and moving beyond epistemic reliabilism (Goldman and Beddor 2021). An illustrative example in the context of scientific inquiry is the problem of protein folding. The three-dimensional structure of a protein defines its function and is determined by an amino-acid sequence. However, the relation between the amino-acid sequence and the resulting structure has been a puzzle of the first order in biology for decades, and there seemed to be no feasible way of proceeding from one to the other by calculation. With the help of DL, this problem has been successfully solved for the majority of known proteins (Jumper et al. 2021), although there is still little knowledge on why a specific structure follows from a respective amino-acid sequence. Nevertheless, biologists in many fields can now work with these predictions, for instance, in drug design (Eisenstein et al. 2021). Thus, DL confronts us with the spectacular practical advances that cannot be theoretically explained. For the scientific community, this is at once exhilarating and demoralizing. We now have a fuller database of crucially important protein structures, unthinkable even a decade ago, but, at the same time, we do not understand how protein sequence leads to protein structure — for all immediate practical purposes we

do not need to understand it, since we have DL. A question of such importance – how sequence determines structure – may now go under-researched, and under-funded, because of DL leaping from one to the other.

This shift in scientific practice – not least away from reliabilism – seems to bring us back closer to more practical notions of 'understanding' (Grimm 2021) as developed by phenomenological philosophers like Martin Heidegger and Maurice Merleau-Ponty. They conceived of the mind not as a detached subject over against a material world to be theoretically dissected, but rather as always already "being-in-the-world" in a way that allows us to practically cope with the world (Heidegger 1996 [1926]; Merleau-Ponty and Smith 1962, on this, see also Dreyfus and Wrathall 2014). This philosophical tradition has influenced both enactivism and salient approaches to science and technology today. Rather than seeing science as a systematic representation of the world (e.g., the "scientific image" in Sellars 1962), such approaches conceptualize our scientific endeavor as a set of human practices that render the world more intelligible by continuously and interactively transforming environments (see Rouse 2019 on the basis of "niche construction" theories, Odling-Smee et al. 2003).

In philosophy of technology, this shift to practice leads to a way of engaging novel technologies – from design to use – in practical, even pragmatic ways that amount to what since antiquity has been called 'wisdom': a combination of practical skill and mastery and rule-based knowledge, *alongside* a sense of one's limits in knowing and ability to handle things (on this, see section 3.3.*iii* below). Such an outlook cannot depend on the rationality of controlled and verifiable procedures alone but faces the need for personal responsibility, virtue, and wisdom in processes of discernment and conjectural explorations guided by values (Hoff 2021; Spiekermann 2023). Some developments in what is called 'hermeneutic technology assessment' and which amounts to an "explorative philosophy of emerging technologies" come close to this (Grunwald 2016; Grunwald et al. 2023).

Follow-Up Questions:

- What makes up the 'holism', with regard to complete human beings and persons, implied in enactivism? What truly makes the difference between an embodied human being as persons, who can be said to possess a mind in the sense that it is adequate to attribute intelligence, knowledge, understanding, will, and the like, and DL applications – or simply to parts of human beings, such as brains?
- Do DL systems represent a novel or perhaps stand-alone form of rationality? Are they indicative of 'how human intelligence works'? What does the strong performance of DL mean for our assessment of 'intelligence', 'rationality', 'understanding', 'sentience', 'consciousness', 'intentionality', 'feeling', 'learning' and the like? What insights from different traditions can the humanities bring to the table when it comes to analyzing such traits?

- What do we mean, conceptually and hermeneutically when we use such terms in human beings, animals and algorithms? What can we learn from the history of interpretations of these terms? How is research concerned with such traits in animals informative of our understanding of it in humans and technology? How are they constituted in technical and cultural settings, in ways that subvert any neat separation of human beings, cultural forms of life, and technological artifacts?
- How can we adequately speak of DL technology in communicative or pedagogical contexts? How do we avoid applying predicates that normally apply to complete human beings or complete animals to parts of human beings or parts of animals, or even electrical systems in a way that is fallacious and risks conceptual and methodological confusion? How, more broadly, can we avoid anthropomorphisms and technomorphisms?
- How do we mediate and communicate between rivaling theoretical outlooks on the world, human beings, technology, and especially intelligence – e.g., between analytical positions, focusing on formal approaches and enactivist positions, focusing on the holistic embeddings of processes that are taken to be irreducible to formalization?
- How does opacity affect the ethics of AI deployment? In biology, for example, results can be tested insofar as they work or they do not. That does not apply in the same way, without a high price, in societal areas where human beings and their freedoms are directly at stake. What factor should 'causal explanations' play in the evaluation, prediction of, or ruling over human behavior? In which areas should corresponding systems be deployed, and in which ones should we refrain from this?
- Does scientific inquiry require causal explanations? What is the role of statistical knowledge in science? What is the qualitative difference between causal knowledge and statistical knowledge? And how does DL factor in such debates?
- How could novel models and modes of knowledge, understanding and coping in terms of practical wisdom look like that would do justice to the relational nature of anthropology of technology?

iii. Human Beings as (Morally) Responsible Agents

The complexity and opacity of DL systems force us to clarify our notions of 'autonomy', 'agency', and 'responsibility'. Who is responsible and should be held accountable for the real-world consequences of deploying algorithms with the power and capabilities we are witnessing in the latest DL applications? (Wagner 2019; Martini 2019; Kaun 2022) This is particularly urgent to ask because the architecture of current DL systems cannot fully prevent unexpected, potentially harmful 'rogue' outputs (see 2.3). In which areas of life should we deploy applications whose results we cannot understand or meaningfully reconstruct? To act upon the output of a statistical model without

the possibility of tracking and understanding sequential causal steps complicates the moral evaluation of those actions. This is aggravated by the lurking possibility of bias, deliberate manipulation, and adversarial attacks, which cannot, in principle, be excluded (see section 2.3). Relying on opaque DL systems thus further complicates the already challenging notion of the responsibility of engineers, labs, or companies, especially, in the latter case, with respect to their increasing weight as global economic agents, able to reshape national and international money flows at large scale. It is clear that we are facing issues here that require not only technical adjustments but also philosophical reflection and practical (societal, political, legal) measures often discussed under the label of a 'trustworthy AI' (on this, see section 3.2 above).

More profoundly, these constellations require us to ask ourselves if and how we can even consider ourselves to be 'autonomous' in our decision-making processes at all. What is the role, range of possibilities, and scope of freedom of human beings in human-technology systems? Prunkl (2022) suggests that 'autonomy' can be analyzed in (at least) two dimensions: Firstly, authenticity, i.e., if beliefs, values, motivations, and reasons held by a person are in a relevant sense authentic to that person, and not the product of external manipulative or distorting influences. And secondly, agency, i.e., if a person is able to act on the beliefs and values they hold. Given our relational approach to anthropology, neither dimension can be construed in a way completely independent of either cultural or technological factors. Here, the humanities have insights to offer into human behavior, motivation, and, more broadly, freedom (Calvo et al. 2020; Wagner 2019, see, e.g.,). Given that technical innovations will continue to transform our societies, we may ask what resources would enable human beings – from stakeholders to designers, engineers, regulators, politicians, and general users – to use them constructively to build more humane societies rather than the opposite.

To make progress on those questions, we need to ask what motivates us to do the 'right thing' in the first place and how we can tap into those resources. A humanities perspective (and particularly from one that is humanistic) opens up vistas for understanding humans as embodied, social, and communal beings. We are shaped and motivated by community and by the stories, symbols, values, and practices we share with others, who, in turn, make us who we are. The disciplines of the humanities have much to contribute here since this is also a question about the social, political, psychological, and spiritual conditions (or worldviews) that support and shape human agency. Here, not least, a realistic assessment of the power of technology is vital (Stiegler 2013, 2018). In trying to resource human beings to develop and cultivate a sense of self, community and agency in a technological world, we suggest that we can draw on the resources of many traditions of philosophy, religion, spirituality, and culture. Those traditions can provide us with practical resources to train, attune, and form human beings to refine their desires, thoughts, and feelings (Hoff 2021). Such virtue – grounded on a relational anthropology of human-technology relations – is the basis of any practical notion of human freedom

and morality around which we can organize our liberal, democratic, and plural societies. It is worth noting that this does not deny the value of other ethical approaches – perhaps deontological, utilitarian, and consequentialist – but rather emphasizes the fact that, ultimately, virtue is instrumental to really *do* what we ethically deem good. Thus, we see virtue ethics and the cultivation of "the technomoral self" and "technomoral wisdom" (Vallor 2016; Kanner 1998) – i.e., morally cultivating the self and wisdom under the influence of technology – as a necessary complement to any practical ethical assessment of DL systems. Here, our analysis of the dynamics of a technicized world goes hand in hand with the question of how such dynamics – insofar as they are unwanted – can also be countered. A virtue-oriented approach, for example, may profit from the spiritual traditions of moral sublimation that focus on money, sex, and power as abiding human temptations toward vice as well as realms in which one can behave virtuously. This moral outlook on human beings, their actions, motivations, and freedom from the negative aspects of those perennial temptations yields a perhaps surprisingly rich assessment of the key ethical challenges of DL systems.

Firstly, it is undeniable that *money* drives DL technology as well as societal changes induced by it (Bughin et al. 2017; Stadelmann 2019; Tricot 2021). Developments in the field go hand in hand with marketing hype cycles and cash-grab investments, as well as dramatic variations in stock value. With a focus on the business models, we can also say that economic dynamics and the incentive structures of the advertisement and attention economy, or – more alarmingly put – "surveillance capitalism" (Zuboff 2019, 2023) – already have destructive, destabilizing and dehumanizing effects on our societies. DL catalyzes such developments and forces us to consider how bad incentive structures and the abuse of economic power can be mitigated – and, positively put, how virtue can be cultivated in economics (Bruni and Sugden 2013; Bruni and Héjj 2011)

Secondly, *sex*, which has always been a driver of technological innovation (Keilty 2018) – from the success and broad implementation of VCR, the dot-com boom, online payment systems, e-commerce, internet-based video streaming platforms, live video-chats, and digital hardware (cameras and devices for faster broadband), all the way to high-speed internet on mobile phones, as well as augmented and virtual reality – is a factor in DL applications. One example of where this manifests is novel possibilities of DL-powered GenAI, which allow for the generation of demeaning and pornographic content (e.g., 'nonconsensual deep fake porn') against the will of victims or even without their knowledge. A virtue-oriented perspective on such technology would not focus only on technical solutions (such as filters and constraints), since technological power can always be circumvented or adversarially deployed. It seems timely, therefore, to revive more traditional humanistic and spiritual ways of engaging with 'the human'; through educational formation (in the *Bildung*-tradition) towards rationality, sociality, morality, and care, which must complement technological innovations (Kergel et al. 2022).

Thirdly, it is vital to assess the relationship of technology and *power* (Coeckelbergh 2022; Sattarov 2019). In a sense, technology can be understood as a (more or less controllable) form of power lent to some, while it renders others (and possibly the rest of nature) more powerless with regards to the former (Lewis 1943). In the last few years, we have increasingly seen the application of DL in the political sphere (Crawford 2021; Crawford and Paglen 2021; Kane 2019; Sætra 2021; Marwala 2023). The manipulative potentials of DL systems (Ienca 2023) clearly have the power to substantially impact our 'freedom' as citizens in modern societies – especially through microtargeting, nudging, adaptive preference formation, and manipulating choice architectures of 'persuasive technology' (Bishop 1985; Fogg 2002; Wilson 2017; Zuiderveen Borgesius et al. 2018; Susser 2019; Milano et al. 2020; Susser et al. 2019; Mele et al. 2021; Ashton and Franklin 2022; Simchon et al. 2023; Smith and de Villiers-Botha 2023; Carroll et al. 2023) – which were impressed on the public mind through the 'Cambridge Analytica Scandal' (Berghel 2018). These potentials are further evinced by the channeling and filtering of accessible information and the algorithmically powered platforming or de-platforming of political actors or opinions, and in some countries, even social scoring and controlling systems (see, e.g., Coeckelbergh 2022; Geller 2022; Heinrichs et al. 2022, pp. 144-168). We have already mentioned fears of corporate totalitarianism, which Shoshana Zuboff describes as "a ubiquitous networked institutional regime that records, modifies, and commodifies everyday experience from toasters to bodies, communication to thought, all with a view to establishing new pathways to monetization and profit" (Zuboff 2019, p. 81). There are similar concerns in the sphere of state-sponsored surveillance and totalitarian power through AI systems (and especially DL systems, since such methods power machine perception). These concerns reach beyond the economic motif of profit and into the political sphere of human rights, dignity, and autonomy. While there is no doubt that such technologies stand to impact our political landscape to an almost seismic degree and that we must respond to this challenge (Véliz 2021; Curzon et al. 2021), it is important to examine the assumptions about the human underlying these fears. Are human beings fully "hackable animals" (Coeckelbergh 2022, pp. 85–86) that can be fully manipulated and controlled? Taken literally, such a view would reduce human beings to quantifiable data, which can be manipulated and controlled through engineering. From a holistic view of the human person, the greater danger seems to be that human beings *believe this* and then treat each other *as if* they were reducible to such data and statistical analyses, profiles, and predictions drawn from them – this is bracketing out the fact, that treating human beings in such a way can be both extremely effective and dehumanizing at the same time. Thus, an ethical assessment of the use of DL, for example, in profiling and predicting behavior – which already finds practical application, e.g., in law, insurance, loan-giving, and health care (see, e.g., Lamberton et al. 2017; Berk 2021; Awotunde et al. 2022; Turiel and Aste 2020; Rong et al. 2020; Yang 2022; Secinaro et al. 2021; Vallès-Peris and Domènech 2023; Gill 2023) – would focus on the insight that

such predictions and profiling can never do justice to human beings, their dignity, and freedom as persons and citizens of our societies. This would be an anthropological analysis, backing the ethical objection to the abusive instrumentalization of DL, rather than just an ethical objection that such abuse should not happen. From a virtue-ethics perspective, an assessment of DL could focus on the following questions:

Follow-Up Questions:

- What are the economic, political, and institutional dynamics related to DL? Who benefits? How do DL systems change the power dynamics? Who is in control, and who is being controlled? Which ideas and values are imposed on society by those who are 'in control'? How do we deal with the fact that many of those dynamics are too complex to even be controlled in any meaningful way?
- How are DL systems being used in exploitative ways? How can they be designed and deployed in more constructive, value-based, and goal-oriented ways? Which incentive structures should be created so that the latter is encouraged and not the former?
- How are we to think about 'autonomy' and 'responsibility' given the opacity of current DL applications? How should we conceptualize such values in light of a relational anthropology (seeing human beings and technologies as co-constitutive)? And how can we motivate ourselves (and design technology that really supports us) to create a more humane future? More broadly still, how is DL affecting our self-understanding?
- How could a humane future look like, and how could DL systems help achieve such a future? Which applications, models, use cases, and best practices are there that lead toward human flourishing?

4 Conclusion

We propose that the most promising way of speaking about (and conceptualizing) DL systems is not as a 'standalone' form of 'intelligence' or 'sentience' but as a form of 'complex information processing' that augments human intelligence (Ford et al. 2015). Given that, historically, this description has been rejected – notably by John McCarthy – in favor of an 'artificial intelligence' description, for marketing and funding purposes, and that this has now become entrenched, we suggest amending this prevailing designation, to become not AI but 'extended intelligence' (Uhl 2021; Karachalios and Ito 2018; Council on Extended Intelligence 2021), seeing the need to stress that we understand such extension in terms of enactivism and a relational anthropology as outlined above (section 3.1.*iii*) and not in terms of the 'extended mind theory'. 'Extended intelligence', in our proposal, would analyse and assess DL technologies within a relational framework of human-technology systems. Such systems include both human actors and algorithms embedded in cultural, technological, societal, and other environmental contexts. Such a perspective avoids the

reification and anthropomorphization of AI, without losing sight of these technologies' powerful dynamics, influence on human beings, and their high degree of practical agency. It complements technical practice and optimization with consideration of the 'human factor', i.e., values, judgments, and our political self-determination as free human beings – but it has no naive conception of a contextless 'freedom', considering how existentially enmeshed we are in our technicized environments. Within an extended intelligence framework, we can combine the question of how to make better technology with more fundamental human questions: what do we actually *want*, and how might we realistically get there? In our view, perhaps the most important question here is: what motivates and enables us to act? Given that we do not conceive of ourselves as fully autonomous subjects independent from external influences (cultural, technological, or biological), how could an entangled freedom look like? More broadly, indeed, this is perhaps the most important question posed to the humanities today – and answers to it will have to draw from intellectual, cultural, and spiritual resources (Hoff 2021). Only in light of answers to these questions can we meaningfully assess whether and which technology helps us to get there.

Such an encompassing view can bear upon all stages of technological development and application: in design, practical implementation, and deployment, in assessing its impact, and finally, in reconsidering regulations, further design, and use. We see such thinking being already fruitfully practiced in approaches of human-centered, 'value-based' and 'value sensitive' systems design (Aurum et al. 2005; Friedman et al. 2013; Spiekermann 2015; Spiekermann and Winkler 2022; Spiekermann 2023; Friedman and Hendry 2019; Shneiderman 2022; Herrmann and Pfeiffer 2023).

A realistic assessment of the promise and peril of DL requires an holistic relational anthropology and thus an encompassing view of the human integration of nature, technology and culture. Such a broader perspective can only fully come into view if we address technical issues, such as those within DL, from a perspective integrating engineering, natural sciences, and the humanities. As a cluster of disciplines, the humanities, particularly with their multifaceted approaches, can help address the pertinent questions in the digital transformation. This work program aims to further this engagement.

DL will never yield the sorts of results that could bring us closer to the future we actually *want* if it is not approached in such an encompassing way. Given the urgency such issues have for our societies, it seems pertinent to note here that such an aim must reach beyond the bounds of scholarly methods in either the natural sciences or the humanities. If we want to realize the potential goods of DL systems, we would do well to draw from other (non-technical and even non-academic) resources – from cultural and spiritual practices and traditions – which can transform human motivation toward care and allow the deployment of DL applications for good.

Appendices

A Deep Learning: A More Mathematical Account

In what follows, we outline the fundamental workings of DL by introducing artificial neural networks (ANNs), or more specifically, multi layer perceptrons (MLP), which comprise the core building block of any DL system (section A.1). We then elaborate how they work by means of 'universal approximation' (section A.2). Next, we introduce the training process and analyze a set of inevitable errors (section A.3). The last section (A.4) introduces some open questions in the theory of DL. In contrast to the non-mathematical introduction to DL in the main text (section 2), this section does not provide a dedicated part on generative language models, such as ChatGPT. For this, we refer to section 2.4.

A.1 Artificial Neural Networks

Artificial neural networks (ANN) are the fundamental building blocks of deep learning. In 1958, Frank Rosenblatt introduced their predecessor, the "perceptron" (Rosenblatt 1958). However, Minsky and Papert (1969) discovered a number of critical weaknesses in perceptrons, such that interest waned for the time being. Almost two decades later, Rosenblatt's idea was revived, with the proposal of the multi layer perceptron (along with a reasonably effective training algorithm), which remains the basic structure of an ANN (Rumelhart et al. 1986; for a complete history, see Schmidhuber 2015). To understand the basic principles of DL, one has to grasp how an MLP works.

An MLP consists of input units $\mathbf{x}$, hidden units $\mathbf{h}$ and output units $\mathbf{z}$ that are connected in a sequence of layers (see Figure 5). Note that bold letters represent variables containing multiple elements, i.e., vectors. The number of units in a layer is referred to as the width of that layer. The number of hidden layers in an MLP is referred to as the depth of that MLP.

Put simply, an ANN is a function f that maps an input to an output. Mathematically, we denote $f : \mathcal{X} \mapsto \mathcal{Y}$, with input $\mathbf{x} \in \mathcal{X}$ and output $f(\mathbf{x}) \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and target spaces respectively (examples following). In the process, each hidden unit computes a simple calculation (described and illustrated in Figure 6). Outputs units are computed similarly, but, depending on the target space, are forced to represent elements of the target space.

As an example, f could predict a person's respiratory function (an important health indicator in medicine, e.g., quantified by the forced expiratory volume, FEV) based on certain measurements. The input would consist of a vector $\mathbf{x}$ containing features (the measurements), such as age (x_0), sex (x_1), height (x_2), and whether the patient smokes (x_3). The input space $\mathcal{X}$ would be considered the set of all possible combinations of features. The output $f(\mathbf{x})$

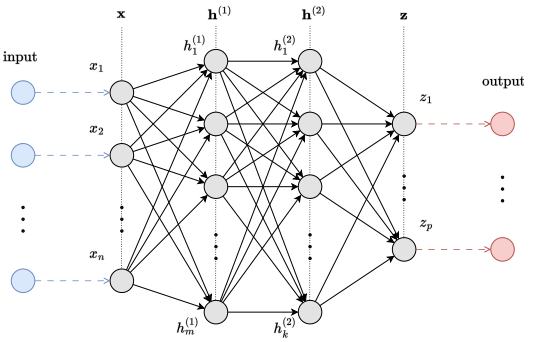


Fig. 5: Illustration of a multi layer perceptron (MLP) with two hidden layers. A hidden unit $h_v^{(u)}$, denotes the vth unit in the uth layer. Note that concerning the notions of the "width" and the "depth" of an MLP, respectively, the depiction is rotated by 90°, such that the mapping from input to output is shown from left to right. This is how ANNs are typically visualized.

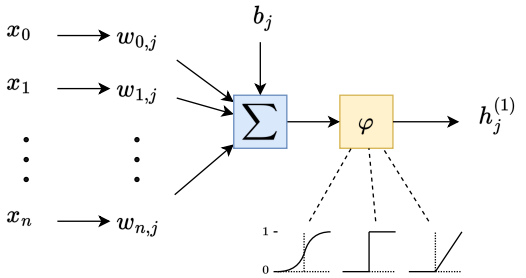


Fig. 6: An artificial neuron, for example the jth unit in the first hidden layer of the multi layer perceptron illustrated in Figure 5. Each hidden unit's output $h_j^{(1)}$ can be computed as a weighted and biased sum of its inputs x_i (blue box), followed by a non-linear activation function φ (yellow box). Three examples of commonly used activation functions are shown below the corresponding box (from left to right: sigmoid function, step function, rectified linear unit). To be exact, every unit $h_j^{(1)}$ is computed as follows: $h_j^{(1)} = \varphi(b_j + \sum_{i=0}^{n} w_{i,j} x_i)$, where $w_{i,j}$ denotes the weight from x_i to h_j and b_j the bias (for each h_j). Intuitively, $h_j^{(1)}$ thus computes a weighted sum of its inputs and passes it on if the sum exceeds some threshold, whereas finding optimal weights and thresholds to achieve the final goal of f is the purpose of the training process.

would consist of one value z representing the FEV. The target space $\mathcal{Y}$ would be considered the set of all possible FEVs.

Alternatively, f could classify images showing handwritten digits into corresponding classes 0 to 9 (a classical problem to e.g., process bank cheques automatically, Lecun et al. 1998). The input $\mathbf{x}$ would contain the gray scale pixel values of a flattened image (going from $w \times w$ pixels to $1 \times w^2$, such that $\mathbf{x}$ is a vector of length w^2), whereas the output $f(\mathbf{x})$ would consist of ten values z_0 to z_9 representing the probabilities of the image for showing the respective digit.

In short, an ANN is a layered network of computationally simple units that encodes a function, which maps any desired input to any desired target space. But how do these straight-forward calculations end up performing complex and meaningful tasks, such as detecting cars in images? In the next section, we address such questions and assess ANNs on a higher level of understanding.

A.2 Universal Approximation

The universal approximation theorem states that every ANN with at least one hidden layer can approximate any continuous function with arbitrary precision (Hornik et al. 1989; Cybenko 1989; the activation function cannot be polynomial, see Leshno et al. 1993). Theoretically, the precision increases with the number of hidden units. However, in practice, it is difficult to achieve a precise approximation by adding width to an ANN. Instead, increasing the depth, i.e., stacking many hidden layers, has been shown to perform far better (Bengio and LeCun 2007). This finding lies at the heart of DL. As the name suggests, "deep" in DL stands for using ANNs with many hidden layers.

Every hidden layer encodes a function itself, such that the function encoded by the ANN consists of a succession of functions. Before taking a closer look at what these successive functions do, it is important to understand that the data dimension at every layer corresponds with the number of units in that layer. For example, if an ANN processes images with pixel count w (remember that in this case the input layer must consists of w units), every image is represented by one point in w-dimensional space (for an illustration, see Figure 7). The function encoded by each layer transforms the space of the antecedent layer (note that every transformation is also a function). The varying width of consecutive layers corresponds with expanding or compressing space to higher and lower dimensions. Non-linear activation functions allow for non-linear transformations of the data space. Thus, if we consider an ANN that classifies images into the classes "shows a car" and "does not show a car", processing an image corresponds to transforming a point in high-dimensional space, until compression to one dimension by the output layer leads to one value that represents the probability of the respective image to show a car. Figure 8 shows how data is transformed on an easy-to-understand low-dimensional ANN, and Figure 9 shows the same effect on a complex task (image classification).

The power of deep ANNs lies in their capacity to approximate high-dimensional and highly non-linear functions by successive data-space transformations combined with the fact that they can be fitted to data, i.e., trained, for specific tasks: the true function underlying the task does not need to be

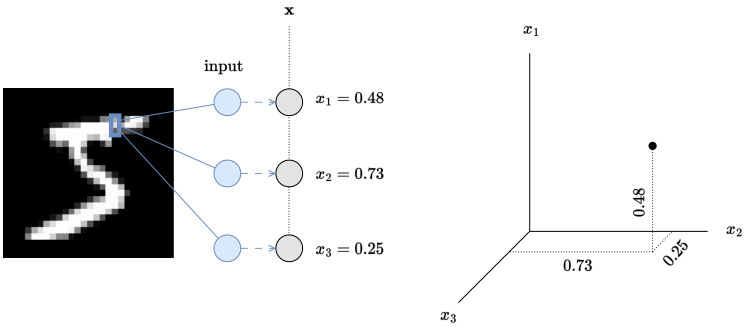


Fig. 7: Representation of an input image by a point in space. On the left is an image showing the handwritten digit '5'. Imagine that only three pixels are fed into the input layer of an ANN (the following layers are not shown), represented by their grey-scale value between $0-1$. On the right, the input units are shown as axes and the unit values as positions on these axes. Thus, one point in three-dimensional space represents the three input pixels. Similarly, all the 784 pixels can be represented (although not visually illustrated) in 784-dimensional space.

known, it suffices to provide enough samples of input-output pairs. As such, they can extract complex patterns and provide human-accessible outputs that represent the underlying patterns in some meaningful way (e.g., the complex patterns underlying images that contain dogs or cats are transformed into two probability values only). But how can an ANN be fitted to data? Or, more precisely: how can an ANN, starting with randomly initialized parameters (weights and biases), approximate a useful function? The next section provides a short insight into the principles of learning in ANNs.

A.3 Training Process and Inevitable Errors

Statistical learning theory provides us with a framework for assessing how ANNs approximate functions and data distributions. It is used to study theoretical properties of learning algorithms from a statistical viewpoint.

Training an ANN can be perceived as an optimization process, in which a loss (or cost) is minimized by altering the underlying model parameters (here, the weights and biases). This requires the definition of a loss function that measures the difference between the output of an ANN and the corresponding target value. The loss function thus measures how well a model performs on a data example. Mathematically speaking, we define a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ which maps the output of a DL model $f(\mathbf{x}) \in \mathcal{Y}$ and its corresponding target $\mathbf{y} \in \mathcal{Y}$ to a real value. For example, if we consider a model f to predict how many days a hospitalized patient has to stay in an ICU based on his or her conditions and demographic data, the loss $\mathcal{L}(f(\mathbf{x}), \mathbf{y})$ would measure the

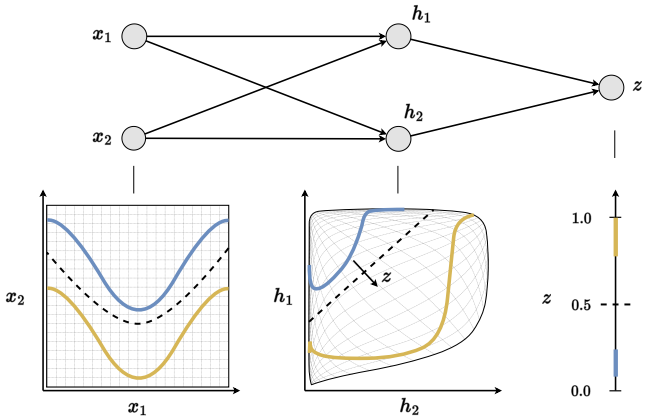


Fig. 8: Illustration of how an ANN transforms data layer by layer, adapted with permission from Olah (2015). In this simple example, the input consists of two dimensions x_1 and x_2. The input space lies on two curves, each belonging to a separate class (class blue and yellow, respectively). Note that one input example would correspond to a point on one of the curves and not to a curve itself. The task of this ANN is to predict, for any given input, to which class it belongs. This is not a linear task, since separation of classes cannot be achieved with a straight line in input space. The hidden layer **h** warps the space, leading to a reshaping of the curves that allows for easier classification, here linear. The output layer z projects the space to a single dimension, representing it on a value between 0 and 1 (corresponding with the probability to belong to class yellow). The decision margin (dashed line) is at 0.5, since all values above 0.5 are considered to belong to class yellow and all values below 0.5 to class blue. To better understand how the decision line separates the input data, its shape is back-projected to all preceding layers.

difference between predicted days and true days in the ICU. The model is optimized in order to minimize that difference.

The goal, however, is not to optimize the model based on the assessment of one individual output, but to optimize it for the whole task, no matter the individual input. This raises the question of how the model can be assessed as a whole. If all possible input-target data pairs were known, i.e., the joint distribution between input and target space $\mathbb{P}_{X,Y}$ (see Figure 10a for an illustration), evaluating and optimizing a model would be straightforward. Obviously, the need for a model would vanish at that point, however, since if we, e.g., had all possible images showing a car, detecting cars in images would no longer be a matter of prediction, but one of simple comparison with known data. Nevertheless, the theoretical case, where we know the data distribution $\mathbb{P}_{X,Y}$ is important in learning theory, because it enables us to formulate the theoretically best model f^*. Computing f^* requires the minimization of the loss between the data pair we expect, knowing $\mathbb{P}_{X,Y}$. This

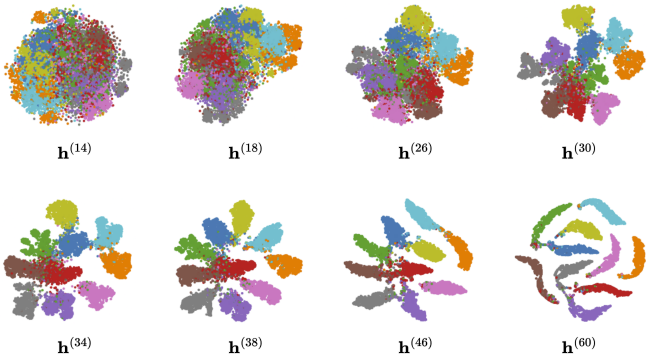


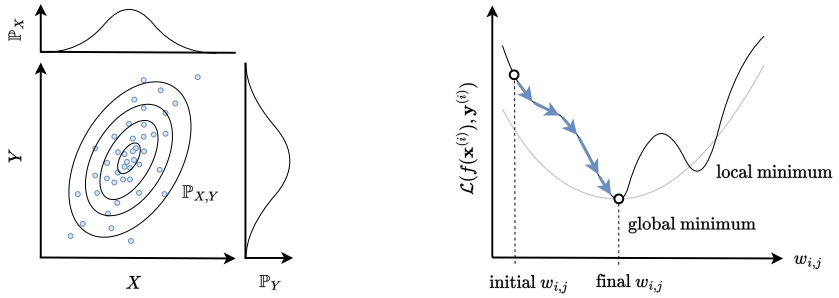
Fig. 9: Effect of the data transformations within an ANN that classifies images into ten classes, such as "plane", "car", "bird", "cat" etc. Every point in the plots corresponds with one image. Colors represent to which class that image belongs, i.e., its target label. To visualize a multidimensional point (an image) in two dimensions, an algorithm was used that produces a low-dimensional representation, where distances between points are reflective of the distances between the original, i.e., multidimensional, points. Looking at the data representation at different hidden layers $\mathbf{h}^{(14)}$ to $\mathbf{h}^{(60)}$, one can see that the data is transformed in a manner that allows for easier separation of classes. The plots are taken from Hoyt and Owen (2021) and are obtained from real data.

loss is called the Bayes risk $\mathcal{R}(f)$. Mathematically, we formulate Bayes risk $\mathcal{R}(f) = \mathbb{E}_{x,y\sim\mathbb{P}_{X,Y}}[\mathcal{L}(f(\mathbf{x}), \mathbf{y})]$ and formulate the best model in theory by a minimization of Bayes risk $f^* = \arg\min_f \mathcal{R}(f)$. Although f^* is the best model in theory, its success is still limited by the Bayes error, referring to a lack of correlation between X and Y. If we, e.g., predict the duration of stay at the ICU based on shoe size, even the best model will not succeed, since the two variables are not correlated in a meaningful way.

Of course, in practice, we only have some examples of data pairs and do not know the joint distribution $\mathbb{P}_{X,Y}$ of these pairs. The solution to this problem is to use a finite set of example pairs $\mathcal{D}$ (called training set) drawn from, and thus (hopefully) representing, the underlying joint distribution. However, the training set can represent the joint distribution only if the number of sampled data pairs is sufficiently large and the sampling is done in an independent and identically distributed (i.i.d.) fashion. If we, e.g., predict ICU days for patients of all ages with a model that is based on a training set $\mathcal{D}$ containing only patients older than 80 years (which is drawn from a subset of $\mathbb{P}_{X,Y}$), the prediction will likely be inaccurate. Note that since $\mathbb{P}_{X,Y}$ is unknown, there is no way to be certain whether these assumptions hold (and they typically do not hold in practice, but are assumed to anyway, with astonishing success regarding the results).

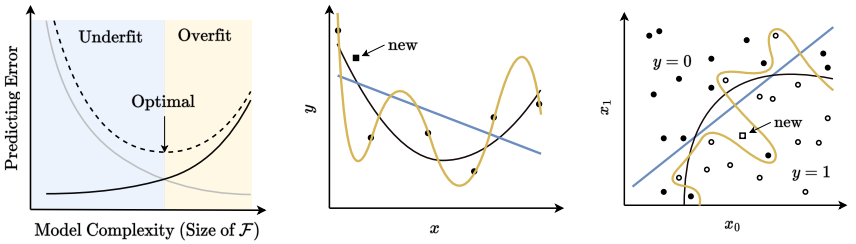
Besides relying on a finite set of example pairs, there is another limiting factor. In practice, the set of functions that can be approximated by a model f is restricted due to the number of parameters and choices of concrete ANN architecture (the architecture of an ANN is defined by the number and arrangement of individual units plus their wiring). The family of functions that can be realized by f is called its hypothesis space $\mathcal{F}$. Thus, instead of minimizing the Bayes risk, the primary goal of machine learning is the minimization of an empirical risk $\mathcal{R}_{\mathcal{D}}(f)$ that is limited by $\mathcal{D}$ and restricted to a hypothesis space $\mathcal{F}$. The best model in practice, $\hat{f}$, is then defined as the function for which the empirical risk is minimal. Mathematically, we formulate the empirical risk $\mathcal{R}_{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=0}^{n}\mathcal{L}(f(\mathbf{x}^{(i)}), \mathbf{y}^{(i)})$ and the best model in practice $\hat{f} = \arg\min_{f\in\mathcal{F}}\mathcal{R}_{\mathcal{D}}(f)$.

In addition to the Bayes error, then, there is also an error caused by dependence on a finite set of examples and an error caused by dependence on a finite set of functions. The former is called the estimation error and the latter the approximation error. The different effects of these two errors on the predictive performance of a model is shown with more detail in Figure 11.



(a) (b)

Fig. 10: **(a)** Illustration of the joint distribution of data pairs. Example input-target data pairs are shown as blue dots, where X represents the input variable and Y the target variable. Their respective underlying distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ as well as the iso-probability lines of the joint distribution $\mathbb{P}_{X,Y}$ are shown as well (all unknown in practice). Note that here, the Bayes error would be high, i.e., knowing X gives relatively little information about Y. **(b)** Shows the loss for one example i with respect to one model parameter $w_{i,j}$ (for a non-convex optimization problem in black and for a convex problem in gray). From its initial value, the parameter is being nudged towards a lower loss in an iterative fashion, eventually reaching a minimum. Note that for non-convex optimizations, there exist multiple minima, and thus, it is almost certain for $w_{i,j}$ to miss the global minimum on convergence.



(a) (b) (c)

Fig. 11: Illustration of the effects of the estimation error and the approximation error. **(a)** Shows the estimation error (black line), the approximation error (gray line), and their sum, the generalization error (dashed line) as a function of model complexity (size of the hypothesis space $\mathcal{F}$). For low model complexities, the model underfits, i.e., the model is too simple to fit the data appropriately (high approximation error). For high model complexities, the model overfits, i.e., the model is too complex and fits the training data too closely, while failing to generalize to data outside the training set (high estimation error). For optimal generalization, the model complexity must usually be balanced between these errors; these balance is found by "regularizing" the models, i.e., giving them side objectives that make bother overfitting and underfitting harder. **(b)** Illustrates the different errors on a regression task, i.e., predicting a real-valued y based on x. An underfitting function (blue) approximates the training examples (black dots) poorly. An overfitting function (yellow) approximates the training examples too closely, failing to generalize to unseen examples (black square). The optimal function is shown in black.**(c)** Illustrates the different errors on a classification task (integer-valued y), i.e., predicting whether an example $\mathbf{x}$ belongs to class $y = 0$ (black dots) or class $y = 1$ (white dots). An underfitting decision margin (blue) separates the training examples poorly. An overfitting decision margin (yellow) follows the observed training examples too closely, failing to generalize to unseen examples appropriately (see white square). The optimal decision margin is shown in black.

Finding $\hat{f}$ from a random starting point f is, as is stated above, an optimization task that involves altering the model parameters $w_{i,j}$ and b_j such that the empirical risk $\mathcal{R}_{\mathcal{D}}(f)$ is minimal. This is done using gradient-based methods (Lecun et al. 1998), such as stochastic gradient descent (SGD). Without going into too much detail, these methods compute the gradient of the loss function with respect to the model parameters. The gradient provides information about how each parameter needs to be adjusted in order to reduce the loss (or risk). Note that the gradient does not provide information about the optimal final value for each parameter (the loss topology with regard to the model parameters is not known and is typically arbitrarily complex, see, e.g., Feizi et al. 2018), but about whether each parameter should be higher or lower. The model parameters can then be nudged in a manner that decreases

the loss. This procedure is repeated many times, moving the parameters in small steps towards a lower and lower loss, eventually arriving at a minimum. However, there is no guarantee that the minimum to which the optimization converges corresponds with the global minimum (the lowest possible risk, see Figure 10b). Thus, the performance of the model that is actually achieved by optimization is likely still below the performance of the best model in practice $\hat{f}$, but typically close enough for practical usability (Choromanska et al. 2015). The gap between those two model is called the optimization error. In sum, we have the following four kinds of commonly known errors or limits that are inherent in statistical learning methods (such as DL):

i. Bayes Error

The Bayes error is the unquantified error occurring due to a lack of correlation between input and target variables. If the input is only loosely related to the target, even the best model in theory f^* performs badly. Every statistical method can only provide good results if its variables are sufficiently related.

ii. Approximation Error

The approximation error is the error due to the restriction on the complexity of the model. Since no model is arbitrarily complex, no model can map arbitrarily complex relations. This error is also called the 'bias' of a model as it is due to a model being biased towards a simplistic solution that is systematically off the true target because of lacking model complexity.

iii. Estimation Error

The estimation error is the error due to the restriction of the joint probability $\mathbb{P}_{X,Y}$ to a finite set of data (the training set) that should represent it. Since there exists no machine learning tasks where all possible data is accessible, the error due to lack of data is inevitable. This error is also called the "variance" of a model as with varying training data, models with different blind spots would be learned, corresponding to different weaknesses.

iv. Optimization Error

The optimization error is the difference between the optimal function in practice $\hat{f}$ and the actual model that was learned. It is due to a suboptimal search process for the best model parameters.

In sum: To minimize the overall error, the model complexity should increase with the complexity of the true underlying input-output relation, and training examples must be representative of it. Since this underlying relation remains unknown, however, there cannot be any guarantee that the trained DL model will not be wildly wrong with new examples (Delétang et al. 2023). Almost the opposite is true: for any statistical classifier, including complex DL models, examples can be generated where it fails dramatically – these are referred

to as 'adversarial examples' (Szegedy et al. 2014; Goodfellow et al. 2015). This is an inevitable characteristic of DL models (Shafahi et al. 2020; Papernot et al. 2017) and poses problems in various applications, such as self-driving cars (Brown et al. 2018; Tu et al. 2022), making the presence of additional processes to detect such out-of-distribution samples necessary (Amirian et al. 2018). While theoretical guarantees are thus absent, however, the success of DL models is built on the empirical finding that, in practice, reasonable generalization to unseen examples usually works quite well if it can be achieved through interpolation between seen training examples (see section A.4).

A.4 Our Shallow Understanding of Why DL Works

Although advances in hardware and the increasing availability of data explain the success of DL to a large extent (see, e.g., Lenzen 2020) and gave rise to numerous algorithmic advances that account for another large part (Stadelmann et al. 2019b), a unified theory that fully justifies the remarkable performance of DL models is still missing (Plebe and Grasso 2019; Sejnowski 2020). To list a few open issues (references follow in the subsections below), it is still not fully clear (*i*) why large models generalize so well despite their vast overparametrization, (*ii*) why deep models are so superior to shallow ones, (*iii*) why models perform well in very high-dimensional environments, and (*iv*) why optimization converges to good local minima despite the non-convexity of the loss topology. For some of these issues, there exist already plausible explanations that are increasingly supported by evidence. We discuss these issues here because this way we gain further interesting insight into how DL work. Note that all addressed issues are not clearly separable, and a solution to one is likely to contribute to the resolution of other issues as well. Here, we only provide a brief outline of these issues. We refer to Plebe and Grasso (2019) or Hodas and Stinis (2018) for a more detailed overview, to Poggio et al. (2019) for a more mathematical approach, and to Berner et al. (2022) for a recent in-depth mathematical investigation.

i. DL Generalizes Surprisingly Well

To interpolate any training data with example count m and example dimension d (e.g., $m = 100$ cases of ICU patients with $d = 10$ vital parameters per case or $m = 5000$ images with a number of $d = 784$ pixels each), a three-layered MLP with a parameter count in the order of $m + d$ (mathematically, this is expressed as $\mathcal{O}(m + d)$) is sufficient (Berner et al. 2022). This means that in theory, increasing the number of parameters beyond $\mathcal{O}(m + d)$ (called overparametrization) will lead to overfitting. However, it was shown that DL models can perfectly fit randomly labeled training data and still generalize well to real labels (Zhang et al. 2017). This means that even models that are proven to overfit can generalize well to unseen data. This behavior seemingly contradicts our previous elaborations in section A.3. In fact, overparametrization was shown to often lead to significant benefits (Soltanolkotabi et al. 2019). It was,

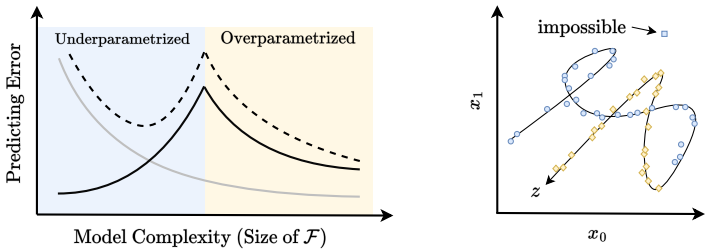
e.g., shown that a DL model with 1.6 million parameters, trained on $m = 50000$ images of size $d = 32 \times 32$ pixels reaches almost state-of-the-art generalization performance (Zhang et al. 2017, 2021). Another example of outstanding generalization performance despite overparametrization is "Noisy Student", a state-of-the-art image recognition system with 480 million parameters, trained on 1.2 million images (Xie et al. 2020).

Current research into this matter focuses on the optimization algorithms, such as SGD (see Figure 10b), suggesting that they exhibit properties of implicit regularization, i.e., a bias that prefers models of low complexity (Soudry et al. 2022; Arora et al. 2019). Furthermore, the probability for optimization to converge to good minima (there seems to be a multiplicity of them in deep models, see section A.4.*iii*) turns out to be higher in deep overparametrized models, i.e., they usually exhibit a smaller optimization error than underparametrized models (Poggio et al. 2019). Analysis of overparametrized models extends the classical U-shaped generalization error in Figure 11a to a "double descent" curve (see Figure 12a). This suggests that if model complexity is increased beyond the point of high estimation error and low approximation error, the estimation error descents again, leading to a lower overall error eventually (see, e.g., Belkin et al. 2019; Feldman 2021; Dosovitskiy et al. 2021). Furthermore, it was observed that the correlation (more precisely the mutual information) between neighboring layers is high, meaning that the hypothesis space collapses to a smaller actual size, implicitly regularizing the optimization task (Hodas and Stinis 2018; Tishby and Zaslavsky 2015).

ii. DL Overcomes the 'Curse of Dimensionality'

Many tasks in computer science become extremely difficult whenever the number of dimensions of the data is high. High dimensional data is typically problematic in that the possible number of configurations of a data example increases *exponentially* with its dimensions. The number of examples required to cover all relevant configurations therefore also increases exponentially. In computer science, this problem is referred to as the curse of dimensionality (Bellman 2015; Novak and Woźniakowski 2009; Goodfellow et al. 2016). Unfortunately, in most DL applications, the data dimension is very high. If we, e.g., want to classify 512×512 pixel gray scale images with each pixel representing a value between 0 and 255, there are 256^{262144} possible images for which we want to compute the probability to belong to certain classes. The data and computing power theoretically needed to solve this task is far beyond reach. However, in practice, high-dimensional tasks have been successfully solved for many applications using deep learning.

An important idea underlying machine learning is that all meaningful data lies on a lower-dimensional manifold embedded in higher-dimensional space. A manifold is a connected region as shown in Figure 12b. An illustrative example, given by Goodfellow et al. (2016), is that although we live in three-dimensional space, we essentially move on a two-dimensional manifold, i.e., the surface of the world. Thus, standing at a random location usually excludes being above



(a) (b)

Fig. 12: **(a)** The double descend curve. In direct comparison to Figure 11a, we see that with increasing model complexity, the generalization error (dashed line) decreases a second time for overparametrized models. **(b)** Illustration of a learned manifold. A data set is shown where every two-dimensional example is concentrated on a one-dimensional manifold (i.e., with the curvature of the manifold z known, the position of a point on the manifold can be given with just one scalar number). It is not possible for an example of that data set to be far from the manifold (e.g., a data set of images showing handwritten digits does not contain examples that show a car). The two-dimensional input space can be transformed such that an output unit z represents the disentangled manifold (figuratively, the curve is stretched out to form a straight line). Separating the two classes (blue and yellow) based on $\mathbf{x}$ is difficult, but separating them based on z is straight forward (just by finding a single threshold on that stretched-out straight line).

or below ground. Driving a car, the probability collapses even further, reducing the problem to one dimension, i.e., all roads. Likewise, the set of all possible images that, e.g., show a face, is far below the set of all possible images. As a consequence, machine learning tasks simplify drastically. Although the manifold hypothesis is not necessarily correct for all problems, there is a lot of evidence confirming it (Goodfellow et al. 2016; Brahma et al. 2016).

iii. Depth is Superior to Width

The universal approximation theorem states that a shallow (or deep) ANN can, in theory, approximate every function if its width is sufficiently large (see section A.3). For a long time, therefore, the focus of analysis was on the width of models. However, the breakthrough for ANNs (and with that, a new Spring for AI in general) came with algorithms that allowed for efficient training of deep models (Hinton et al. 2006, followed by, e.g., Bengio et al. 2006; Ranzato et al. 2006). The focus shifted from width to depth (Bengio 2009), as DL outperformed other machine learning methods on important tasks (Krizhevsky et al. 2012). Today, models can contain over 1000 layers (He et al. 2016). Interestingly, we still do not have a tangible explanation for the strength of deep over shallow models.

Recent research, however, has shed some light on this issue. It was shown, for instance, that the function complexity of deep models grows exponentially with depth, and that shallow models need significantly more parameters to exhibit similar complexity (Eldan and Shamir 2016; Raghu et al. 2017; Berner et al. 2022; Lin et al. 2017). Furthermore, not all parameters are equally important (Frankle and Carbin 2019). The model output has been shown to be more sensitive to lower layer parameters (Raghu et al. 2017). These findings affirm a widely supported theory that data representation must gradually progress through the layers from rudimentary to more complex (Hodas and Stinis 2018; Shwartz-Ziv and Tishby 2017). Figure 13 shows this effect in an image classifier. For many researchers, the benefit of this compositional structure of functions makes intuitive (and quantitative) sense (Hodas and Stinis 2018; Mhaskar et al. 2017; Lee et al. 2009). A further hypothesis is that by using many narrow layers (i.e., low-dimensional functions) instead of a few wide layers (i.e., high-dimensional functions), models can better avoid the curse of dimensionality (Poggio et al. 2019; for curse of dimensionality, see section A.4.*ii*).

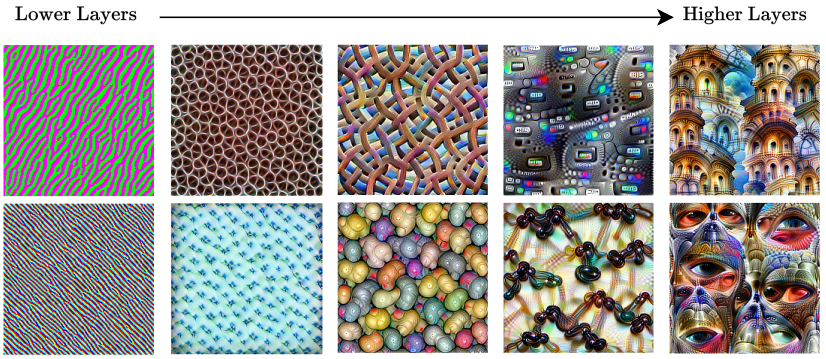


Fig. 13: Progression of data representation in the convolutional neural network (CNN) "GoogleLeNet" (Szegedy et al. 2015) (used with permission from Olah et al. 2017). These images were achieved by fixing the trained model parameters and instead optimize the pixel values of the input image in such a manner that maximizes certain hidden channels (in CNNs, convolutional layers usually consist of channels, which consist of units). Thus, the obtained images show what the respective channels are detecting, i.e., how the respective channels represent input images. Going from lower to higher layers, we see that channels represent edges, then textures, then patterns, then parts, then objects, such as archways and eyes. Remembering section A.3, we can see that the image representations in higher layers serve to simplify classification. e.g., detecting cars based on raw pixel values or edges is hard, but detecting cars based on channels that represent objects, such as tires, lights, and streets (and ultimately cars themselves) is much easier.

iii. Training Reaches Near-Global Minima Despite Non-Convex Optimization
In order to be guaranteed to reach the global minimum with SGD, the loss topology has to be convex. This is typically not the case in the optimization of deep models (see, e.g., Elbrächter et al. (2019); Petersen et al. (2021)). With a non-convex loss topology, the issues with optimization include converging at suboptimal local minima (see Figure 10b and, e.g., Auer et al. 1995; Safran and Shamir 2018), converging at saddle points, and high time consumption due to escaping very small gradients (see, e.g., Berner et al. (2022)). Even today, there is little theoretical explanation for why, in practice, convergence to a quasi-global minimum seems to be observed quite often.

One insight is that by adding more depth and width to a model the quality of local minima has been shown to improve, i.e., they are no worse than global minima (Kawaguchi et al. 2019). Thus, for sufficiently large models, the focus shifts from finding the minimum loss, i.e., the global minimum, to finding some low loss, i.e., a good minimum (Goodfellow et al. 2016).

Further reading: This brief introduction into the theory of DL only scratches the surface of the varieties of concepts, models, and applications. For a more in-depth account, we suggest reading Goodfellow et al. (2016); Prince (2023); Murphy (2022).

Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability: This manuscript has no associated data.

References

Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable Artificial Intelligence (XAI). IEEE Access 6:52,138–52,160. https://doi.org/10.1109/ACCESS.2018.2870052

Agüera y Arcas B (2022) Do Large Language Models Understand Us? Daedalus 151(2):183–197. https://doi.org/10.1162/daed_a_01909

Amirian M, Schwenker F, Stadelmann T (2018) Trace and detect adversarial attacks on cnns using feature response maps. In: Pancioni L, Schwenker F, Trentin E (eds) Artificial Neural Networks in Pattern Recognition. Springer International Publishing, pp 346–358, https://doi.org/10.1007/978-3-319-99978-4_27

Amirian M, Füchslin RM, Herzig I, Hotz PE, Lichtensteiger L, Montoya-Zegarra JA, et al. (2023) Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks. Medical Physics https://doi.org/10.1002/mp.16405

Andler D (2009) Philosophy of cognitive science. In: French Studies in the Philosophy of Science: Contemporary Research in France. Springer, p 255–300

Antweiler C (2012) Inclusive Humanism: Anthropological Basics for a Realistic Cosmopolitanism. Vandenhoeck & Ruprecht

Antweiler C (2013) Pan-cultural universals. a fundament for an inclusive humanism. In: Rüsen J (ed) Approaching Humankind. Towards an Intercultural Humanism. Vandenhoeck & Ruprecht, p 37–68

Arora S, Cohen N, Hu W, Luo Y (2019) Implicit regularization in deep matrix factorization. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc.

Aroyo AM, de Bruyne J, Dheu O, Fosch-Villaronga E, Gudkov A, Hoch H, et al. (2021) Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. Paladyn, Journal of Behavioral Robotics 12(1):423–436. https://doi.org/doi:10.1515/pjbr-2021-0029

Ashton H, Franklin M (2022) The problem of behaviour and preference manipulation in AI systems. In: Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022), CEUR Workshop Proceedings, URL https://discovery.ucl.ac.uk/id/eprint/10146136

Auer P, Herbster M, Warmuth MKK (1995) Exponentially many local minima for single neurons. In: Touretzky D, Mozer M, Hasselmo M (eds) Advances in Neural Information Processing Systems, vol 8. MIT Press

Aurum A, Biffl S, Boehm B, Erdogmus H, Grünbacher P (2005) Value-Based Software Engineering. Springer

Awotunde JB, Misra S, Ayeni F, Maskeliunas R, Damasevicius R (2022) Artificial Intelligence based system for bank loan fraud prediction. In: Abraham A, Siarry P, Piuri V, Gandhi N, Casalino G, Castillo O, et al. (eds) Hybrid Intelligent Systems. Springer International Publishing, pp 463–472, https://doi.org/10.1007/978-3-030-96305-7_43

Barrat J (2015) Our Final Invention: Artificial Intelligence and the End of the Human Era. St. Martin's Publishing Group

Baum SD (2018) Reconciliation between factions focused on near-term and long-term Artificial Intelligence. AI & Society 33(4):565–572. https://doi.org/10.1007/s00146-017-0734-3

Baumann J, Heitz C (2022) Group fairness in prediction-based decision making: From moral assessment to implementation. In: 2022 9th Swiss

Conference on Data Science (SDS), IEEE, pp 19–25

Bélisle-Pipon JC, Monteferrante E, Roy MC, Couture V (2022) Artificial Intelligence ethics has a black box problem. AI & Society pp 1–16. https://doi.org/10.1007/s00146-021-01380-0

Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences of the United States of America 116(32):15,849–15,854. https://doi.org/10.1073/pnas.1903070116

Bellman RE (2015) Adaptive Control Processes. Princeton University Press

Bender EM, 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. Association for Computational Linguistics, pp 5185–5198, https://doi.org/10.18653/v1/2020.acl-main.463

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, FAccT '21, p 610–623, https://doi.org/10.1145/3442188.3445922

Bengio Y (2009) Learning Deep Architectures for AI. Now Publishers Inc

Bengio Y, LeCun Y (2007) Scaling learning algorithms toward AI. In: Bottou L, Chapelle O, DeCoste D, Weston J (eds) Large-Scale Kernel Machines. MIT Press, https://doi.org/10.7551/mitpress/7496.001.0001

Bengio Y, Lamblin P, Popovici D, Larochelle H (2006) Greedy layer-wise training of deep networks. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in Neural Information Processing Systems, vol 19. MIT Press, URL https://proceedings.neurips.cc/paper_files/paper/2006/file/5da713a690c067105aeb2fae32403405-Paper.pdf

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. https://doi.org/10.1109/TPAMI.2013.50

Bengio Y, et al. (2023) Pause giant AI experiments: An open letter. Future of Life Institute Open Letter, https://futureoflife.org/open-letter/pause-giant-ai-experiments

Bennett MR, Hacker PMS (2022) Philosophical Foundations of Neuroscience. John Wiley & Sons

Berghel H (2018) Malice domestic: The cambridge analytica dystopia. Computer 51(5):84–89. https://doi.org/10.1109/MC.2018.2381135

Berk RA (2021) Artificial Intelligence, predictive policing, and risk assessment for law enforcement. Annual Review of Criminology 4(1):209–237. https://doi.org/10.1146/annurev-criminol-051520-012342

Berner J, Grohs P, Kutyniok G, Petersen P (2022) The modern mathematics of Deep Learning. In: Grohs P, Kutyniok G (eds) Mathematical Aspects of Deep Learning. Cambridge University Press, p 1–111, https://doi.org/10.1017/9781009025096.002

Besold TR, Uckelman SL (2018) The what, the why, and the how of artificial explanations in automated decision-making. Preprint, https://doi.org/10.48550/arXiv.1808.07074

Bishop J (1985) Elster, j.: "sour grapes: Studies in the subversion of rationality". Australasian Journal of Philosophy 63(n/a):245

Bisk Y, Holtzman A, Thomason J, Andreas J, Bengio Y, Chai J, et al. (2020) Experience grounds language. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp 8718–8735, https://doi.org/10.18653/v1/2020.emnlp-main.703

Black D (2014) Embodiment and Mechanisation: Reciprocal Understandings of Body and Machine from the Renaissance to the Present. Ashgate Press

Boden MA (1988) Computer Models of Mind: Computational Approaches in Theoretical Psychology. Cambridge University Press

Boden MA (2008) Mind as Machine: A History of Cognitive Science. Oxford University Press

Boden MA (2016) AI: Its Nature and Future. Oxford University Press

Bogert E, Schecter A, Watson RT (2021) Humans rely more on algorithms than social influence as a task becomes more difficult. Scientific Reports 11(1):8028. https://doi.org/10.1038/s41598-021-87480-9

Borji A (2023) Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. Preprint, https://doi.org/10.48550/arXiv.2210.00586

Bostrom N (2013) Existential risk prevention as global priority. Global Policy 4(1):15–31

Bostrom N (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press

Brahma PP, Wu D, She Y (2016) Why Deep Learning Works: A Manifold Disentanglement Perspective. IEEE Transactions on Neural Networks and Learning Systems 27(10):1997–2008. https://doi.org/10.1109/TNNLS.2015.2496947

Braidotti R (2013) The Posthuman. Polity Press

Bray D (2011) Wetware: A Computer in Every Living Cell. Yale University Press

Brey P (2005) Artifacts as social agents. In: Harbers H (ed) Inside the Politics of Technology: Agency and Normativity in the Co-production of Technology and Society. Amsterdam University Press, p 61–84, URL http://www.jstor.org/stable/j.ctt45kcv7.6

Brodie ML (2019) What is data science? In: Braschler M, Stadelmann T, Stockingers K (eds) Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, p 101–130, https://doi.org/10.1007/978-3-030-11821-1_8

Brooks R (2017) The seven deadly sins of predicting the future of AI. URL https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai

Brown TB, Mané D, Roy A, Abadi M, Gilmer J (2018) Adversarial patch. Preprint, https://doi.org/10.48550/arXiv.1712.09665

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. (2018) The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation. Preprint, https://doi.org/10.48550/arXiv.1802.07228

Bruni L, Héjj T (2011) The economy of communion. In: Handbook of Spirituality and Business. Springer, p 378–386, https://doi.org/10.1057/9780230321458_45

Bruni L, Sugden R (2013) Reclaiming virtue ethics for economics. Journal of Economic Perspectives 27(4):141–164. https://doi.org/10.1257/jep.27.4.141

Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4. Preprint, https://doi.org/10.48550/arXiv.2303.12712

Bughin J, Hazan E, Ramaswamy S, Chui M, Allas T, Dahlstrom P, et al. (2017) Artificial Intelligence: the next digital frontier? McKinsey Global Institute

Butlin P, Long R, Elmoznino E, Bengio Y, Birch J, Constant A, et al. (2023) Consciousness in Artificial Intelligence: Insights from the science of consciousness. Preprint, https://doi.org/10.48550/arXiv.2308.08708

Cali DD (2017) Mapping Media Ecology. Peter Lang Verlag, https://doi.org/10.3726/978-1-4539-1871-5

Calvo RA, Peters D, Vold K, Ryan RM (2020) Supporting human autonomy in AI systems: A framework for ethical enquiry. In: Burr C, Floridi L (eds) Ethics of Digital Well-Being: A Multidisciplinary Approach. Springer International Publishing, p 31–54, https://doi.org/10.1007/978-3-030-50585-1_2

Campolo A, Crawford K (2020) Enchanted Determinism: Power without Responsibility in Artificial Intelligence. Engaging Science, Technology, and Society 6:1–19. https://doi.org/10.17351/ests2020.277

Cappuccio ML (2017) Mind-upload. the ultimate challenge to the embodied mind theory. Phenomenology and the Cognitive Sciences 16:425–448. https://doi.org/10.1007/s11097-016-9464-0

Carroll M, Chan A, Ashton H, Krueger D (2023) Characterizing manipulation from AI systems. Preprint, https://doi.org/10.48550/arXiv.2303.09387

Caruana R, Lundberg S, Ribeiro MT, Nori H, Jenkins S (2020) Intelligible and explainable machine learning: Best practices and practical challenges. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, KDD '20, p 3511–3512, https://doi.org/10.1145/3394486.3406707

Cave S, Coughlan K, Dihal K (2019) "Scary robots": Examining public responses to AI. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, AIES '19, p 331–337, https://doi.org/10.1145/3306618.3314232

Cave S, Dihal K, Dillon S (2020) AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford University Press

Chalmers DJ (2010) The singularity: A philosophical analysis. Journal of Consciousness Studies 17(9-10):9–10

Chalmers DJ (2011) A computational foundation for the study of cognition. Journal of Cognitive Science 12(4):325–359. https://doi.org/10.17791/jcs.2011.12.4.325

Chatila R, Dignum V, Fisher M, Giannotti F, Morik K, Russell S, et al. (2021) Trustworthy AI. In: Braunschweig B, Ghallab M (eds) Reflections on Artificial Intelligence for Humanity. Springer International Publishing, p 13–39,

https://doi.org/10.1007/978-3-030-69128-8_2

Chollet F (2019) On the measure of intelligence. Preprint, https://doi.org/10.48550/arXiv.1911.01547

Choromanska A, Henaff M, Mathieu M, Arous GB, LeCun Y (2015) The loss surfaces of multilayer networks. Preprint, https://doi.org/https://doi.org/10.48550/arXiv.1412.0233

Churchland PS (2013) Touching a Nerve: The Self as Brain. W. W. Norton & Company

Churchland PS, Sejnowski TJ (1992) The Computational Brain. MIT Press

Clark A (2008) Pressing the flesh: A tension in the study of the embodied, embedded mind? Philosophy and Phenomenological Research 76(1):37–59. https://doi.org/10.1111/j.1933-1592.2007.00114.x

Clark A (2013) Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36(3):181–204. https://doi.org/10.1017/S0140525X12000477

Clark A (2016) Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press

Cobbe J, Lee MSA, Singh J (2021) Reviewable automated decision-making: A framework for accountable algorithmic systems. Preprint, https://doi.org/10.48550/arXiv.2102.04201, 2102.04201

Coeckelbergh M (2020) AI Ethics. MIT Press

Coeckelbergh M (2022) The Political Philosophy of AI: An Introduction. John Wiley & Sons

Coenen C, Grunwald A (2017) Responsible research and innovation (rri) in quantum technology. Ethics and Information Technology 19:277–294. https://doi.org/10.1007/s10676-017-9432-6

Confalonieri R, Coba L, Wagner B, Besold TR (2021) A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11(1):e1391. https://doi.org/10.1002/widm.1391

Council on Extended Intelligence (2021) Our vision. URL https://globalcxi.org/vision/

Courchamp F, Mizrahi L, Morin C, Courchamp F, Bernard J, Lambert O (2018) Eine überschätzte Spezies. URL https://www.arte.tv/de/videos/

RC-014177/eine-ueberschaetzte-spezies/

Crawford K (2021) The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press

Crawford K, Paglen T (2021) Excavating AI: The politics of images in machine learning training sets. AI & Society 36:1399. https://doi.org/10.1007/s00146-021-01301-1

Crolic C, Thomaz F, Hadi R, Stephen AT (2022) Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. Journal of Marketing 86(1):132–148. https://doi.org/10.1177/00222429211045687

Crutzen PJ, Stoermer EF (2013) The anthropocene [2000]. In: Robin L, Sörlin S, Warde P (eds) The Future of Nature. Yale University Press, p 479–490, https://doi.org/10.12987/9780300188479-041

Curzon J, Kosa TA, Akalu R, El-Khatib K (2021) Privacy and Artificial Intelligence. IEEE Transactions on Artificial Intelligence 2(2):96–108. https://doi.org/10.1109/TAI.2021.3088084

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2(4):303–314. https://doi.org/10.1007/BF02551274

Damasio A (2010) Self Comes to Mind: Constructing the Conscious Brain. Goodreads

Darling K (2017) "Who's Johnny?" Anthropomorphic Framing in Human–Robot Interaction, Integration, and Policy. In: Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence. Oxford University Press, https://doi.org/10.1093/oso/9780190652951.003.0012

Davison A (2021) Machine learning and theological traditions of analogy. Modern Theology 37(2):254–274. https://doi.org/https://doi.org/10.1111/moth.12682

Delétang G, Ruoss A, Grau-Moya J, Genewein T, Wenliang LK, Catt E, et al. (2023) Neural networks and the chomsky hierarchy. Preprint, https://doi.org/10.48550/arXiv.2207.02098

Dennett DC (1989) The Intentional Stance. MIT Press

Dennett DC (1991) Consciousness Explained. Penguin Books

Dennett DC (2007) Philosophy as naive anthropology: Comment on bennett and hacker. In: Bennett M, Dennett DC, Hacker PMS, Searle JR (eds) Neuroscience and Philosophy: Brain, Mind, and Language. Columbia University

Press, p 73–96, URL http://www.jstor.org/stable/10.7312/benn14044

Dennett DC (2013) Intuition Pumps and Other Tools for Thinking. WW Norton & Company

Di Paolo E, Buhrmann T, Barandiaran X (2017) Sensorimotor Life: An Enactive Proposal. Oxford University Press

Domingos P (2012) A few useful things to know about machine learning. Communications of the ACM 55(10):78–87. https://doi.org/10.1145/2347736.2347755

Donepudi PK (2017) Machine Learning and Artificial Intelligence in Banking. Engineering International 5(2):83–86. https://doi.org/10.18034/ei.v5i2.490

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=YicbFdNTTy

Dosovitsky G, Bunge EL (2021) Bonding with bot: User feedback on a chatbot for social isolation. Frontiers in Digital Health 3:735,053. https://doi.org/10.3389/fdgth.2021.735053

Došilović FK, Brčić M, Hlupić N, Došilović FK, Brčić M, Hlupić N (2018) Explainable Artificial Intelligence: A survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp 210–215, https://doi.org/10.23919/MIPRO.2018.8400040

Dreyfus H, Taylor C (2015) Retrieving Realism. Harvard University Press

Dreyfus HL, Wrathall MA (2014) Skillful Coping: Essays on the phenomenology of everyday perception and action. Oxford University Press, https://doi.org/10.1093/acprof:oso/9780199654703.001.0001

Dubber MD, Pasquale F, Das S (2020) The Oxford Handbook of Ethics of AI. Oxford University Press, https://doi.org/10.1093/oxfordhb/9780190067397.001.0001

Dürr O (2021) Homo Novus: Vollendlichkeit im Zeitalter des Transhumanismus. No. 108 in Studia Oecumenica Friburgensia, Aschendorff Verlag

Dürr O (2023) Transhumanismus – Traum oder Alptraum? Herder

Dürr O, Segessenmann J, Steinmann J (2023) Meaning, form, and the limits of natural language processing, upublished manuscript under review

Durt C, Froese T, Fuchs T (2023) Against AI understanding and sentience: Large language models, meaning, and the patterns of human language use. Preprint, URL http://philsci-archive.pitt.edu/21983/

Durán JM, Formanek N (2018) Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds & Machines 28(4):645–666. https://doi.org/10.1007/s11023-018-9481-6

Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. (2023) "so what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management 71:102,642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eden A, Steinhart E, Pearce D, Moor J (2012) Singularity hypotheses: An overview. In: Eden A, Pearce D, Moor J, Søraker J, Steinhart E (eds) Singularity Hypotheses. The Frontiers Collection, Springer, p 1–12, https://doi.org/10.1007/978-3-642-32560-1_1

Edwards D, Edwarts H (2018) Google's engineers say that "magic spells" are ruining AI research. Quartz URL https://qz.com/1274131/googles-engineers-say-that-lack-of-rigor-is-ruining-ai-research/

Eisenstein M, et al. (2021) Artificial Intelligence powers protein-folding predictions. Nature 599(7886):706–708. https://doi.org/10.1038/d41586-021-03499-y

Elbrächter DM, Berner J, Grohs P (2019) How degenerate is the parametrization of neural networks with the ReLU activation function? In: Advances in Neural Information Processing Systems, vol 32. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/hash/04115ec378e476c56d19d827bcf8db56-Abstract.html

Eldan R, Shamir O (2016) The power of depth for feedforward neural networks. Preprint, https://doi.org/10.48550/arXiv.1512.03965

Ellul J (2021 [1954]) The Technological Society. Vintage

Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychological Review 114(4):864. https://doi.org/10.1037/0033-295X.114.4.864

European Parliament (2017) REPORT with recommendations to the Commission on Civil Law Rules on Robotics. European Parliament, URL https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html

European Parliament, Council of the European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. URL https://data.europa.eu/eli/reg/2016/679/oj

Feizi S, Javadi H, Zhang J, Tse D (2018) Porcupine neural networks: Approximating neural network landscapes. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2018/file/b6cda17abb967ed28ec9610137aa45f7-Paper.pdf

Feldman V (2021) Does learning require memorization? A short tale about a long tail. Preprint, https://doi.org/10.48550/arXiv.1906.05271

Felt U, Fouché R, Miller CA, Smith-Doerr L (2017) The Handbook of Science and Technology Studies, 4th edn. MIT Press

Flessner B (2018) Die Rückkehr der Magier: Die KI als Lapis philosophorum des 21. Jahrhunderts. In: Die Rückkehr der Magier: Die KI als Lapis philosophorum des 21. Jahrhunderts. Transcript Verlag, p 63–106, https://doi.org/10.1515/9783839442876-003

Floridi L (2019) Establishing the rules for building trustworthy AI. Nature Machine Intelligence 1(6):261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi L (2020) AI and Its New Winter: from Myths to Realities. Philosophy & Technology 33(1):1–3. https://doi.org/10.1007/s13347-020-00396-6

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. (2018) An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds & Machines 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Flynn T (2002) A secular humanist definition setting the record straight. Free Inquiry

Fodor JA (1975) The Language of Thought. Harvard University Press

Fogg BJ (2002) Persuasive technology: Using computers to change what we think and do. Ubiquity https://doi.org/10.1145/764008.763957

Ford KM, Hayes PJ, Glymour C, Allen J (2015) Cognitive orthoses: Toward human-centered AI. AI Magazine 36(4):5–8. https://doi.org/10.1609/aimag.v36i4.2629

Ford M (2018) Architects of Intelligence: The truth about AI from the people building it. Packt Publishing Ltd

Foucault M (1990) Les mots et les choses. Gallimard Paris

Frank M (2002) Self-consciousness and self-knowledge: On some difficulties with the reduction of subjectivity. Constellations 9(3):390–408. https://doi.org/10.1111/cons.2002.9.issue-3

Frank M (2007) Non-objectal subjectivity. Journal of Consciousness Studies 14(5-6):152–173

Frankle J, Carbin M (2019) The lottery ticket hypothesis: Finding sparse, trainable neural networks. Preprint, https://doi.org/10.48550/arXiv.1803.03635

Friedman B, Hendry DG (2019) Value Sensitive Design: Shaping Technology With Moral Imagination. MIT Press

Friedman B, Kahn PH, Borning A, Huldtgren A (2013) Value sensitive design and information systems. In: Doorn N, Schuurbiers D, van de Poel I, Gorman ME (eds) Early engagement and new technologies: Opening up the laboratory. Springer Netherlands, p 55–95, https://doi.org/10.1007/978-94-007-7844-3_4

Fuchs T (2018) Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind. Oxford University Press

Fuchs T (2020) The circularity of the embodied mind. Frontiers in Psychology 11. https://doi.org/10.3389/fpsyg.2020.01707

Fuchs T (2021) In Defence of the Human Being: Foundational Questions of an Embodied Anthropology. Oxford University Press

Fuchs T (2022) Understanding sophia? on human interaction with artificial agents. Phenomenology and the Cognitive Sciences https://doi.org/10.1007/s11097-022-09848-0

Gallagher S (2011) Interpretations of embodied cognition. In: Tschacher W, Bergomi C (eds) The Implications of Embodiment: Cognition and Communication. Imprint Academic, p 59–70

Gallagher S (2017) Enactivist Interventions: Rethinking the Mind. Oxford University Press, https://doi.org/10.1093/oso/9780198794325.001.0001

Gallagher S (2018) The extended mind: State of the question. Southern Journal of Philosophy 56(4):421–447. https://doi.org/10.1111/sjp.12308

Geller A (2022) Social Scoring durch Staaten. PhD thesis, Ludwig-Maximilians-Universität, München

Gill KS (2023) Seeing beyond the lens of platonic embodiment. AI & Society 38(4):1261–1266. https://doi.org/10.1007/s00146-023-01711-3

Glüge S, Amirian M, Flumini D, Stadelmann T (2020) How (not) to measure bias in face recognition networks. In: Schilling FP, Stadelmann T (eds) Artificial Neural Networks in Pattern Recognition. Springer International Publishing, pp 125–137, https://doi.org/10.1007/978-3-030-58309-5_10

Goldman A, Beddor B (2021) Reliabilist Epistemology. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Summer 2021 edn. Metaphysics Research Lab, Stanford University

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press, URL www.deeplearningbook.org

Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. Preprint, https://doi.org/10.48550/arXiv.1412.6572

Gordon JS, Pasvenskiene A (2021) Human rights for robots? a literature review. AI and Ethics 1(4):579–591. https://doi.org/10.1007/s43681-021-00050-7

de Graaf MM, Hindriks FA, Hindriks KV (2021) Who wants to grant robots rights? In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, HRI '21 Companion, p 38–46, https://doi.org/10.1145/3434074.3446911

Greaves H, MacAskill W (2021) The case for strong longtermism. Tech. rep., Global Priorities Institute, University of Oxford

Grey C, Dürr O (2023) On changing the subject: Secularity, religion, and the idea of the human. Religions 14(4). https://doi.org/10.3390/rel14040466

Grimm S (2021) Understanding. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Summer 2021 edn. Metaphysics Research Lab, Stanford University

Grunwald A (2007) Converging technologies: Visions, increased contingencies of the conditio humana, and search for orientation. Futures 39(4):380–392. https://doi.org/https://doi.org/10.1016/j.futures.2006.08.001

Grunwald A (2009) Technology assessment: Concepts and methods. In: Meijers A (ed) Philosophy of Technology and Engineering Sciences. Handbook of the Philosophy of Science, North-Holland, p 1103–1146, https://doi.org/https:

//doi.org/10.1016/B978-0-444-51667-1.50044-6

Grunwald A (ed) (2016) The Hermeneutic Side of Responsible Research and Innovation. John Wiley & Sons, https://doi.org/https://doi.org/10.1002/9781119340898

Grunwald A (2019a) Technology Assessment in Practice and Theory. Routledge

Grunwald A (2019b) The inherently democratic nature of technology assessment. Science and Public Policy 46(5):702–709. https://doi.org/10.1093/scipol/scz023

Grunwald A, Nordmann A, Sand M (eds) (2023) Hermeneutics, History, and Technology: The Call of the Future. Routledge, https://doi.org/10.4324/9781003322290

Gunkel DJ (2018) Robot rights. MIT Press

Hagner M (1997) Homo Cerebralis: Der Wandel vom Seelenorgan zum Gehirn. Suhrkamp

Hardré PL (2016) When, how, and why do we trust technology too much? In: Tettegah SY, Espelage DL (eds) Emotions, Technology, and Behaviors. Emotions and Technology, Academic Press, p 85–106, https://doi.org/10.1016/B978-0-12-801873-6.00005-4

Haring KS, Mougenot C, Ono F, Watanabe K (2014) Cultural differences in perception and attitude towards robots. International Journal of Affective Engineering 13(3):149–157. https://doi.org/10.1007/s12369-022-00920-y

Haslam N (2006) Dehumanization: An integrative review. Personality and Social Psychology Review 10(3):252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-inspired Artificial Intelligence. Neuron 95(2):245–258. https://doi.org/10.1016/j.neuron.2017.06.011

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778, https://doi.org/10.1109/CVPR.2016.90

Hebb DO (1949) The Organization of Behavior; A Neuropsychological Theory. Wiley

Heidegger M (1996 [1926]) Being and Time. Suny Press

Heidenreich F, Weber-Stein F (2022) The Politics of Digital Pharmacology: Exploring the Craft of Collective Care. Transcript Verlag

Heil J (2020) Philosophy of Mind: A Contemporary Introduction, 4th edn. Routledge

Heinrichs B, Heinrichs JH, Rüther M (2022) Künstliche Intelligenz. De Gruyter, https://doi.org/10.1515/9783110746433

Herbrechter S (2009) Posthumanismus: Eine kritische Einführung. WBG

Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. (2015) Teaching Machines to Read and Comprehend. In: Advances in Neural Information Processing Systems, vol 28. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html

Herrmann T, Pfeiffer S (2023) Keeping the organization in the loop: a socio-technical extension of human-centered Artificial Intelligence. AI & Society 38(4):1523–1542. https://doi.org/10.1007/s00146-022-01391-5

Hinton GE, Osindero S, Teh YW (2006) A Fast Learning Algorithm for Deep Belief Nets. Neural Computation 18(7):1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527, URL https://doi.org/10.1162/neco.2006.18.7.1527

Hodas NO, Stinis P (2018) Doing the Impossible: Why Neural Networks Can Be Trained at All. Frontiers in Psychology 9. https://doi.org/10.3389/fpsyg.2018.01185

Hoff J (2021) Verteidigung des Heiligen: Anthropologie der digitalen Transformation. Herder

Hohwy J (2013) The Predictive Mind. Oxford University Press

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Networks 2(5):359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hoyt CR, Owen AB (2021) Probing neural networks with t-sne, class-specific projections and a guided tour. Preprint, https://doi.org/10.48550/arXiv.2107.12547

Huerta MF, Koslow SH, Leshner AI (1993) The human brain project: An international resource. Trends in Neurosciences 16(11):436–438. https://doi.org/https://doi.org/10.1016/0166-2236(93)90069-X

Hughes TP (1987) The evolution of large technological systems. In: Bijker W, Hughes T, Pinch T (eds) The Social Construction of Technological Systems:

New Directions in the Sociology and History of Technology. MIT Press, p 51–82

Hutson M (2018) Has Artificial Intelligence become alchemy? Science 360(6388):478–478. https://doi.org/10.1126/science.360.6388.478, publisher: American Association for the Advancement of Science

Hutto DD, Myin E (2012) Radicalizing Enactivism: Basic Minds Without Content. MIT Press

Ienca M (2023) On Artificial Intelligence and manipulation. Topoi 42:833–842. https://doi.org/10.1007/s11245-023-09940-3

Ihde D (1990) Technology and the Lifeworld: From Garden to Earth. Indiana University Press

Ihde D (1995) Postphenomenology: Essays in the Postmodern Context. Northwestern University Press

Janich P (2009) Kein neues Menschenbild: Zur Sprache der Hirnforschung. Suhrkamp Verlag

Jank M (2014) Der homme machine des 21. Jahrhunderts: Von lebendigen Maschinen im 18. Jahrhundert zur humanoiden Robotik der Gegenwart. Brill Fink, https://doi.org/https://doi.org/10.30965/9783846756577

Jenkins R, Hammond K, Spurlock S, Gilpin L (2023) Separating facts and evaluation: motivation, account, and learnings from a novel approach to evaluating the human impacts of machine learning. AI & Society 38:1415–1428. https://doi.org/10.1007/s00146-022-01417-y

Johansen J, Pedersen T, Johansen C (2023) Studying human-to-computer bias transference. AI & Society 38:1659–1683. https://doi.org/10.1007/s00146-021-01328-4

Jonas H (1966) The Phenomenon of Life. Toward a Philosophical Biology. Harper & Row

Joshi G, Walambe R, Kotecha K (2021) A review on explainability in multimodal deep neural nets. IEEE Access 9:59,800–59,821. https://doi.org/10.1109/ACCESS.2021.3070212

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. (2021) Highly accurate protein structure prediction with alphafold. Nature 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

Kane TB (2019) Artificial Intelligence in politics: Establishing ethics. IEEE Technology and Society Magazine 38(1):72–80. https://doi.org/10.1109/

MTS.2019.2894474

Kanner AD (1998) Technological wisdom. ReVision 20(4):45–46

Kaplan M (2022) After Google chatbot becomes 'sentient,' MIT prof says Alexa could too. New York Post URL https://nypost.com/2022/06/13/mit-prof-says-alexa-could-become-sentient-like-google-chatbot/

Karachalios K, Ito J (2018) Human intelligence and autonomy in the era of 'extended intelligence'. Council on Extended Intelligence URL https://globalcxi.org/wp-content/uploads/CXI_Essay.pdf

Karanasiou AP, Pinotsis DA (2017) A study into the layers of automated decision-making: Emergent normative and legal aspects of Deep Learning. International Review of Law, Computers & Technology 31(2):170–187. https://doi.org/10.1080/13600869.2017.1298499

Katz DM, Bommarito MJ, Gao S, Arredondo P (2023) GPT-4 passes the bar exam. Preprint, https://doi.org/10.2139/ssrn.4389233

Kaun A (2022) Suing the algorithm: the mundanization of automated decision-making in public services through litigation. Information, Communication & Society 25(14):2046–2062. https://doi.org/10.1080/1369118X.2021.1924827

Kaur D, Uslu S, Rittichier KJ, Durresi A (2022) Trustworthy Artificial Intelligence: A review. ACM Computing Surveys 55(2). https://doi.org/10.1145/3491209

Kawaguchi K, Huang J, Kaelbling LP (2019) Effect of Depth and Width on Local Minima in Deep Learning. Neural Computation 31(7):1462–1498. https://doi.org/10.1162/neco_a_01195

Keilty P (2018) Desire by design: pornography as technology industry. Porn Studies 5(3):338–342. https://doi.org/10.1080/23268743.2018.1483208

Kenny A (1984) The Legacy of Wittgenstein. Oxford University Press

Kergel D, Paulsen M, Garsdal J, Heidkamp-Kergel B (eds) (2022) Bildung in the Digital Age. Routledge

Kitchin R, Dodge M (2011) Code/Space: Software and everyday life. Software studies. MIT Press, https://doi.org/10.7551/mitpress/9780262042482.001.0001

Kostopoulos L (2021) Decoupling Human Characteristics from Algorithmic Capabilities. Tech. rep., IEEE Standards Association, URL https://standards.{IEEE}.org/initiatives/artificial-intelligence-systems/decoupling-human-characteristics/

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges C, Bottou L, Weinberger K (eds) Advances in Neural Information Processing Systems, vol 25. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kroes P, Verbeek PP (2014) Introduction: The moral status of technical artefacts. In: Kroes P, Verbeek PP (eds) The Moral Status of Technical Artefacts. Springer Netherlands, p 1–9, https://doi.org/10.1007/978-94-007-7914-3_1

Krüger S, Wilson C (2023) The problem with trust: on the discursive commodification of trust in AI. AI & Society pp 1753—-1761. https://doi.org/10.1007/s00146-022-01401-6

Kuljian OR, Hohman ZP (2023) Warmth, competence, and subtle dehumanization: Comparing clustering patterns of warmth and competence with animalistic and mechanistic dehumanization. British Journal of Social Psychology 62(1):181–196. https://doi.org/10.1111/bjso.12565

Kurzweil R (2005) The Singularity Is Near: When Humans Transcend Biology. Penguin Publishing Group

Lamberton C, Brigo D, Hoy D (2017) Impact of robotics, rpa and AI on the insurance industry: Challenges and opportunities. Journal of Financial Perspectives 4(1). URL https://ssrn.com/abstract=3079495

Latour B (2012) We Have Never Been Modern. Harvard university press

LeCun Y (2022) A path towards autonomous machine intelligence. Preprint, URL https://openreview.net/pdf?id=BZ5a1r-kVsf

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

LeCun Y, Bengio Y, Hinton G (2015) Deep Learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning. Association for Computing Machinery, ICML '09, p 609–616, https://doi.org/10.1145/1553374.1553453

Legg S, Hutter M (2007) A collection of definitions of intelligence. In: Proceedings of the 2007 Conference on Advances in Artificial General Intelligence:

Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006. IOS Press, p 17–24

Lemoine B (2022) Is LaMDA sentient? An interview. Medium URL https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Lenzen M (2020) Künstliche Intelligenz: Fakten, Chancen, Risiken. C.H. Beck

Leroi-Gourhan A (1993) Gesture and Speech. MIT Press

Leshno M, Lin VY, Pinkus A, Schocken S (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6(6):861–867. https://doi.org/10.1016/S0893-6080(05)80131-5

Leung KH (2019) The Picture of Artificial Intelligence and the Secularization of Thought. Political Theology 20(6):457–471. https://doi.org/10.1080/1462317X.2019.1605725

Lewis CS (1943) The Abolition of Man. Oxford University Press

Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. (2023) Trustworthy AI: From principles to practices. ACM Computing Surveys 55(9). https://doi.org/10.1145/3555803

Li M, Leidner B, Castano E (2014) Toward a comprehensive taxonomy of dehumanization: Integrating two senses of humanness, mind perception theory, and stereotype content model. TPM: Testing, Psychometrics, Methodology in Applied Psychology 21(3):285—-300

Liggieri K, Müller O (eds) (2019) Mensch-Maschine-Interaktion: Handbuch Zu Geschichte - Kultur - Ethik. J.B. Metzler

Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G (2020) Backpropagation and the brain. Nature Reviews Neuroscience 21(6):335–346. https://doi.org/10.1038/s41583-020-0277-3

Lin HW, Tegmark M, Rolnick D (2017) Why does deep and cheap learning work so well? Journal of Statistical Physics 168(6):1223–1247. https://doi.org/10.1007/s10955-017-1836-5

Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Lipton ZC, Steinhardt J (2019) Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public

and stymie future research. Queue 17(1):45–77. https://doi.org/10.1145/3317287.3328534

Lipton ZC, Azizzadenesheli K, Kumar A, Li L, Gao J, Deng L (2018) Combating reinforcement learning's sisyphean curse with intrinsic fear. Preprint, https://doi.org/10.48550/arXiv.1611.01211

Liu X, Xie L, Wang Y, Zou J, Xiong J, Ying Z, et al. (2021) Privacy and security issues in Deep Learning: A survey. IEEE Access 9:4566–4593. https://doi.org/10.1109/ACCESS.2020.3045078

Liu Z, Kitouni O, Nolte N, Michaud EJ, Tegmark M, Williams M (2022) Towards understanding grokking: An effective theory of representation learning. Preprint, https://doi.org/10.48550/arXiv.2205.10343

Loi M, Heitz C, Ferrario A, Schmid A, Christen M (2019) Towards an ethical code for data-based business. In: 6th Swiss Conference on Data Science (SDS), pp 6–12, https://doi.org/10.1109/SDS.2019.00-15

Ma Y, Tsao D, Shumm HY (2022) On the principles of parsimony and self-consistency for the emergence of intelligence. Frontiers of Information Technology & Electronic Engineering 23(9):1298–1323. https://doi.org/10.1631/FITEE.2200297

Madsen A, Reddy S, Chandar S (2022) Post-hoc interpretability for neural nlp: A survey. ACM Computing Surveys 55(8). https://doi.org/10.1145/3546577

Man K, Damasio A (2019) Homeostasis and soft robotics in the design of feeling machines. Nature Machine Intelligence 1(10):446–452. https://doi.org/10.1038/s42256-019-0103-7

Marcus G (2018) Deep Learning: A critical appraisal. Preprint, https://doi.org/10.48550/arXiv.1801.00631

Marcus G, Leivada E, Murphy E (2023) A sentence is worth a thousand pictures: Can large language models understand human language? Preprint, https://doi.org/10.48550/arXiv.2308.00109

Margolis E, Samuels R, Stich SP (2012) The Oxford Handbook of Philosophy of Cognitive Science. Oxford University Press

Martinetz J, Martinetz T (2023) Highly over-parameterized classifiers generalize since bad solutions are rare. Preprint, https://doi.org/10.48550/arXiv.2211.03570

Martini M (2019) Blackbox Algorithmus: Grundfragen einer Regulierung Künstlicher Intelligenz. Springer, https://doi.org/10.1007/

978-3-662-59010-2

Marwala T (2023) Artificial Intelligence in politics. In: Artificial Intelligence, Game Theory and Mechanism Design in Politics. Springer Nature Singapore, p 41–58, https://doi.org/10.1007/978-981-99-5103-1_4

Matsuo Y, LeCun Y, Sahani M, Precup D, Silver D, Sugiyama M, et al. (2022) Deep Learning, reinforcement learning, and world models. Neural Networks 152(C):267–275. https://doi.org/10.1016/j.neunet.2022.03.037

Mazzone M, Elgammal A (2019) Art, creativity, and the potential of Artificial Intelligence. Arts 8(1). https://doi.org/10.3390/arts8010026

McCarthy J, Minsky ML, Rochester N, Shannon CE (1955) A proposal for the dartmouth summer research project on Artificial Intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, URL http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics 5(4):115–133. https://doi.org/10.1007/BF02478259

McDaniel J, Pease K (2021) Predictive Policing and Artificial Intelligence. Routledge

McLuhan M (1994 [1964]) Understanding Media. The Extensions of Man. MIT Press

Mele C, Russo Spena T, Kaartemo V, Marzullo ML (2021) Smart nudging: How cognitive technologies enable choice architectures for value co-creation. Journal of Business Research 129:949–960. https://doi.org/https://doi.org/10.1016/j.jbusres.2020.09.004

Merleau-Ponty M (1964) The child's relation with others. In: Edie JM (ed) The Primacy of Perception. Northwestern University Press, p 96–155

Merleau-Ponty M, Smith C (1962) Phenomenology of Perception. Routledge

Mhaskar H, Liao Q, Poggio T (2017) When and why are deep networks better than shallow ones? Proceedings of the AAAI Conference on Artificial Intelligence 31(1). https://doi.org/10.1609/aaai.v31i1.10913

Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. AI & Society 35:957–967. https://doi.org/10.1007/s00146-020-00950-y

Miller B (2021) Is technology value-neutral? Science, Technology, & Human Values 46(1):53–80. https://doi.org/10.1177/01622439199009

Minsky M, Papert SA (1969) Perceptrons: An Introduction to Computational Geometry. MIT Press

Mitchell T (1997) Machine Learning. McGraw Hill, URL https://www.cs.cmu.edu/~tom/mlbook.html

Morozov E (2014) To save everything, click here. J Inf Policy

Moyal-Sharrock D (2021) Certainty in Action: Wittgenstein on Language, Mind and Epistemology. Bloomsbury Publishing

Mozer MC (1994) Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. Connection Science 6(2-3):247–280. https://doi.org/10.1080/09540099408915726

Müller O, Liggieri K (2019) Mensch-Maschine-Interaktion seit der Antike: Imaginationsräume, Narrationen und Selbstverständnisdiskurse. In: Liggieri K, Müller O (eds) Mensch-Maschine-Interaktion: Handbuch zu Geschichte, Kultur, Ethik. J.B. Metzler, p 3–14

Munn N, Weijers D (2023) Corporate responsibility for the termination of digital friends. AI & Society 38(4):1501–1502. https://doi.org/10.1007/s00146-021-01276-z

Murphie A, Potts J (2017) Culture and Technology. Bloomsbury Publishing

Murphy KP (2022) Probabilistic Machine Learning: An Introduction. MIT Press, URL probml.ai

Nagel T (1989) The View From Nowhere. Oxford University Press

Nguyen AM, Yosinski J, Clune J (2015) Innovation Engines: Automated Creativity and Improved Stochastic Optimization via Deep Learning. In: Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation. Association for Computing Machinery, GECCO '15, pp 959–966, https://doi.org/10.1145/2739480.2754703

Noble R, Noble D (2023) Understanding Living Systems. Cambridge University Press

Noë A (2023) The Entanglement: How Art and Philosophy Make Us What We Are. Princeton University Press

Notovich A, Chalutz-Ben Gal H, Ben-Gal I (2023) Explainable Artificial Intelligence (XAI): Motivation, terminology, and taxonomy. In: Rokach L, Maimon O, Shmueli E (eds) Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook. Springer International

Publishing, p 971–985, https://doi.org/10.1007/978-3-031-24628-9_41

Novak E, Woźniakowski H (2009) Approximation of infinitely differentiable multivariate functions is intractable. Journal of Complexity 25(4):398–404. https://doi.org/10.1016/j.jco.2008.11.002

Novelli C (2023) Legal personhood for the integration of AI systems in the social context: a study hypothesis. AI & Society 38(4):1347–1359. https://doi.org/10.1007/s00146-021-01384-w

Núñez R, Allen M, Gao R, Miller Rigoli C, Relaford-Doyle J, Semenuks A (2019) What happened to cognitive science? Nature Human Behaviour 3(8):782–791. https://doi.org/10.1038/s41562-019-0626-2

Odling-Smee FJ, Lala KN, Feldman M (2003) Niche Construction: The Neglected Process in Evolution. Princeton University Press

Olah C (2015) Visualizing representations: Deep Learning and human beings. URL https://colah.github.io/posts/2015-01-Visualizing-Representations/

Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill https://doi.org/10.23915/distill.00007

OpenAI (2023) GPT-4 technical report. Preprint, https://doi.org/10.48550/arXiv.2303.08774

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A (2017) Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Association for Computing Machinery, ASIA CCS '17, p 506–519, https://doi.org/10.1145/3052973.3053009

Pavlick E (2023) Symbols and grounding in large language models. Philosophical Transactions A Math Phys Eng Sci 381(2251):20220,041. https://doi.org/10.1098/rsta.2022.0041

Petersen P, Raslan M, Voigtlaender F (2021) Topological Properties of the Set of Functions Generated by Neural Networks of Fixed Size. Foundations of Computational Mathematics 21(2):375–444. https://doi.org/10.1007/s10208-020-09461-0

Pflanzer M, Dubljević V, Bauer WA, Orcutt D, List G, Singh MP (2023) Embedding AI in society: ethics, policy, governance, and impacts. AI & Society 38:1267—1271. https://doi.org/10.1007/s00146-023-01704-2

Piantadosi ST, Hill F (2022) Meaning without reference in large language models. Preprint, https://doi.org/10.48550/arXiv.2208.02957

Pitt D (2022) Mental Representation. In: Zalta EN, Nodelman U (eds) The Stanford Encyclopedia of Philosophy, Fall 2022 edn. Metaphysics Research Lab, Stanford University

Pitt JC (2014) "Guns don't kill, people kill": Values in and/or around technologies. In: Kroes P, Verbeek PP (eds) The Moral Status of Technical Artefacts. Springer Netherlands, p 89–101, https://doi.org/10.1007/978-94-007-7914-3_6

Plebe A, Grasso G (2019) The Unbearable Shallow Understanding of Deep Learning. Minds & Machines 29(4):515–553. https://doi.org/10.1007/s11023-019-09512-8

Poggio T, Banburski A, Liao Q (2019) Theoretical issues in deep networks: Approximation, optimization and generalization. Preprint, https://doi.org/10.48550/arXiv.1908.09375

Polanyi M (1967) The Tacit Dimension: Michael Polanyi. Routledge & Kegan Paul

Postman N (2006) Media ecology education. Explorations in Media Ecology 5(1):5–14. https://doi.org/https://doi.org/10.1386/eme.5.1.5_1

Prabhakaran V, Mitchell M, Gebru T, Gabriel I (2022) A human rights-based approach to responsible AI. Preprint, https://doi.org/10.48550/arXiv.2210.02667

Prescott TJ, Camilleri D (2019) The synthetic psychology of the self. In: Aldinhas Ferreira MI, Silva Sequeira J, Ventura R (eds) Cognitive Architectures. Springer International Publishing, p 85–104, https://doi.org/10.1007/978-3-319-97550-4_7

Prince SJ (2023) Understanding Deep Learning. MIT Press, URL www.udlbook.github.io/udlbook/

Proudfoot D (2011) Anthropomorphism and AI: Turing's much misunderstood imitation game. Artificial Intelligence 175(5):950–957. https://doi.org/https://doi.org/10.1016/j.artint.2011.01.006, special Review Issue

Prunkl C (2022) Human autonomy in the age of Artificial Intelligence. Nature Machine Intelligence 4(2):99–101. https://doi.org/10.1038/s42256-022-00449-9

Putnam H (1960) Minds & machines. In: Hook S (ed) Dimensions of Mind. Collier Books, p 138–164

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. URL https://openai.com/research/language-unsupervised

Raghu M, Poole B, Kleinberg J, Ganguli S, Dickstein JS (2017) On the expressive power of deep neural networks. In: Proceedings of the 34th International Conference on Machine Learning. JMLR.org, ICML'17, p 2847–2854, https://doi.org/10.5555/3305890.3305975

Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. Preprint, https://doi.org/10.48550/arXiv.2204.06125

Ranzato Ma, Poultney C, Chopra S, Cun Y (2006) Efficient learning of sparse representations with an energy-based model. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in Neural Information Processing Systems, vol 19. MIT Press, URL https://proceedings.neurips.cc/paper_files/paper/2006/file/87f4d79e36d68c3031ccf6c55e9bbd39-Paper.pdf

Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A, Barth-Maron G, et al. (2022) A generalist agent. Preprint, https://doi.org/10.48550/arXiv.2205.06175

Reizinger P, Szemenyei M (2020) Attention-Based Curiosity-Driven Exploration in Deep Reinforcement Learning. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 3542–3546, https://doi.org/10.1109/ICASSP40776.2020.9054546

Rescorla M (2020) The Computational Theory of Mind. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Fall 2020 edn. Metaphysics Research Lab, Stanford University

Reutlinger A, Saatsi J (eds) (2018) Explanation Beyond Causation: Philosophical Perspectives on Non-causal Explanations. Oxford University Press

Roberts DA, Yaida S, Hanin B (2022) The Principles of Deep Learning Theory. Cambridge University Press

Robertson J (2014) Human rights vs. robot rights: Forecasts from japan. Critical Asian Studies 46(4):571–598. https://doi.org/10.1080/14672715.2014.960707

Robertson J (2018) Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation. University of California Press

Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. (2022) Tackling climate change with machine learning. ACM Computing

Surveys 55(2). https://doi.org/10.1145/3485128

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. Preprint, https://doi.org/10.48550/arXiv.2112.1075

Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M (2020) Artificial Intelligence in healthcare: Review and prediction case studies. Engineering 6(3):291–301. https://doi.org/https://doi.org/10.1016/j.eng.2019.08.015

Rosenberger R, Verbeek P (eds) (2015) Postphenomenological Investigations: Essays on Human-Technology Relations. Lexington Books

Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65(6):386–408. https://doi.org/10.1037/h0042519

Rouse J (2019) Articulating the World: Conceptual Understanding and the Scientific Image. University of Chicago Press

Rovetto RJ (2023) The ethics of conceptual, ontological, semantic and knowledge modeling. AI & Society https://doi.org/10.1007/s00146-022-01563-3

Rowlands M (2009) Enactivism and the extended mind. Topoi 28:53–62. https://doi.org/10.1007/s11245-008-9046-z

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0

Russell S (2019) Human Compatible: Artificial Intelligence and the Problem of Control. Penguin Books

Russell S, Norvig P (2021) Artificial Intelligence: A Modern Approach, Global Edition. Pearson Education

Ryberg J, Roberts JV (2022) Sentencing and Artificial Intelligence. Oxford University Press

Sætra HS (2021) A typology of AI applications in politics. In: Visvizi A, Bodziany M (eds) Artificial Intelligence and Its Contexts: Security, Business and Governance. Springer International Publishing, p 27–43, https://doi.org/10.1007/978-3-030-88972-2_3

Safran I, Shamir O (2018) Spurious Local Minima are Common in Two-Layer ReLU Neural Networks. In: Proceedings of the 35th International Conference on Machine Learning. PMLR, pp 4433–4441, URL https://proceedings.mlr.press/v80/safran18a.html, iSSN: 2640-3498

Salles A, Evers K, Farisco M (2020) Anthropomorphism in AI. AJOB Neuroscience 11(2):88–95. https://doi.org/10.1080/21507740.2020.1740350

Salmi J (2023) A democratic way of controlling artificial general intelligence. AI & Society 38:1785–1791. https://doi.org/10.1007/s00146-022-01426-x

Sarasin P (2001) Reizbare Maschinen: Eine Geschichte des Körpers 1765–1914. Suhrkamp

Sattarov F (2019) Power and Technology: A Philosophical and Ethical Analysis. Rowman & Littlefield

Schaeffer R, Miranda B, Koyejo S (2023) Are emergent abilities of large language models a mirage? Preprint, https://doi.org/10.48550/arXiv.2304.15004

Schmidgall S, Achterberg J, Miconi T, Kirsch L, Ziaei R, Hajiseyedrazi SP, et al. (2023) Brain-inspired learning in artificial neural networks: a review. Preprint, https://doi.org/10.48550/arXiv.2305.11252

Schmidhuber J (2015) Deep Learning in neural networks: An overview. Neural Networks 61:85–117. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003

Schmidhuber J (2022) Self-aware and conscious AI. Talk at ETH Zürich, https://www.idsia.ch/idsia_en/highlights/news/2022/2022-12-15.html

Searle J (2007) Putting consciousness back in the brain. In: Bennett M, Dennett DC, Hacker PMS, Searle JR (eds) Neuroscience and Philosophy: Brain, Mind, and Language. Columbia University Press, p 97–124, URL https://www.jstor.org/stable/10.7312/benn14044.7

Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P (2021) The role of Artificial Intelligence in healthcare: a structured literature review. BMC Medical Informatics and Decision Making 21:125. https://doi.org/10.1186/s12911-021-01488-9

Sejnowski TJ (2020) The unreasonable effectiveness of Deep Learning in Artificial Intelligence. Proceedings of the National Academy of Sciences 117(48):30,033–30,038. https://doi.org/10.1073/pnas.1907373117

Sellars WS (1962) Philosophy and the scientific image of man. In: Colodny R (ed) Science, Perception, and Reality. Humanities Press, p 35–78

Shafahi A, Huang WR, Studer C, Feizi S, Goldstein T (2020) Are adversarial examples inevitable? Preprint, https://doi.org/10.48550/arXiv.1809.02104

Sharon T (2013) Human Nature in an Age of Biotechnology: The Case for Mediated Posthumanism, Philosophy of Engineering and Technology, vol 14. Springer

Shneiderman B (2022) Human-Centered AI. Oxford University Press

Shwartz-Ziv R, Tishby N (2017) Opening the black box of deep neural networks via information. Preprint, https://doi.org/10.48550/arXiv.1703.00810

Simchon A, Edwards M, Lewandowsky S (2023) The persuasive effects of political microtargeting in the age of generative AI. Preprint, https://doi.org/10.31234/osf.io/62kxq

Sismondo S (2010) An Introduction to Science and Technology Studies. Wiley-Blackwell

Skjuve M, Følstad A, Brandtzæg PB (2023) A Longitudinal Study of Self-Disclosure in Human–Chatbot Relationships. Interacting with Computers 35(1):24–39. https://doi.org/10.1093/iwc/iwad022

Smit H, Hacker PM (2014) Seven misconceptions about the mereological fallacy: A compilation for the perplexed. Erkenntnis 79:1077–1097. https://doi.org/0.1007/s10670-013-9594-5

Smith J, de Villiers-Botha T (2023) Hey, google, leave those kids alone: Against hypernudging children in the age of big data. AI & Society 38:1639—1649. https://doi.org/10.1007/s00146-021-01314-w

Soltanolkotabi M, Javanmard A, Lee JD (2019) Theoretical Insights Into the Optimization Landscape of Over-Parameterized Shallow Neural Networks. IEEE Transactions on Information Theory 65(2):742–769. https://doi.org/10.1109/TIT.2018.2854560

Soudry D, Hoffer E, Nacson MS, Gunasekar S, Srebro N (2022) The implicit bias of gradient descent on separable data. Preprint, https://doi.org/10.48550/arXiv.1710.10345

Spaemann R (2006) Personen. Klett-Cotta

Spiekermann S (2015) Ethical It Innovation: A Value-Based System Design Approach. CRC Press

Spiekermann S (2019) Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert. Droemer

Spiekermann S (2023) Value-Based Engineering: A Guide to Building Ethical Technology for Humanity. De Gruyter

Spiekermann S, Winkler T (2022) Value-based engineering with IEEE 7000. IEEE Technology and Society Magazine 41(3):71–80. https://doi.org/10.1109/MTS.2022.3197116

Stadelmann T (2019) Wie maschinelles lernen den markt verändert. In: Haupt R, Schmitz S (eds) Digitalisierung: Datenhype mit Werteverlust?: ethische Perspektiven für eine Schlüsseltechnologie. SCM Hänssler, p 67–79

Stadelmann T, Amirian M, Arabaci I, Arnold M, Duivesteijn GF, Elezi I, et al. (2018) Deep Learning in the wild. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, pp 17–38

Stadelmann T, Braschler M, Stockinger K (2019a) Introduction to applied data science. In: Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, p 3–16, https://doi.org/10.1007/978-3-030-11821-1_1

Stadelmann T, Tolkachev V, Sick B, Stampfli J, Dürr O (2019b) Beyond imagenet: Deep Learning in industrial practice. In: Braschler M, Stadelmann T, Stockinger K (eds) Applied Data Science: Lessons Learned for the Data-Driven Business. Springer International Publishing, p 205–232, https://doi.org/10.1007/978-3-030-11821-1_12

Stewart J, Gapenne O, Di Paolo EA (eds) (2010) Enaction: Toward a New Paradigm for Cognitive Science. MIT Press

Stiegler B (1998) Technics and Time, 1: The Fault of Epimetheus. Stanford University Press

Stiegler B (2013) What Makes Life Worth Living: On Pharmacology. John Wiley & Sons

Stiegler B (2017) What is called caring? beyond the anthropocene. Techné: Research in Philosophy & Technology 21. https://doi.org/10.5840/techne201712479

Stiegler B (2018) Automatic Society, Volume 1: The Future of Work. John Wiley & Sons

Strate L (2017) Media Ecology. An Approach to Understanding the Human Condition. Understanding Media Ecology, Peter Lang Press

Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for Deep Learning in NLP. Preprint, https://doi.org/https://doi.org/10.48550/arXiv.1906.02243

Susser D (2019) Invisible influence: Artificial Intelligence and the ethics of adaptive choice architectures. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, AIES '19, p 403–408, https://doi.org/10.1145/3306618.3314286

Susser D, Roessler B, Nissenbaum H (2019) Technology, autonomy, and manipulation. Internet Policy Review 8(2). https://doi.org/10.14763/2019.2.1410

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. (2014) Intriguing properties of neural networks. Preprint, https://doi.org/https://doi.org/10.48550/arXiv.1312.6199

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. (2015) Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1–9, https://doi.org/10.1109/CVPR.2015.7298594

Tallis R (2004) Why the Mind is Not a Computer: A Pocket Lexicon of Neuromythology. Societas

Tallis R (2020) Seeing Ourselves: Reclaiming Humanity From God and Science. Agenda Publishing

Taylor C (2016) The Language Animal: The Full Shape of the Human Linguistic Capacity. Harvard University Press

Tegmark M (2018) Life 3.0. Being Human in the Age of Artificial Intelligence. Penguin Books

The Royal Society (2018) AI narratives: Portrayals and perceptions of Artificial Intelligence and why they matter. URL https://royalsociety.org/topics-policy/projects/ai-narratives/

Thompson E (2010) Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press

Thompson E, Stapleton M (2009) Making sense of sense-making: Reflections on enactive and extended mind theories. Topoi 28:23–30

Tiku N (2022) The google engineer who thinks the company's AI has come to life. The Washington Post URL https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine

Tishby N, Zaslavsky N (2015) Deep Learning and the information bottleneck principle. In: IEEE Information Theory Workshop (ITW), pp 1–5, https://doi.org/10.1109/ITW.2015.7133169

Todorov T (2016) Hope and Memory: Lessons From the Twentieth Century. Princeton University Press

Tricot R (2021) Venture capital investments in Artificial Intelligence. OECD Digital Economy Papers (319). https://doi.org/10.1787/f97beae7-en

Tu J, Li H, Yan X, Ren M, Chen Y, Liang M, et al. (2022) Exploring adversarial robustness of multi-sensor perception systems in self driving. Preprint, https://doi.org/10.48550/arXiv.2101.06784

Turiel J, Aste T (2020) Peer-to-peer loan acceptance and default prediction with Artificial Intelligence. Royal Society open science 7(6):191,649. https://doi.org/10.1098/rsos.191649

Turing A (1950) Computing Machinery and Intelligence. Mind LIX(236):433–460. https://doi.org/10.1093/mind/LIX.236.433

Turner JS (2017) Purpose & Desire: What Makes Something "Alive" and Why Modern Darwinism Has Failed to Explain It. Harper One

Uhl A (2021) Extended intelligence: Awareness-based interventions into the ecology of autonomous and intelligent systems. PhD thesis, Harvard University Graduate School of Arts and Sciences, URL https://dash.harvard.edu/handle/1/37368514

Ullman S (2019) Using neuroscience to develop Artificial Intelligence. Science 363(6428):692–693. https://doi.org/10.1126/science.aau6595

Vallès-Peris N, Domènech M (2023) Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. AI & Society 38(4):1685–1695. https://doi.org/10.1007/s00146-021-01330-w

Vallor S (2016) Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, https://doi.org/10.1093/acprof:oso/9780190498511.001.0001

Varela FJ, Thompson E, Rosch E (1992) The Embodied Mind: Cognitive Science and Human Experience. MIT Press

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al. (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Véliz C (2021) Privacy is Power. Melville House

Verbeek PP (2005) What Things Do. Philosophical Reflections on Technology, Agency, and Design. Pennsylvania State University Press

Verbeek PP (2015) Beyond interaction: A short introduction to mediation theory. Interactions 22(3):26–31. https://doi.org/10.1145/2751314

von der Malsburg C (2023) Fodor and Pylyshyn's critique of connectionism and the brain as basis of the mind. Preprint, https://doi.org/10.48550/arXiv.2307.14736

von der Malsburg C, Stadelmann T, Grewe BF (2022) A theory of natural intelligence. Preprint, https://doi.org/10.48550/arXiv.2205.00002

Véliz C (ed) (2023) The Oxford Handbook of Digital Ethics. Oxford University Press, https://doi.org/10.1093/oxfordhb/9780198857815.001.0001

Wagner B (2019) Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. Policy & Internet 11(1):104–122. https://doi.org/10.1002/poi3.198

Waldrop MM (2012) Computer modelling: Brain in a box. Nature 482(7386):456–458. https://doi.org/10.1038/482456a

Ward D, Silverman D, Villalobos M (2017) Introduction: The varieties of enactivism. Topoi 36:365–375. https://doi.org/10.1007/s11245-017-9484-6

Watson D (2019) The rhetoric and reality of anthropomorphism in Artificial Intelligence. Minds & Machines 29(3):417–440. https://doi.org/10.1007/s11023-019-09506-6

Waytz A, Gray K, Epley N, Wegner DM (2010) Causes and consequences of mind perception. Trends in Cognitive Sciences 14(8):383–388. https://doi.org/10.1016/j.tics.2010.05.006

Wehrli S, Hertweck C, Amirian M, Glüge S, Stadelmann T (2021) Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics pp 1–14. https://doi.org/10.1007/s43681-021-00108-6

Weinberg AM (1966) Can technology replace social engineering? Bulletin of the Atomic Scientists 22(10):4–8. https://doi.org/10.1080/00963402.1966.11454993

Weizenbaum J (1976) Computer Power and Human Reason: From Judgement to Calculation. W.H.Freeman & Co Ltd

Weld DS, Bansal G (2019) The challenge of crafting intelligible intelligence. Communications of the ACM 62(6):70–79. https://doi.org/10.1145/3282486

Wilson AD, Golonka S (2013) Embodied cognition is not what you think it is. Frontiers in Psychology 4:58. https://doi.org/10.3389/fpsyg.2013.00058

Wilson DG (2017) The ethics of automated behavioral microtargeting. AI Matters 3(3):56–64. https://doi.org/10.1145/3137574.3139451

Wing JM (2021) Trustworthy AI. Communications of the ACM 64(10):64–71. https://doi.org/10.1145/3448248

Winner L (2020) The Whale and the Reactor. A Search for Limits in an Age of High Technology, 2nd edn. University of Chicago Press

Wittgenstein L (2013 [1921]) Tractatus Logico-Philosophicus. Routledge

Wolfe C (2010) What is Posthumanism? University of Minnesota Press

Xiang J, Tao T, Gu Y, Shu T, Wang Z, Yang Z, et al. (2023) Language models meet world models: Embodied experiences enhance language models. Preprint, https://doi.org/10.48550/arXiv.2305.10626

Xie Q, Luong MT, Hovy E, Le QV (2020) Self-training with noisy student improves imagenet classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10,684–10,695, https://doi.org/10.1109/CVPR42600.2020.01070

Yan P, Abdulkadir A, Rosenthal M, Schatte GA, Grewe BF, Stadelmann T (2023) A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. Preprint, https://doi.org/10.48550/arXiv.2307.05638

Yang CC (2022) Explainable Artificial Intelligence for predictive modeling in healthcare. Journal of Healthcare Informatics Research 6(2):228–239. https://doi.org/10.1007/s41666-022-00114-1

Yasnitsky LN (2020) Whether Be New "Winter" of Artificial Intelligence? In: Antipova T (ed) Integrated Science in Digital Age. Springer International Publishing, Lecture Notes in Networks and Systems, pp 13–17, https://doi.org/10.1007/978-3-030-22493-6_2

Yazdanpanah V, Gerding EH, Stein S, Dastani M, Jonker CM, Norman TJ, et al. (2023) Reasoning about responsibility in autonomous systems: challenges and opportunities. AI & Society 38(4):1453–1464. https://doi.org/10.1007/s00146-022-01607-8

Yudkowski E (2023) Will superintelligent AI end the world? Youtube, URL https://www.youtube.com/watch?v=Yd0yQ9yxSYY

Zahavi D (2006) Thinking about (self-)consciousness: Phenomenological perspectives. In: Kriegel U, Williford K (eds) Self-Representational Approaches to Consciousness. MIT Press, p 273–296

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014. Springer International Publishing, pp 818–833

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2017) Understanding Deep Learning requires rethinking generalization. Preprint, https://doi.org/10.48550/arXiv.1611.03530

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2021) Understanding Deep Learning (still) requires rethinking generalization. Communications of the ACM 64(3):107–115. https://doi.org/10.1145/3446776

Zhou DX (2020) Universality of deep convolutional neural networks. Applied and computational harmonic analysis 48(2):787–794

Zuboff S (2019) The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power. Public Affairs

Zuboff S (2023) The age of surveillance capitalism. In: Longhofer W, Winchester D (eds) Social Theory Re-Wired. Routledge, p 203–213

Zuiderveen Borgesius FJ, Möller J, Kruikemeier S, Ó Fathaigh R, Irion K, Dobber T, et al. (2018) Online political microtargeting: Promises and threats for democracy. Utrecht Law Review https://doi.org/10.18352/ulr.420