AI and Ethics
https://doi.org/10.1007/s43681-023-00408-z

REVIEW

Assessing deep learning: a work program for the humanities in the age of artificial intelligence

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

Received: 25 September 2023 / Accepted: 29 November 2023

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. This paper contains a thematic introduction (Sect. 1), an introduction to the workings of DL for non-technical readers (Sect. 2), and a main part, containing the outlines of a work program for the humanities (Sect. 3). Readers familiar with DL might want to ignore 2 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 [1–3], the field of artificial intelligence (AI) entered what has proven to be its most impressive period of advancement. The methods introduced with DL perform remarkably well in identifying complex patterns in large data sets to make predictions. Today, DL has found its way from research into our daily lives in a multitude of applications [4, 5], 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 [6]. 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' [7–9]), while generalist models (e.g., 'GATO' [10]), and the unprecedented utility of multimodal 'large language models'

✉ Jan Segessenmann
jan.segessenmann@unifr.ch

Thilo Stadelmann
stdm@zhaw.ch

Andrew Davison
apd31@cam.ac.uk

Oliver Dürr
oliver.duerr@unifr.ch

[1] Center for Faith & Society, University of Fribourg, Fribourg, Switzerland

[2] Centre for Artificial Intelligence, Zurich University of Applied Sciences, Zurich, Switzerland

[3] European Centre for Living Technology, University of Venice, Venice, Italy

[4] Faculty of Divinity, University of Cambridge, Cambridge, UK

[5] Institute of Hermeneutics and Philosophy of Religion, University of Zurich, Zurich, Switzerland

Published online: 21 December 2023

(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 [11, 12]. 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 [13–15]. LLMs, such as 'ChatGPT', 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 [16]).

Despite these successes, our theoretical insight into why DL performs so well is still shallow, and some of its success remains a mystery [17–23]. 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 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' [24–30]. 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 [31]. 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" [32, 33] or a "technological singularity" [34–37]. Not surprisingly, the astonishing performance of DL applications has given rise to anthropomorphisms and even a longing for—or fear of [38]—*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, 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 [39] 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 [40–42]. Next to such concerns, engineers also have technical reasons for wanting to understand the 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) [43–49]—which is sometimes also referred to as 'intelligible' [50, 51] or 'reviewable' AI [52] (on this, see also the 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 [49, 53]. 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' [54, 55], 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 [56], societal, political, and ecological concerns remain essential [57] and have, for example, led to bans on facial recognition, and a consequent slowing of research in that area [58]. 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 [59, 60]. The danger here is conceptual and methodological confusion, with socio-political 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 among other machines [61–64]? Although the confusion of human beings with machines, and especially computers, has a long history [65–68], notable recent achievements in DL have greatly contributed to the myth of the 'electronic person'—as seen, for instance, in a work by the European Commission to address the status of sophisticated robots in terms of 'persons' [69]. 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 [70, 71]. Thus, the perceived comparability of human and computational forms of intelligence has propelled the anthropomorphization of DL language [72–74]. Now running in the opposite direction, definitions of intelligence in purely technical terms [75, 76] are often projected back onto humans and perceived as the norm of intelligence *tout court* [77–80]. Evidence that ‘intelligence’ and other characteristics of the mind can indeed be modeled as computational processes seem to be increasing [81], as DL models continue to deliver impressive results (notable, for instance, in the tendency to ascribe previously unknown ‘creativity’ to AI-generated ‘art’ [82]).

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 [83]. 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 [84]. 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 Sect. 2, we provide a brief and up-to-date introduction of the known and unknown aspects of DL, written with 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 Sect. 2.5 refers to some gaps in this theory. Readers already familiar with these concepts might want to skip Sect. 2. In Sect. 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 (Sect. 3.1), efforts to contextualize DL more broadly (Sect. 3.2), and exemplary resources, provided by the humanities in dealing with questions arising from DL deployment (Sect. 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 [85]. 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 [86], which itself is a form of AI [87]. 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 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 (Sect. 2.1). We then elaborate how they work by means of ‘universal approximation’ (Sect. 2.2). Next, we analyze a set of inevitable errors that apply to every such system, based on their architecture and training algorithm (Sect. 2.3). Section 2.4 aims, more specifically, to familiarize readers with the core concepts of current generative language models, like ChatGPT. The last Sect. 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 [88–90]; for a historical summary, see [91], and for a contemporary introduction, see [92]). 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 Fig. 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 [93]). In this case, the input would contain the grayscale 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, the input and output layers are chosen to represent something meaningful (in this case, images and the 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’).

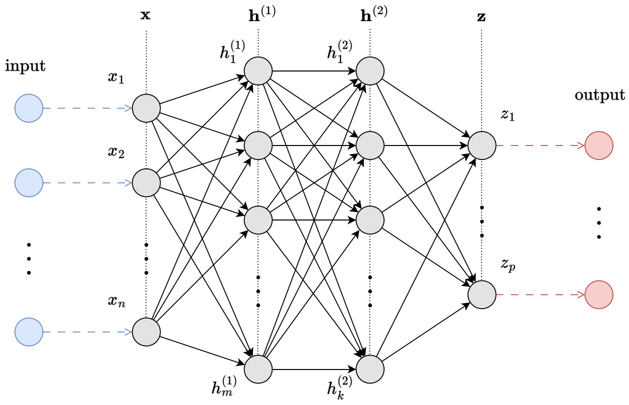

Fig. 1 Illustration of an ANN with two hidden layers $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$. The mapping from input to output is shown from left to right. The input and output consist of an array of numbers, respectively, e.g., corresponding with the pixel values of an image. Each hidden unit $h_j^{(i)}$ 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

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 [94]. 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 [95–97]. 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 Fig. 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 multidimensional data space (see Fig. 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 multidimensional 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 fewer units corresponds to collapsing the data space. Based on the

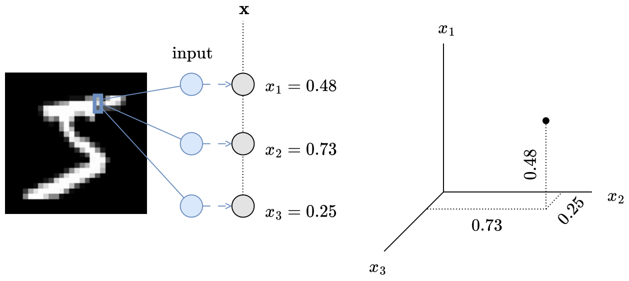

Fig. 2 Representation of an input image by a point in space. On the left is an image showing the handwritten digit '5'. Imagine that only three pixels are fed into the input layer of an ANN (the following layers are not shown), represented by their grayscale value between 0 and 1. On the right, the input 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

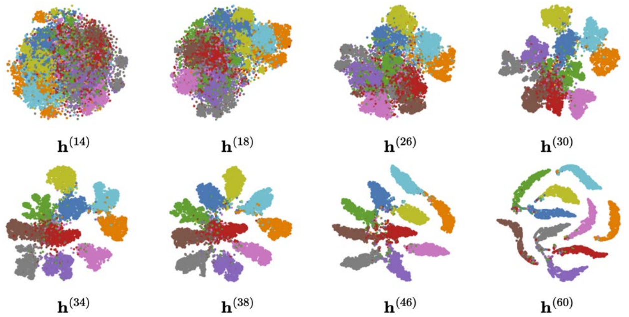

Fig. 3 Effect of the data space transformations within an ANN that classifies images into ten classes, such as 'plane', 'car', 'bird', 'cat', etc. Every point in the plots corresponds to one image. Colors represent the respective class. Looking at the data representation at different hidden layers $\mathbf{h}^{(14)}$ 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 [108] with permission 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

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 Sect. 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 [20, 98–101]. 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 [98, 102].

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 in [103, 104], with explanatory approaches proposed in [18, 105–107] and summarized in [92]). Figure 3 shows the effect that this feature abstraction has on classification capabilities, while 4 visualizes the respective features themselves.

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 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

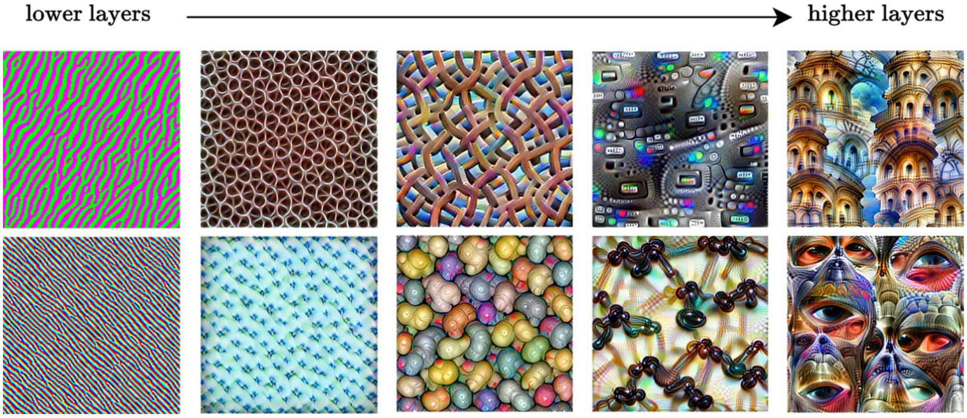

Fig. 4 Progression of data representation in the DL network 'GoogLeNet' [109]. GoogLeNet is an instance of a convolutional neural network (CNN), which achieves state-of-the-art performance in image analysis tasks (for an accessible introduction, see [110]). The shown images were achieved by fixing the 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. [111] with permission

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. Here, we rather want to give a general idea (see Sect. 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 is 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 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 below), although the current state of DL theory suggests that any found optimum should be 'good enough' in practice [92]. The empiricist process of iterating through examples and nudging model parameters thus involves inevitable errors. These can be categorized as follows.

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.

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.

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. 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 [112]. 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' [113, 114]. This is an inevitable characteristic of DL models [115, 116] and poses problems in various applications, such as self-driving cars [117, 118], making the presence of additional processes to detect such out-of-distribution samples necessary [119]. 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 Sect. 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 [120, 121]. 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 [122]—see [92] for a more detailed introduction.

Generative Pretrained Transfomer (GPT) models [123] 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 in [124], 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' ('attention' here is a purely technical term that describes the statistical importance of word representations for one another as computed from their co-occurrence in the training data). 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 [124]). 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 to 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 [125–129], 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 [12] 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 Sect. 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 [127], 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 [130]. 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" [131, 132], critically discussed in [129]).

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 [133] and gave rise to numerous algorithmic advances, which account for another large part [110], a unified theory that fully justifies the remarkable performance of DL models is still missing [17, 22, 23], although progress seems to be made [134, 135]. Two 'unknowns' remain particularly significant, which we will discuss here—albeit only briefly. For a more complete and detailed overview, we refer to [17] or [18], for a more mathematical approach to [19], and for an in-depth mathematical investigation to [20]. What *is* known theoretically about DL workings is summarized, e.g., in [92, 136].

DL generalizes surprisingly well: As elaborated in Sect. 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 [21, 22, 137, 138]. 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 [139]. 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 [19, 140, 141]. 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 [18, 142]. 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 [135]—nevertheless, overfitting remains an issue when training deep models [143], and there exist several methods to counter this by penalizing complexity during training.

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 to 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' [144–146]. 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 [102]). What does this mean? Goodfellow et al. [146] provide 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 [146, 147].

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 (otherwise often implicit) 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.

In the following, we work within a 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* accounts of humanism [148] (see, e.g., the website 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 religious–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 [149]. (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' [150, 151] as well as with some concerns of (critical) 'post-humanism' [152–155], 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 [156, 157].

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 [33, 158]. 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 (Sect. 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 (Sect. 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 (Sect. 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' [159] 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: a tendency to anthropomorphize AI (Sect. 3.1.1), and the corresponding tendency to technomorphize human beings (Sect. 3.1.2)—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 technology is part of and shapes human life (nature and culture), such that human beings must be understood as inherently related to them (Sect. 3.1.3). Further practical and applied questions and suggestions stemming from these observations are provided in Sect. 3.3.

3.1.1 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 and supernatural phenomena [160]. 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' [161]. This is a notable propensity not only in the public sphere, but also in AI research [72, 162–165].

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 [13, 166]. Others argue that robotic AI systems are candidates for personal rights [61], which is true, particularly for people under thirty who believe that future robots will develop cognition and affect [167]. 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 [168–170]. 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 [171]. Cultural variations of these phenomena can be observed, for instance, in a comparison between Europe and Japan [172, 173]. 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 [174]. Thomas Fuchs even diagnoses a novel form of "digital animism" in Western societies [175]. There is a notable tendency of people to trust computers more than human beings in decision-making processes [176], leading to an 'overtrust' [177, 178]. 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 [129, 175] is most consequential if it leads to deployment in decision-making processes that existentially affect human lives, e.g., in jurisprudence [179], policing [180], banking [181], and insurance [182]. 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 [163]. Kostopoulos [73] 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 [183–187]. 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 [188]). 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 [72] 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 [189].

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 [190]. 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 (e.g., see [38]). 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.

3.1.2 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' [163, 191]. 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., [192–194]). 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 [195–197].

DL anthropomorphism, however, and the dynamics of seeing ourselves in the image of our technology, has a pedigree reaching back to antiquity. [68, 198]. It gained modern plausibility with the scientific and industrial revolutions, and the ascent of an all-encompassing mechanistic world picture since at least the seventeenth century [65–67, 199–201]. 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 twentieth 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., [33, 202, 203]; and, for a critical perspective, see [204]). 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 [205]. 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 [206]). However, 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 [207]—had to overcome the purely behaviorist approach to intelligence. This was achieved on the basis of functionalism [208] with the central concept of mental representations [81, 209–211]. 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' [78, 212, 213]. 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 to four 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 [65]. Kenny [214] has provided helpful clarifications in addressing what he termed the "homunculus fallacy" (pp. 125–136)—otherwise also addressed as "mereological fallacy" [215, pp. 79–93], or more broadly as 'cerebrocentrism' [216–218]. 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" [214, 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." (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" (p. 129). This holds against critics, who say that knowing something *is* to be in a neural state (e.g., [219, 220]), because "to know something is ability-like, and hence more akin to a potentiality than to an actuality (a state)" [221, p. 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 [222]. Janich [223] 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., [224, 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 [222, 225–227].

Thirdly, computer-anthropologies neglect the phenomenon of life for subjectivity. According to behaviorist and functionalist accounts, we ascribe subjectivity to things based on solipsism and inference, i.e., we take the 'intentional stance' toward an object by deducing that it is a subject [206, 228]. However, research on 'embodied cognition' indicates that this is not true [229]. Rather, we presuppose selfhood from the outset as we engage embodied participants in a common form of living [230, 231]. 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 [232, 233]—which is why some cognitive scientists place an increasing 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 [175] terms this sharing of a form of living 'conviviality' [175]. According to this view, even today's most advanced language models, with their surprisingly human-like outputs, do not 'understand' anything any more than a pocket calculator or a stone can. In this view, substrate does matter, and a simulated body in a virtual space—which some label 'embodiment' (see, e.g., [234])—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 by computer-anthropologies. In the same manner, Fuchs [222] argues that 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.

The fourth concern is more grave still. Modeling humans on computers can have dehumanizing effects [59, 60, 67, 217, 235]. This is sometimes referred to as 'mechanistic dehumanization' [236–238] The historical record of those who saw and treated people as machines, programmable at will, is sinister [236, 239]. 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., [240]). 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 [62]). 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.

In sum: The exclusive focus on the brain, the neglect of subjective experience and the phenomenon of life, and the dehumanizing effects are just four 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.

3.1.3 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 [241]. 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 [242–245]. Technologies do something to us as we do something with them [246] and thus make vital an encompassing analysis of the structure of human–technology systems as well as their 'co-evolution' [235, 247, 248].

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 [249–251], media philosophy and media ecology [252–255], phenomenology and postphenomenology [246, 256–258], and the interdisciplinary field of 'science and technology studies' ([259, 260], see also [261, 262]). 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" [263, p. 31]. In transforming our environments, DL applications are not merely neutral or passive instruments, but have their own kind of agency [257, 259]. They transform our experiential, cultural, and social environments with lasting effect [235, 264–266]. 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 [267]. This practically forces us to revise our notion of human 'autonomy' [268] (on this see Sects. 3.3. 3.3.3 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 [235, 269, 270]. This goes against the grain of both 'technological determinism', for which technology is the only decisive factor [271] or 'socio-cultural determinism', for which it is only social and economic factors and human action which determine outcomes [241]. Empirically, both sides seem to have a point but are lopsided in their exclusivity of other factors [272].

The case for a more holistic and relational anthropology of technology sets out phenomenologically from the experience of lived embodiment (*Leiblichkeit*, see [222, 235]). 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 [273, 274]. Another example is prostheses, which has led philosophers of technology to speak of the 'prostheticity' of technology more broadly [275]. 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 [276]. (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 Sects. 3.3 and 4.)

A promising anthropological starting point for such a project seems to be a line of thought under the previously mentioned notion "embodied cognition" (see Sect. 3.1.2), which has recently attracted significant attention within and outside cognitive science, and which is most distinctly represented by theories of "enactivism" [222, 229, 277–282] (for an introduction to the varieties of enactivism, see [283], for an overview over the very dispersed field of cognitive science in general, see [284–287]).

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 [175, 233] as enactivism does not sit too well with the idea of 'extended' or 'substrate independent minds' (see [288–290], against, e.g., [34, 203]). 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' [291–293], 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' [294]. 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 [277, pp. 205-206] (see also [278, 295, 296]): one that involves the whole organism–environment system and, not least, considers that every explanatory perspective is subject to this co-constitutive interrelation 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 [222, 297]). This organic and phenomenological 'process' 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 Sect. 3.1.2), 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 socioculturally malleable frame of reference for how we shape our common life. Bracketing out either the 'human factor', the 'technological factor', or the 'socio-cultural frame' 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' [250, 298], 'value-sensitive design' [299], 'value-based engineering' [276], and privacy and security assessment [39, 300–302], 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 [303–311].

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 [312–315]). One strong emphasis has fallen on our inability to understand the outputs and decisions of DL models (on this, see Sect. 2.5), drawing attention to questions around harmful bias and discrimination in data-based assessments or decision-making support systems [316–318]. 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 [319], and ecological considerations (positively in as much as DL can help to work toward ecological sustainability [320], and negatively, given the ecological impact of training DL systems themselves [57, 321, 322]). Such work does not succumb to the idea that technology on its own could be the solution to our societal and planetary challenges. Just as important is *how* technology is designed, regulated, implemented, and used in our societies [323, 324]. The challenges of the digital transformation require more than a 'technical fix' [325, 326] 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.

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 above-mentioned '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 [327] 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 [11] (for a critical perspective on such claims, see, e.g., [127, 328]) and the subsequent open letter to pause "giant AI experiments", signed by leading AI researchers and CEOs [329]. While 'presentists'—in Baum's terminology—argue for the need to mitigate current societal and ecological harm (see, e.g., [322, 330, 331]), 'futurists' urge concentrating all resources on mitigating 'existential risks' (see, e.g., [332, 333]) not least from an out-of-control and misaligned superhuman intelligence [32, 33, 324] or even a so-called "singularity" [34, 36]. 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' [334] 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.

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

In this section, we outline some of the questions and issues we deem important with regard to assessing DL, drawing from all of the above-mentioned threads. 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. These aspects deepen some of the philosophical issues addressed in Sect. 3.1 above and are interrelated, i.e., they elucidate each other. Firstly, we look at the implications of an understanding of human beings as always embodied and embedded in natural, technological, and cultural environments, for assessing DL and we offer some questions (Sect. 3.3.1). Secondly, we consider the challenges we, as rational and responsible beings, face as we try to understand a world shaped by technologies we cannot comprehend (Sect. 3.3.2). 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 (Sect. 3.3.3).

3.3.1 Human beings as embodied and embedded in an environment

We have already seen that findings in embodied cognition and enactivism suggest that subjective experience is closely linked with the phenomenon of life (see Sect. 3.1.2) and the complex co-constitution between an organism and its environment (Sect. 3.1.3). What would such a view of human beings indicate for an assessment of DL systems?

The divide between living beings and DL shifts the attention away from the fear of having sentient AI anytime soon. This includes fears of making increasingly sophisticated artificial 'agents' suffer (so-called "mind crime" [32], which—if occurring on an astronomical scale—falls under the notion of "suffering risks" or "s-risks" [335]). Instead, the attention shifts toward the more realistic concern that DL systems could catastrophically impact our societies and ecology as powerful (but mindless) technologies (see Sect. 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 [129]. This goes to show that anthropological considerations, far from being distractions, actually set the course for further inquiry and action. On the contrary, considerations based on the speculative hypothesis of actual artificial agents are what distract us from setting that course toward human flourishing.

On the other side, the lived and embodied embedding of human beings in the world underpins the dramatic effect technology has on us. It reinforces efforts toward formulating and solving the problems *we* have by developing and deploying *suitable* technology. This includes problems that arise when we treat AI *as if* it were sentient, i.e., if we treat AI as others, as *someone*, despite it being *something* (on this distinction, see [336]). Furthermore, this includes efforts toward ecological sustainability because such a view 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:

- 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?
- What is at risk, if AI is perceived as others and how should we deploy AI such that these risks can be minimized?
- 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 in the context of an increasingly digitally perceived world?
- 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 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? 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?
- 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?

3.3.2 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 is usually translated as 'rational animal' but might also, as Charles Taylor correctly suggests, be rendered as 'animal possessing language' [337, 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.

Some think that the generated content of prominent LLMs amounts to understanding, knowledge, and intelligence in a human-like sense (see, e.g., [12, 338]), while others are more skeptical [128], and believe that there are other ways to explain these capabilities [127], or think we are dealing here with 'stochastic parrots' [125, 322]. 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 [175] (see also [60]). Recalling Sect. 2.4, one can explicitly state what 'understanding' constitutes if applied to LLMs: 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, the probability density function (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, see Sect. 3.3.1). 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 only 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.

The combination of technical mastery and explanatory mystery in DL marks a significant step in the history of human inquiry into reality. As we have seen in Sect. 2, the workings of trained DL systems remain opaque to our understanding. Since DL systems themselves do not understand anything, we can now engineer and deploy working systems whose inner workings remain fully opaque 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* [339]. This corresponds to debates in the philosophy of science, which increasingly question the dominance of causal explanations [340] and moving beyond epistemic reliabilism [341]. 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 [94], 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 [342]. 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 seems to bring us back closer to more practical notions of 'understanding' [343] 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 [273, 344] (on this, see also [345]). 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 [346]), 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 [347] on the basis of "niche construction" theories [348]).

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. 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 [235, 276] (we will turn to this issue in the next section).

Follow-up questions:

- Do DL systems represent a novel or perhaps stand-alone form of rationality? Are they indicative of 'how human intelligence works'?
- 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?

3.3.3 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? [28, 349, 350] This is particularly urgent to ask because the architecture of current DL systems cannot fully prevent unexpected, potentially harmful ‘rogue’ outputs (see Sect. 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. Relying on opaque DL systems thus further complicates the already challenging notion of the responsibility of engineers, laboratories, 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 Sect. 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 [268] 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 to human behavior, motivation, and, more broadly, freedom (see, e.g., [349, 351]). 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 [264, 352]. 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 [235]. 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—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” [353, 354]—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 [355–357]. 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” [358, 359]— 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 [360, 361]

Secondly, *sex*, which has always been a driver of technological innovation [362]—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 generative AI, 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) toward rationality, sociality, morality, and care, which must complement technological innovations [363].

Thirdly, it is vital to assess the relationship of technology and *power* [364, 365]. 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 [366]. In the last few years, we have increasingly seen the application of DL in the political sphere [57, 367–370]. The manipulative potentials of DL systems [371] 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' [372–383]—which were impressed on the public mind through the 'Cambridge Analytica Scandal' [384]. 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., [364, 385, 386]). 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" [358, 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 [301, 302], it is important to examine the assumptions about the human underlying these fears. Are human beings fully "hackable animals" [364, 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., [182, 331, 387–393])—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 begin by focusing on the following questions:

Follow-up questions:

- Around which values, standards, and future visions are we creating, designing, and deploying novel technologies? Who sets those markers, and with which legitimacy?
- 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 [394]. Historically, this description has been rejected—notably by John McCarthy—in favor of an 'artificial intelligence' description, for marketing and funding purposes. Given that this has now become entrenched, we suggest amending this prevailing designation, to become not AI but 'extended intelligence' [395–397]. We understand such extension in terms of enactivism and a relational anthropology as outlined above (Sect. 3.1.3) and not in terms of the 'extended mind theory' [398]. Speaking of AI as if it were truly intelligent implies a reduction of the human condition to closed systems and processes [396]. 'Extended intelligence', in our proposal, would analyze and assess DL technologies within a relational framework of human–technology systems, instead of seeing them as ontological entities *sui generis*. 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. Emphasizing the inherent complexity of such systems limits the longing for control by accentuating the deficiency of rigid optimization processes of "single currencies" (such as GDP, see [399]). Our proposal 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 [235]. 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 [276, 299, 400–405].

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.

Acknowledgements We would like to thank Sarah Spiekermann, Johannes Hoff, Thomas Fuchs, Matthias Egg, Carina Prunkl, Cyril Stoller, and Armin Grunwald for their valuable input to different sections of this paper.

Author Contributions All authors contributed to the conception and design of the article. JS drafted the first versions of chapters 1 and 2, and OD drafted the first versions of chapters 3 and 4. All authors reviewed previous versions of the manuscript and contributed from their field of expertise. All authors read and approved the final version of the manuscript.

Funding Open access funding provided by University of Fribourg. We received no funding for this paper, but were supported and provided with resources by the Institute of Hermeneutics and Philosophy of Religion at the University of Zurich, the Center for Faith & Society at the University of Fribourg, the Harris Manchester College at the University of Oxford, the Center for Theological Inquiry at the University of Princeton, and the Centre for Artificial Intelligence at the Zurich University of Applied Sciences.

Declarations

Conflict of interest We have no conflict of interest do declare.

Ethical standard Approval not applicable.

Data availability Not applicable.

Code availability Not applicable.

References

1. Hinton, G.E., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006). https://doi.org/10.1162/neco.2006.18.7.1527
2. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: 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 (2006). https://proceedings.neurips.cc/paper_files/paper/2006/file/5da713a690c067105aeb2fae32403405-Paper.pdf
3. Ranzato, M.a., Poultney, C., Chopra, S., Cun, Y.: 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 (2006). https://proceedings.neurips.cc/paper_files/paper/2006/file/87f4d79e36d68c3031ccf6c55e9bbd39-Paper.pdf
4. Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., Duivesteijn, G.F., Elezi, I., Geiger, M., Lörwald, S., Meier, B.B., Rombach, K., et al.: Deep Learning in the wild. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, pp. 17–38 (2018). Springer
5. Yan, P., Abdulkadir, A., Rosenthal, M., Schatte, G.A., Grewe, B.F., Stadelmann, T.: A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: methods, applications, and directions. Preprint (2023). https://doi.org/10.48550/arXiv.2307.05638
6. Amirian, M., Füchslin, R.M., Herzig, I., Hotz, P.E., Lichtensteiger, L., Montoya-Zegarra, J.A., Morf, M., Paysan, P., Peterlik, I., Scheib, S., et al.: Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks. Med. Phys. (2023). https://doi.org/10.1002/mp.16405
7. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. Preprint (2022). https://doi.org/10.48550/arXiv.2204.06125
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. Preprint (2022). https://doi.org/10.48550/arXiv.2112.1075
9. Borji, A.: Generated faces in the wild: quantitative comparison of stable diffusion, Midjourney and DALL-E 2. Preprint (2023). https://doi.org/10.48550/arXiv.2210.00586
10. Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., de Freitas, N.: A Generalist agent. Preprint (2022). https://doi.org/10.48550/arXiv.2205.06175
11. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: early experiments with GPT-4. Preprint (2023). https://doi.org/10.48550/arXiv.2303.12712
12. Agüera y Arcas, B.: Do large language models understand us? Daedalus **151**(2), 183–197 (2022). https://doi.org/10.1162/daed_a_01909
13. Tiku, N.: The Google engineer who thinks the company's AI has come to life. The Washington Post (2022). https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine Accessed 2022-07-31
14. Kaplan, M.: After Google chatbot becomes 'sentient,' MIT prof says Alexa could too. New York Post (2022). https://nypost.com/2022/06/13/mit-prof-says-alexa-could-become-sentient-like-google-chatbot/ Accessed 2022-07-31
15. Schmidhuber, J.: Self-aware and conscious AI. Talk at ETH Zürich, https://www.idsia.ch/idsia_en/highlights/news/2022/2022-12-15.html (2022)
16. Katz, D.M., Bommarito, M.J., Gao, S., Arredondo, P.: GPT-4 passes the bar exam. Elsevier (2023). https://doi.org/10.2139/ssrn.4389233
17. Plebe, A., Grasso, G.: The Unbearable Shallow Understanding of Deep Learning. Minds Mach. **29**(4), 515–553 (2019). https://doi.org/10.1007/s11023-019-09512-8
18. Hodas, N.O., Stinis, P.: Doing the impossible: why neural networks can be trained at all. Front. Psychol. **9** (2018). https://doi.org/10.3389/fpsyg.2018.01185
19. Poggio, T., Banburski, A., Liao, Q.: Theoretical issues in deep networks: approximation, optimization and generalization. Preprint (2019). https://doi.org/10.48550/arXiv.1908.09375
20. Berner, J., Grohs, P., Kutyniok, G., Petersen, P.: The modern mathematics of deep learning. In: Grohs, P., Kutyniok, G. (eds.) Mathematical Aspects of Deep Learning, pp. 1–111. Cambridge University Press (2022). https://doi.org/10.1017/9781009025096.002
21. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. Preprint (2017). https://doi.org/10.48550/arXiv.1611.03530
22. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021). https://doi.org/10.1145/3446776
23. Sejnowski, T.J.: The unreasonable effectiveness of deep learning in artificial intelligence. Proc. Natl. Acad. Sci. **117**(48), 30033–30038 (2020). https://doi.org/10.1073/pnas.1907373117
24. Hutson, M.: Has artificial intelligence become alchemy? Science **360**(6388), 478–478 (2018). https://doi.org/10.1126/science.360.6388.478 (**Publisher: American Association for the Advancement of Science**)
25. Ford, M.: Architects of Intelligence: The Truth About AI from the People Building it. Packt Publishing Ltd, Birmingham (2018)
26. Edwards, D., Edwarts, H.: Google's engineers say that "magic spells" are ruining AI research. Quartz (2018). Accessed 2022-05-03

27. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012). https://doi.org/10.1145/2347736.2347755
28. Martini, M.: Blackbox Algorithmus: Grundfragen Einer Regulierung Künstlicher Intelligenz. Springer (2019). https://doi.org/10.1007/978-3-662-59010-2
29. Flessner, B.: 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, pp. 63–106. Transcript Verlag (2018). https://doi.org/10.1515/9783839442876-003
30. von der Malsburg, C., Stadelmann, T., Grewe, B.F.: A theory of natural intelligence. Preprint (2022). https://doi.org/10.48550/arXiv.2205.00002
31. Campolo, A., Crawford, K.: Enchanted determinism: power without responsibility in artificial intelligence. Engaging Sci. Technol. Soc. **6**, 1–19 (2020). https://doi.org/10.17351/ests2020.277
32. Bostrom, N.: Superintelligence: Paths, Dangers. Oxford University Press, Strategies (2014)
33. Tegmark, M.: Life 3.0. Being Human in the Age of Artificial Intelligence. Penguin Books (2018)
34. Kurzweil, R.: The Singularity Is Near: When Humans Transcend Biology. Penguin Publishing Group (2005)
35. Chalmers, D.J.: The singularity: a philosophical analysis. J. Consciousness Stud. **17**(9–10), 9–10 (2010)
36. Eden, A., Steinhart, E., Pearce, D., Moor, J.: Singularity hypotheses: an overview. In: Eden, A., Pearce, D., Moor, J., Søraker, J., Steinhart, E. (eds.) Singularity Hypotheses. The Frontiers Collection, pp. 1–12. Springer (2012). https://doi.org/10.1007/978-3-642-32560-1_1
37. Barrat, J.: Our final invention: artificial intelligence and the end of the human era. St. Martin's Publishing Group (2015)
38. Yudkowski, E.: Will Superintelligent AI End the World? Youtube (2023). https://www.youtube.com/watch?v=Yd0yQ9yxSYY Accessed 2023-08-23
39. European Parliament, Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council (2016). https://data.europa.eu/eli/reg/2016/679/oj Accessed 2023-08-21
40. Grunwald, A.: The inherently democratic nature of technology assessment. Sci. Publ. Policy **46**(5), 702–709 (2019). https://doi.org/10.1093/scipol/scz023
41. Pflanzer, M., Dubljević, V., Bauer, W.A., Orcutt, D., List, G., Singh, M.P.: Embedding AI in society: ethics, policy, governance, and impacts. AI Soc. **38**, 1267–1271 (2023). https://doi.org/10.1007/s00146-023-01704-2
42. Salmi, J.: A democratic way of controlling artificial general intelligence. AI Soc. **38**, 1785–1791 (2023). https://doi.org/10.1007/s00146-022-01426-x
43. Došilović, F.K., Brčić, M., Hlupić, N., Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 210–215 (2018). https://doi.org/10.23919/MIPRO.2018.8400040
44. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
45. Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Rev. **11**(1), 1391 (2021). https://doi.org/10.1002/widm.1391
46. Joshi, G., Walambe, R., Kotecha, K.: A review on explainability in multimodal deep neural nets. IEEE Access **9**, 59800–59821 (2021). https://doi.org/10.1109/ACCESS.2021.3070212
47. Madsen, A., Reddy, S., Chandar, S.: Post-hoc interpretability for neural nlp: a survey. ACM Comput. Surveys **55**(8), (2022). https://doi.org/10.1145/3546577
48. Notovich, A., Chalutz-Ben Gal, H., Ben-Gal, I.: 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, pp. 971–985. Springer (2023). https://doi.org/10.1007/978-3-031-24628-9_41
49. Besold, T.R., Uckelman, S.L.: The what, the why, and the how of artificial explanations in automated decision-making. Preprint (2018). https://doi.org/10.48550/arXiv.1808.07074
50. Weld, D.S., Bansal, G.: The challenge of crafting intelligible intelligence. Commun. ACM **62**(6), 70–79 (2019). https://doi.org/10.1145/3282486
51. Caruana, R., Lundberg, S., Ribeiro, M.T., Nori, H., Jenkins, S.: Intelligible and explainable machine learning: Best practices and practical challenges. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '20, pp. 3511–3512. Association for Computing Machinery (2020). https://doi.org/10.1145/3394486.3406707
52. Cobbe, J., Lee, M.S.A., Singh, J.: Reviewable automated decision-making: a framework for accountable algorithmic systems. Preprint (2021). https://doi.org/10.48550/arXiv.2102.04201
53. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018). https://doi.org/10.1145/3236386.3241340
54. Floridi, L.: AI and its new Winter: from Myths to Realities. Philosophy Technol. **33**(1), 1–3 (2020). https://doi.org/10.1007/s13347-020-00396-6
55. Yasnitsky, L.N.: Whether Be New "Winter" of artificial intelligence? In: Antipova, T. (ed.) Integrated Science in Digital Age. Lecture Notes in Networks and Systems, pp. 13–17. Springer (2020). https://doi.org/10.1007/978-3-030-22493-6_2
56. Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.-C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O'Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.-V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., Wright, R.: "so what if ChatGPT wrote it?'' multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inform. Manag. **71**, 102642 (2023). https://doi.org/10.1016/j.ijinfomgt.2023.102642
57. Crawford, K.: The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press (2021)
58. Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., Stadelmann, T.: Bias, awareness, and ignorance in deep-learning-based face recognition. AI and Ethics, 1–14 (2021). https://doi.org/10.1007/s43681-021-00108-6
59. Tallis, R.: Why the Mind Is Not a Computer: A Pocket Lexicon of Neuromythology. Societas (2004)
60. Tallis, R.: Seeing Ourselves: Reclaiming Humanity From God and Science. Agenda Publishing (2020)

61. Gunkel, D.J.: Robot Rights. MIT Press (2018)
62. Gordon, J.-S., Pasvenskiene, A.: Human rights for robots? a literature review. AI and Ethics **1**(4), 579–591 (2021). https://doi.org/10.1007/s43681-021-00050-7
63. Munn, N., Weijers, D.: Corporate responsibility for the termination of digital friends. AI Soc. **38**(4), 1501–1502 (2023). https://doi.org/10.1007/s00146-021-01276-z
64. Novelli, C.: Legal personhood for the integration of AI systems in the social context: a study hypothesis. AI Soc. **38**(4), 1347–1359 (2023). https://doi.org/10.1007/s00146-021-01384-w
65. Boden, M.A.: Mind as Machine: A History of Cognitive Science. Oxford University Press (2008)
66. Black, D.: Embodiment and Mechanisation: Reciprocal Understandings of Body and Machine from the Renaissance to the Present. Ashgate Press (2014)
67. Dürr, O.: Homo Novus: Vollendlichkeit Im Zeitalter des Transhumanismus. Studia Oecumenica Friburgensia, vol. 108. Aschendorff Verlag (2021)
68. Cave, S., Dihal, K., Dillon, S.: AI Narratives: A History of Imaginative Thinking About Intelligent Machines. Oxford University Press (2020)
69. European Parliament: REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2017). https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html Accessed 2023-08-21
70. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **5**(4), 115–133 (1943). https://doi.org/10.1007/BF02478259
71. Hebb, D.O.: The Organization of Behavior. Wiley, A Neuropsychological Theory (1949)
72. Lipton, Z.C., Steinhardt, J.: 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 (2019). https://doi.org/10.1145/3317287.3328534
73. Kostopoulos, L.: Decoupling human characteristics from algorithmic capabilities. Technical report, IEEE Standards Association (2021). https://standards.IEEE.org/initiatives/artificial-intelligence-systems/decoupling-human-characteristics/ Accessed 2022-05-18
74. The Royal Society: AI Narratives: portrayals and perceptions of artificial intelligence and why they matter (2018). https://royalsociety.org/topics-policy/projects/ai-narratives/ Accessed 2023-08-21
75. Legg, S., Hutter, M.: 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, pp. 17–24. IOS Press (2007)
76. Chollet, F.: On the measure of intelligence. Preprint (2019). https://doi.org/10.48550/arXiv.1911.01547
77. Dennett, D.C.: Consciousness Explained. Penguin Books (1991)
78. Churchland, P.S., Sejnowski, T.J.: The Computational Brain. MIT Press (1992)
79. Chalmers, D.J.: A computational foundation for the study of cognition. J. Cognit. Sci. **12**(4), 325–359 (2011). https://doi.org/10.17791/jcs.2011.12.4.325
80. Boden, M.A.: Computer Models of Mind: Computational Approaches in Theoretical Psychology. Cambridge University Press (1988)
81. von der Malsburg, C.: Fodor and Pylyshyn's Critique of Connectionism and the Brain as Basis of the Mind. Preprint (2023). https://doi.org/10.48550/arXiv.2307.14736
82. Mazzone, M., Elgammal, A.: Art, creativity, and the potential of Artificial Intelligence. Arts **8**(1), (2019). https://doi.org/10.3390/arts8010026
83. Liggieri, K., Müller, O. (eds.): Mensch-Maschine-Interaktion: Handbuch Zu Geschichte - Kultur - Ethik. J.B, Metzler (2019)
84. Stiegler, B.: What is called caring? beyond the anthropocene. Techné: Research in Philosophy & Technology **21**, (2017). https://doi.org/10.5840/techne201712479
85. Marcus, G.: Deep learning: a critical appraisal. Preprint (2018). https://doi.org/10.48550/arXiv.1801.00631
86. Mitchell, T.: Machine Learning. McGraw Hill (1997)
87. Russell, S., Norvig, P.: Artificial intelligence: a modern approach. Global Edition, Pearson Education (2021)
88. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**(6), 386–408 (1958). https://doi.org/10.1037/h0042519
89. Minsky, M., Papert, S.A.: Perceptrons: An Introduction to Computational Geometry. MIT Press (1969)
90. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986). https://doi.org/10.1038/323533a0
91. Schmidhuber, J.: Deep Learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015). https://doi.org/10.1016/j.neunet.2014.09.003
92. Prince, S.J.D.: Understanding Deep Learning. MIT Press (2023)
93. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
94. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
95. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989). https://doi.org/10.1016/0893-6080(89)90020-8
96. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**(4), 303–314 (1989). https://doi.org/10.1007/BF02551274
97. Zhou, D.-X.: Universality of deep convolutional neural networks. Appl. Comput. Harmonic Anal. **48**(2), 787–794 (2020)
98. Bengio, Y., LeCun, Y.: Scaling learning algorithms toward AI. In: Bottou, L., Chapelle, O., DeCoste, D., Weston, J. (eds.) Large-Scale Kernel Machines. MIT Press (2007). https://doi.org/10.7551/mitpress/7496.001.0001
99. Eldan, R., Shamir, O.: The power of depth for feedforward neural networks. Preprint (2016). https://doi.org/10.48550/arXiv.1512.03965
100. Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., Dickstein, J.S.: On the expressive power of deep neural networks. In: Proceedings of the 34th International Conference on Machine Learning. ICML'17, pp. 2847–2854. JMLR.org (2017). https://doi.org/10.5555/3305890.3305975
101. Lin, H.W., Tegmark, M., Rolnick, D.: Why does deep and cheap learning work so well? J. Stat. Phys. **168**(6), 1223–1247 (2017). https://doi.org/10.1007/s10955-017-1836-5
102. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
103. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09, pp. 609–616. Association for Computing Machinery (2009). https://doi.org/10.1145/1553374.1553453
104. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B.,

Tuytelaars, T. (eds.) Computer Vision – ECCV 2014, pp. 818–833. Springer (2014)

105. Mhaskar, H., Liao, Q., Poggio, T.: When and why are deep networks better than shallow ones? Proceedings of the AAAI Conference on Artificial Intelligence **31**(1), (2017). https://doi.org/10.1609/aaai.v31i1.10913
106. Frankle, J., Carbin, M.: The lottery ticket hypothesis: finding sparse, trainable neural networks. Preprint (2019). https://doi.org/10.48550/arXiv.1803.03635
107. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. Preprint (2017). https://doi.org/10.48550/arXiv.1703.00810
108. Hoyt, C.R., Owen, A.B.: Probing neural networks with t-SNE, class-specific projections and a guided tour. Preprint (2021). https://doi.org/10.48550/arXiv.2107.12547
109. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
110. Stadelmann, T., Tolkachev, V., Sick, B., Stampfli, J., Dürr, O.: 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, pp. 205–232. Springer (2019). https://doi.org/10.1007/978-3-030-11821-1_12
111. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017). https://doi.org/10.23915/distill.00007
112. Delétang, G., Ruoss, A., Grau-Moya, J., Genewein, T., Wenliang, L.K., Catt, E., Cundy, C., Hutter, M., Legg, S., Veness, J., Ortega, P.A.: Neural networks and the chomsky hierarchy. Preprint (2023). https://doi.org/10.48550/arXiv.2207.02098
113. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. Preprint (2014). https://doi.org/10.48550/arXiv.1312.6199
114. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. Preprint (2015). https://doi.org/10.48550/arXiv.1412.6572
115. Shafahi, A., Huang, W.R., Studer, C., Feizi, S., Goldstein, T.: Are adversarial examples inevitable? Preprint (2020). https://doi.org/10.48550/arXiv.1809.02104
116. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ASIA CCS '17, pp. 506–519. Association for Computing Machinery (2017). https://doi.org/10.1145/3052973.3053009
117. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial Patch. Preprint (2018). https://doi.org/10.48550/arXiv.1712.09665
118. Tu, J., Li, H., Yan, X., Ren, M., Chen, Y., Liang, M., Bitar, E., Yumer, E., Urtasun, R.: Exploring adversarial robustness of multi-sensor perception systems in self driving. Preprint (2022). https://doi.org/10.48550/arXiv.2101.06784
119. Amirian, M., Schwenker, F., Stadelmann, T.: 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, pp. 346–358. Springer (2018). https://doi.org/10.1007/978-3-319-99978-4_27
120. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25. Curran Associates, Inc. (2012). https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
121. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
122. OpenAI: GPT-4 Technical Report. Preprint (2023). https://doi.org/10.48550/arXiv.2303.08774
123. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-training. https://openai.com/research/language-unsupervised Accessed 2023-08-23
124. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
125. Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5185–5198. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.463
126. Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., Turian, J.: Experience grounds language. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8718–8735. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.703
127. Durt, C., Froese, T., Fuchs, T.: Against AI understanding and sentience: large language models, meaning, and the patterns of human language use. Preprint (2023). http://philsci-archive.pitt.edu/21983/
128. Marcus, G., Leivada, E., Murphy, E.: A Sentence is worth a thousand pictures: can large language models understand human language? Preprint (2023). https://doi.org/10.48550/arXiv.2308.00109
129. Dürr, O., Segessenmann, J., Steinmann, J.J.: Meaning, form, and the limits of natural language processing. Philosophy Theol. Sci. **10**(1), 42–72 (2023). https://doi.org/10.1628/ptsc-2023-0005
130. Pavlick, E.: Symbols and grounding in large language models. Philosophical Trans. A Math. Phys. Eng. Sci. **381**(2251), 20220041 (2023). https://doi.org/10.1098/rsta.2022.0041
131. LeCun, Y.: A Path towards autonomous machine intelligence. Preprint (2022). https://openreview.net/pdf?id=BZ5a1r-kVsf
132. Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., Morimoto, J.: Deep Learning, reinforcement learning, and world models. Neural Netw. **152**(C), 267–275 (2022). https://doi.org/10.1016/j.neunet.2022.03.037
133. Lenzen, M.: Künstliche Intelligenz: Fakten, Chancen. Risiken. C.H, Beck (2020)
134. Ma, Y., Tsao, D., Shumm, H.Y.: On the principles of parsimony and self-consistency for the emergence of intelligence. Front. Inform. Technol. Electron. Eng. **23**(9), 1298–1323 (2022). https://doi.org/10.1631/FITEE.2200297
135. Liu, Z., Kitouni, O., Nolte, N., Michaud, E.J., Tegmark, M., Williams, M.: Towards understanding grokking: an effective theory of representation learning. Preprint (2022). https://doi.org/10.48550/arXiv.2205.10343
136. Roberts, D.A., Yaida, S., Hanin, B.: The Principles of Deep Learning Theory. Cambridge University Press (2022)
137. Soltanolkotabi, M., Javanmard, A., Lee, J.D.: Theoretical Insights Into the Optimization Landscape of Over-Parameterized Shallow Neural Networks. IEEE Trans. Inform. Theory **65**(2), 742–769 (2019). https://doi.org/10.1109/TIT.2018.2854560

138. Martinetz, J., Martinetz, T.: Highly over-parameterized classifiers generalize since bad solutions are rare. Preprint (2023). https://doi.org/10.48550/arXiv.2211.03570
139. Xie, Q., Luong, M.-T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2020). https://doi.org/10.1109/CVPR42600.2020.01070
140. Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S., Srebro, N.: The implicit bias of gradient descent on separable data. Preprint (2022). https://doi.org/10.48550/arXiv.1710.10345
141. Arora, S., Cohen, N., Hu, W., Luo, Y.: Implicit regularization in deep matrix factorization. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc. (2019)
142. Tishby, N., Zaslavsky, N.: Deep Learning and the information bottleneck principle. In: IEEE Information Theory Workshop (ITW), pp. 1–5 (2015). https://doi.org/10.1109/ITW.2015.7133169
143. Tuggener, L., Schmidhuber, J., Stadelmann, T.: Is it enough to optimize CNN architectures on ImageNet? Front. Comput. Sci. **4**, 1041703 (2022)
144. Bellman, R.E.: Adaptive Control Processes. Princeton University Press (2015)
145. Novak, E., Woźniakowski, H.: Approximation of infinitely differentiable multivariate functions is intractable. J. Complexity **25**(4), 398–404 (2009). https://doi.org/10.1016/j.jco.2008.11.002
146. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
147. Brahma, P.P., Wu, D., She, Y.: Why deep learning works: a manifold disentanglement perspective. IEEE Trans. Neural Netw. Learn. Syst. **27**(10), 1997–2008 (2016). https://doi.org/10.1109/TNNLS.2015.2496947
148. Flynn, T.: A secular humanist definition setting the record straight. Free Inquiry (2002)
149. Grey, C., Dürr, O.: On changing the subject: Secularity, religion, and the idea of the human. Religions **14**(4), (2023). https://doi.org/10.3390/rel14040466
150. Antweiler, C.: Inclusive Humanism: Anthropological Basics for a Realistic Cosmopolitanism. Vandenhoeck & Ruprecht (2012)
151. Antweiler, C.: Pan-cultural universals. a fundament for an inclusive humanism. In: Rüsen, J. (ed.) Approaching Humankind. Towards an In-tercultural Humanism, pp. 37–68. Vandenhoeck & Ruprecht (2013)
152. Foucault, M.: Les Mots et les Choses. Gallimard Paris (1990)
153. Herbrechter, S.: Posthumanismus: Eine Kritische Einführung. WBG (2009)
154. Wolfe, C.: What Is Posthumanism? University of Minnesota Press (2010)
155. Braidotti, R.: The Posthuman. Polity Press (2013)
156. Wittgenstein, L.: Tractatus Logico-Philosophicus. Routledge (2013 [1921])
157. Leung, K.-H.: The picture of artificial intelligence and the secularization of thought. Political Theol. **20**(6), 457–471 (2019). https://doi.org/10.1080/1462317X.2019.1605725
158. Boden, M.A.: AI: Its Nature and Future. Oxford University Press (2016)
159. Crutzen, P.J., Stoermer, E.F.: The anthropocene [2000]. In: Robin, L., Sörlin, S., Warde, P. (eds.) The future of nature, pp. 479–490. Yale University Press (2013). https://doi.org/10.12987/9780300188479-041
160. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. **114**(4), 864 (2007). https://doi.org/10.1037/0033-295X.114.4.864
161. Waytz, A., Gray, K., Epley, N., Wegner, D.M.: Causes and consequences of mind perception. Trends Cognit. Sci. **14**(8), 383–388 (2010). https://doi.org/10.1016/j.tics.2010.05.006
162. Proudfoot, D.: Anthropomorphism and AI: turing's much misunderstood imitation game. Artificial Intell. **175**(5), 950–957 (2011). https://doi.org/10.1016/j.artint.2011.01.006 (**Special Review Issue**)
163. Salles, A., Evers, K., Farisco, M.: Anthropomorphism in AI. AJOB Neurosci. **11**(2), 88–95 (2020). https://doi.org/10.1080/21507740.2020.1740350
164. Watson, D.: The rhetoric and reality of anthropomorphism in Artificial Intelligence. Minds Mach. **29**(3), 417–440 (2019). https://doi.org/10.1007/s11023-019-09506-6
165. Cave, S., Coughlan, K., Dihal, K.: "Scary robots": Examining public responses to AI. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES '19, pp. 331–337. Association for Computing Machinery (2019). https://doi.org/10.1145/3306618.3314232
166. Lemoine, B.: Is LaMDA Sentient? An Interview (2022). https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 Accessed 2023-08-23
167. de Graaf, M.M.A., Hindriks, F.A., Hindriks, K.V.: Who wants to grant robots rights? In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI '21 Companion, pp. 38–46. Association for Computing Machinery (2021). https://doi.org/10.1145/3434074.3446911
168. Dosovitsky, G., Bunge, E.L.: Bonding with bot: User feedback on a chatbot for social isolation. Front. Digital Health **3**, 735053 (2021). https://doi.org/10.3389/fdgth.2021.735053
169. Skjuve, M., Følstad, A., Brandtzæg, P.B.: A longitudinal study of self-disclosure in human-chatbot relationships. Interacting Comput. **35**(1), 24–39 (2023). https://doi.org/10.1093/iwc/iwad022
170. Crolic, C., Thomaz, F., Hadi, R., Stephen, A.T.: Blame the bot: anthropomorphism and anger in customer-chatbot interactions. J. Marketing **86**(1), 132–148 (2022). https://doi.org/10.1177/00222429211045687
171. Darling, K.: "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 (2017). https://doi.org/10.1093/oso/9780190652951.003.0012
172. Haring, K.S., Mougenot, C., Ono, F., Watanabe, K.: Cultural differences in perception and attitude towards robots. Int. J. Affective Eng. **13**(3), 149–157 (2014). https://doi.org/10.1007/s12369-022-00920-y
173. Robertson, J.: Human rights vs. robot rights: forecasts from Japan. Critical Asian Stud. **46**(4), 571–598 (2014). https://doi.org/10.1080/14672715.2014.960707
174. Robertson, J.: Robo Sapiens Japanicus: Robots. Family, and the Japanese Nation. University of California Press, Gender (2018)
175. Fuchs, T.: Understanding sophia? on human interaction with artificial agents. Phenomenol. Cognit. Sci. (2022). https://doi.org/10.1007/s11097-022-09848-0
176. Bogert, E., Schecter, A., Watson, R.T.: Humans rely more on algorithms than social influence as a task becomes more difficult. Sci. Rep. **11**(1), 8028 (2021). https://doi.org/10.1038/s41598-021-87480-9
177. Hardré, P.L.: When, how, and why do we trust technology too much? In: Tettegah, S.Y., Espelage, D.L. (eds.) Emotions, technology, and behaviors. Emotions Technol., pp. 85–106. Academic Press (2016). https://doi.org/10.1016/B978-0-12-801873-6.00005-4
178. Aroyo, A.M., de Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M., Tamò-Larrieux, A.: Overtrusting robots: setting a research agenda to mitigate overtrust in automation. Paladyn J.

Behav. Robot. **12**(1), 423–436 (2021). https://doi.org/10.1515/pjbr-2021-0029

179. Ryberg, J., Roberts, J.V.: Sentencing and Artificial Intelligence. Oxford University Press (2022)
180. McDaniel, J., Pease, K.: Predictive Policing and Artificial Intelligence. Routledge (2021)
181. Donepudi, P.K.: Machine learning and artificial intelligence in banking. Eng. Int. **5**(2), 83–86 (2017). https://doi.org/10.18034/ei.v5i2.490
182. Lamberton, C., Brigo, D., Hoy, D.: Impact of robotics, rpa and AI on the insurance industry: Challenges and opportunities. J. Financial Perspectives **4**(1), (2017)
183. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching Machines to Read and Comprehend. In: Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. (2015). https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html Accessed 2022-05-04
184. Mozer, M.C.: Neural network music composition by prediction: exploring the benefits of psychoacoustic constraints and multiscale processing. Connect. Sci. **6**(2–3), 247–280 (1994). https://doi.org/10.1080/09540099408915726
185. Reizinger, P., Szemenyei, M.: 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 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054546
186. Nguyen, A.M., Yosinski, J., Clune, J.: Innovation Engines: Automated Creativity and Improved Stochastic Optimization via Deep Learning. In: Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation. GECCO '15, pp. 959–966. Association for Computing Machinery (2015). https://doi.org/10.1145/2739480.2754703
187. Lipton, Z.C., Azizzadenesheli, K., Kumar, A., Li, L., Gao, J., Deng, L.: Combating Reinforcement learning's sisyphean curse with intrinsic fear. Preprint (2018). https://doi.org/10.48550/arXiv.1611.01211
188. Davison, A.: Machine learning and theological traditions of analogy. Modern Theol. **37**(2), 254–274 (2021). https://doi.org/10.1111/moth.12682
189. Stadelmann, T., Braschler, M., Stockinger, K.: Introduction to applied data science. In: Applied data science: lessons learned for the data-driven business, pp. 3–16. Springer (2019). https://doi.org/10.1007/978-3-030-11821-1_1
190. Brooks, R.: The seven deadly sins of predicting the future of AI. https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai Accessed 2023-08-22
191. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired Artificial Intelligence. Neuron **95**(2), 245–258 (2017). https://doi.org/10.1016/j.neuron.2017.06.011
192. Huerta, M.F., Koslow, S.H., Leshner, A.I.: The human brain project: an international resource. Trends Neurosci. **16**(11), 436–438 (1993). https://doi.org/10.1016/0166-2236(93)90069-X
193. Waldrop, M.M.: Computer modelling: brain in a box. Nature **482**(7386), 456–458 (2012). https://doi.org/10.1038/482456a
194. Prescott, T.J., Camilleri, D.: The synthetic psychology of the self. In: Aldinhas Ferreira, M.I., Silva Sequeira, J., Ventura, R. (eds.) Cognitive Architectures, pp. 85–104. Springer (2019). https://doi.org/10.1007/978-3-319-97550-4_7
195. Schmidgall, S., Achterberg, J., Miconi, T., Kirsch, L., Ziaei, R., Hajiseyedrazi, S.P., Eshraghian, J.: Brain-inspired learning in artificial neural networks: a review. Preprint (2023). https://doi.org/10.48550/arXiv.2305.11252
196. Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., Hinton, G.: Backpropagation and the brain. Nat. Rev. Neurosci. **21**(6), 335–346 (2020). https://doi.org/10.1038/s41583-020-0277-3
197. Ullman, S.: Using neuroscience to develop artificial intelligence. Science **363**(6428), 692–693 (2019). https://doi.org/10.1126/science.aau6595
198. Müller, O., Liggieri, K.: Mensch-Maschine-Interaktion seit der Antike: Imaginationsräume, Narrationen und Selbstverständnisdiskurse. In: Liggieri, K., Müller, O. (eds.) Mensch-Maschine-Interaktion: Handbuch zu Geschichte, Kultur, Ethik, pp. 3–14. J.B. Metzler (2019)
199. Jank, M.: Der Homme Machine des 21. Jahrhunderts: Von Lebendigen Maschinen Im 18. Jahrhundert zur Humanoiden Robotik der Gegenwart. Brill Fink (2014). https://doi.org/10.30965/9783846756577
200. Dürr, O.: Transhumanismus—Traum Oder Alptraum? Herder (2023)
201. Sarasin, P.: Reizbare Maschinen: Eine Geschichte des Körpers 1765–1914. Suhrkamp (2001)
202. Bray, D.: Wetware: A Computer in Every Living Cell. Yale University Press (2011)
203. Clark, A.: Pressing the flesh: a tension in the study of the embodied, embedded mind? Philosophy Phenomenol. Res. **76**(1), 37–59 (2008). https://doi.org/10.1111/j.1933-1592.2007.00114.x
204. Weizenbaum, J.: Computer Power and Human Reason: From Judgement to Calculation. W.H.Freeman & Co Ltd (1976)
205. Rescorla, M.: The Computational Theory of Mind. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Fall, 2020th edn. Stanford University, Metaphysics Research Lab (2020)
206. Turing, A.: Computing machinery and intelligence. Mind **LIX**(236), 433–460 (1950). https://doi.org/10.1093/mind/LIX.236.433
207. McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (1955). http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
208. Putnam, H.: Minds & machines. In: Hook, S. (ed.) Dimensions of Mind, pp. 138–164. Collier Books (1960)
209. Fodor, J.A.: The Language of Thought. Harvard University Press (1975)
210. Heil, J.: Philosophy of Mind: A Contemporary Introduction, 4th edn. Routledge (2020)
211. Pitt, D.: Mental Representation. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy, Fall, 2022nd edn. Stanford University, Metaphysics Research Lab (2022)
212. Churchland, P.S.: Touching a Nerve: The Self as Brain. W. W. Norton & Company (2013)
213. Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences **36**(3), 181–204 (2013). https://doi.org/10.1017/S0140525X12000477
214. Kenny, A.: The Legacy of Wittgenstein. Oxford University Press (1984)
215. Bennett, M.R., Hacker, P.M.S.: Philosophical Foundations of Neuroscience. Wiley (2022)
216. Hagner, M.: Homo Cerebralis: Der Wandel Vom Seelenorgan zum Gehirn. Suhrkamp (1997)
217. Fuchs, T.: In Defence of the Human Being: Foundational Questions of an Embodied Anthropology. Oxford University Press, UK (2021)
218. Dreyfus, H., Taylor, C.: Retrieving Realism. Harvard University Press (2015)
219. Dennett, D.C.: Philosophy as naive anthropology: Comment on bennett and hacker. In: Bennett, M., Dennett, D.C., Hacker,

P.M.S., Searle, J.R.. (eds.) Neuroscience and Philosophy: Brain, Mind, and Language, pp. 73–96. Columbia University Press (2007). http://www.jstor.org/stable/10.7312/benn14044

220. Searle, J.: Putting consciousness back in the brain. In: Bennett, M., Dennett, D.C., Hacker, P.M.S., Searle, J.R.. (eds.) Neuroscience and Philosophy: Brain, Mind, and Language, pp. 97–124. Columbia University Press (2007). https://www.jstor.org/stable/10.7312/benn14044.7
221. Smit, H., Hacker, P.M.: Seven misconceptions about the mereological fallacy: a compilation for the perplexed. Erkenntnis **79**, 1077–1097 (2014). https://doi.org/10.1007/s10670-013-9594-5
222. Fuchs, T.: Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind. Oxford University Press (2018)
223. Janich, P.: Kein Neues Menschenbild: Zur Sprache der Hirnforschung. Suhrkamp Verlag (2009)
224. Dennett, D.C.: Intuition Pumps and Other Tools for Thinking. WW Norton & Company (2013)
225. Frank, M.: Self-consciousness and self-knowledge: On some difficulties with the reduction of subjectivity. Constellations **9**(3), 390–408 (2002). https://doi.org/10.1111/cons.2002.9.issue-3
226. Frank, M.: Non-objectal subjectivity. J. Consciousness Stud. **14**(5–6), 152–173 (2007)
227. Zahavi, D.: Thinking about (self-)consciousness: Phenomenological perspectives. In: Kriegel, U., Williford, K. (eds.) Self-Representational Approaches to Consciousness, pp. 273–296. MIT Press (2006)
228. Dennett, D.C.: The Intentional Stance. MIT Press (1989)
229. Gallagher, S.: Interpretations of embodied cognition. In: Tschacher, W., Bergomi, C. (eds.) The Implications of Embodiment: Cognition and Communication, pp. 59–70. Imprint Academic (2011)
230. Merleau-Ponty, M.: The child's relation with others. In: Edie, J.M. (ed.) The Primacy of Perception, pp. 96–155. Northwestern University Press (1964)
231. Moyal-Sharrock, D.: Certainty in Action: Wittgenstein on Language. Bloomsbury Publishing, Mind and Epistemology (2021)
232. Jonas, H.: The Phenomenon of Life. Toward a Philosophical Biology. Harper & Row (1966)
233. Thompson, E., Stapleton, M.: Making sense of sense-making: reflections on enactive and extended mind theories. Topoi **28**, 23–30 (2009)
234. Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., Hu, Z.: Language Models Meet World Models: Embodied Experiences Enhance Language Models. Preprint (2023). https://doi.org/10.48550/arXiv.2305.10626
235. Hoff, J.: Verteidigung des Heiligen: Anthropologie der Digitalen Transformation. Herder (2021)
236. Haslam, N.: Dehumanization: an integrative review. Personality Soc. Psychol. Rev. **10**(3), 252–264 (2006). https://doi.org/10.1207/s15327957pspr1003_4
237. Li, M., Leidner, B., Castano, E.: 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 (2014)
238. Kuljian, O.R., Hohman, Z.P.: Warmth, competence, and subtle dehumanization: Comparing clustering patterns of warmth and competence with animalistic and mechanistic dehumanization. Br. J. Social Psychol. **62**(1), 181–196 (2023). https://doi.org/10.1111/bjso.12565
239. Todorov, T.: Hope and Memory: Lessons From the Twentieth Century. Princeton University Press (2016)
240. Courchamp, F., Mizrahi, L., Morin, C., Courchamp, F., Bernard, J., Lambert, O.: Eine überschätzte Spezies. https://www.arte.tv/de/videos/RC-014177/eine-ueberschaetzte-spezies/ Accessed 2023-08-22
241. Pitt, J.C.: "Guns don't kill, people kill": Values in and/or around technologies. In: Kroes, P., Verbeek, P.-P. (eds.) The Moral Status of Technical Artefacts, pp. 89–101. Springer (2014). https://doi.org/10.1007/978-94-007-7914-3_6
242. Brey, P.: Artifacts as social agents. In: Harbers, H. (ed.) Inside the Politics of Technology: Agency and Normativity in the Coproduction of Technology and Society, pp. 61–84. Amsterdam University Press (2005). http://www.jstor.org/stable/j.ctt45kcv7.6
243. Miller, B.: Is technology value-neutral? Sci. Technol. Hum. Values **46**(1), 53–80 (2021). https://doi.org/10.1177/01622439199009
244. Kroes, P., Verbeek, P.-P.: Introduction: The moral status of technical artefacts. In: Kroes, P., Verbeek, P.-P. (eds.) The Moral Status of Technical Artefacts, pp. 1–9. Springer (2014). https://doi.org/10.1007/978-94-007-7914-3_1
245. Jenkins, R., Hammond, K., Spurlock, S., Gilpin, L.: Separating facts and evaluation: motivation, account, and learnings from a novel approach to evaluating the human impacts of machine learning. AI Soc. **38**, 1415–1428 (2023). https://doi.org/10.1007/s00146-022-01417-y
246. Ihde, D.: Technology and the Lifeworld: From Garden to Earth. Indiana University Press (1990)
247. Hughes, T.P.: 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, pp. 51–82. MIT Press (1987)
248. Murphie, A., Potts, J.: Culture and Technology. Bloomsbury Publishing (2017)
249. Grunwald, A.: Technology assessment: Concepts and methods. In: Meijers, A. (ed.) Philosophy of Technology and Engineering Sciences. Handbook of the Philosophy of Science, pp. 1103–1146. North-Holland (2009). https://doi.org/10.1016/B978-0-444-51667-1.50044-6
250. Grunwald, A.: Technology Assessment in Practice and Theory. Routledge (2019)
251. Winner, L.: The Whale and the Reactor. A Search for Limits in an Age of High Technology, 2nd edn. University of Chicago Press (2020)
252. McLuhan, M.: Understanding Media. The Extensions of Man. MIT Press (1994 [1964])
253. Postman, N.: Media ecology education. Explorations Media Ecol. **5**(1), 5–14 (2006). https://doi.org/10.1386/eme.5.1.5_1
254. Strate, L.: Media Ecology. Peter Lang Press, An Approach to Understanding the Human Condition. Understanding Media Ecology (2017)
255. Cali, D.D.: Mapping Media Ecology. Peter Lang Verlag (2017). https://doi.org/10.3726/978-1-4539-1871-5
256. Ihde, D.: Postphenomenology: Essays in the Postmodern Context. Northwestern University Press (1995)
257. Verbeek, P.-P.: What Things Do. Agency, and Design. Pennsylvania State University Press, Philosophical Reflections on Technology (2005)
258. Rosenberger, R., Verbeek, P. (eds.): Postphenomenological Investigations: Essays on Human-Technology Relations. Lexington Books (2015)
259. Latour, B.: We Have Never Been Modern. Harvard University Press (2012)
260. Sharon, T.: Human Nature in an Age of Biotechnology: The Case for Mediated Posthumanism. Philosophy of Engineering and Technology, vol. 14. Springer (2013)
261. Sismondo, S.: An Introduction to Science and Technology Studies. Wiley-Blackwell (2010)
262. Felt, U., Fouché, R., Miller, C.A., Smith-Doerr, L.: The Handbook of Science and Technology Studies, 4th edn. MIT Press (2017)

263. Verbeek, P.-P.: Beyond interaction: a short introduction to mediation theory. Interactions **22**(3), 26–31 (2015). https://doi.org/10.1145/2751314
264. Stiegler, B.: What Makes Life Worth Living: On Pharmacology. Wiley (2013)
265. Kitchin, R., Dodge, M.: Code/Space: Software and Everyday Life. Software Studies. MIT Press (2011). https://doi.org/10.7551/mitpress/9780262042482.001.0001
266. Heidenreich, F., Weber-Stein, F.: The Politics of Digital Pharmacology: Exploring the Craft of Collective Care. Transcript Verlag (2022)
267. Karanasiou, A.P., Pinotsis, D.A.: A study into the layers of automated decision-making: Emergent normative and legal aspects of Deep Learning. Int. Rev. Law Comput. Technol. **31**(2), 170–187 (2017). https://doi.org/10.1080/13600869.2017.1298499
268. Prunkl, C.: Human autonomy in the age of Artificial Intelligence. Nat. Mach. Intell. **4**(2), 99–101 (2022). https://doi.org/10.1038/s42256-022-00449-9
269. Leroi-Gourhan, A.: Gesture and Speech. MIT Press (1993)
270. Noë, A.: The Entanglement: How Art and Philosophy Make Us What We Are. Princeton University Press (2023)
271. Ellul, J.: The Technological Society. Vintage (2021 [1954])
272. Grunwald, A.: Converging technologies: visions, increased contingencies of the conditio humana, and search for orientation. Futures **39**(4), 380–392 (2007). https://doi.org/10.1016/j.futures.2006.08.001
273. Merleau-Ponty, M., Smith, C.: Phenomenology of Perception. Routledge (1962)
274. Polanyi, M.: The Tacit Dimension: Michael Polanyi. Routledge & Kegan Paul (1967)
275. Stiegler, B.: Technics and Time, 1: The Fault of Epimetheus. Stanford University Press (1998)
276. Spiekermann, S.: Value-Based Engineering: A Guide to Building Ethical Technology for Humanity. De Gruyter (2023)
277. Varela, F.J., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press (1992)
278. Thompson, E.: Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press (2010)
279. Di Paolo, E., Buhrmann, T., Barandiaran, X.: Sensorimotor Life: An Enactive Proposal. Oxford University Press (2017)
280. Hutto, D.D., Myin, E.: Radicalizing Enactivism: Basic Minds Without Content. MIT Press (2012)
281. Stewart, J., Gapenne, O., Di Paolo, E.A. (eds.): Enaction: Toward a New Paradigm for Cognitive Science. MIT Press (2010)
282. Gallagher, S.: Enactivist Interventions: Rethinking the Mind. Oxford University Press (2017). https://doi.org/10.1093/oso/9780198794325.001.0001
283. Ward, D., Silverman, D., Villalobos, M.: Introduction: the varieties of enactivism. Topoi **36**, 365–375 (2017). https://doi.org/10.1007/s11245-017-9484-6
284. Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., Semenuks, A.: What happened to cognitive science? Nat. Hum. Behav. **3**(8), 782–791 (2019). https://doi.org/10.1038/s41562-019-0626-2
285. Andler, D.: Philosophy of cognitive science. In: French Studies in the Philosophy of Science: Contemporary Research in France, pp. 255–300. Springer (2009)
286. Wilson, A.D., Golonka, S.: Embodied cognition is not what you think it is. Front. Psychol. **4**, 58 (2013). https://doi.org/10.3389/fpsyg.2013.00058
287. Margolis, E., Samuels, R., Stich, S.P.: The Oxford Handbook of Philosophy of Cognitive Science. Oxford University Press (2012)
288. Rowlands, M.: Enactivism and the extended mind. Topoi **28**, 53–62 (2009). https://doi.org/10.1007/s11245-008-9046-z
289. Cappuccio, M.L.: Mind-upload. the ultimate challenge to the embodied mind theory. Phenomenol. Cognit. Sci. **16**, 425–448 (2017). https://doi.org/10.1007/s11097-016-9464-0
290. Gallagher, S.: The extended mind: state of the question. Southern J. Philosophy **56**(4), 421–447 (2018). https://doi.org/10.1111/sjp.12308
291. Hohwy, J.: The Predictive Mind. Oxford University Press (2013)
292. Clark, A.: Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press (2016)
293. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., VanRullen, R.: Consciousness in artificial intelligence: insights from the science of consciousness. Preprint (2023). https://doi.org/10.48550/arXiv.2308.08708
294. Nagel, T.: The View From Nowhere. Oxford University Press (1989)
295. Turner, J.S.: Purpose & Desire: What Makes Something "Alive" and Why Modern Darwinism Has Failed to Explain It. Harper One (2017)
296. Noble, R., Noble, D.: Understanding Living Systems. Cambridge University Press (2023)
297. Fuchs, T.: The circularity of the embodied mind. Frontiers in Psychology **11** (2020). https://doi.org/10.3389/fpsyg.2020.01707
298. Coenen, C., Grunwald, A.: Responsible research and innovation (rri) in quantum technology. Ethics Inform. Technol. **19**, 277–294 (2017). https://doi.org/10.1007/s10676-017-9432-6
299. Friedman, B., Hendry, D.G.: Value Sensitive Design: Shaping Technology With Moral Imagination. MIT Press (2019)
300. Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., Vasilakos, A.V.: Privacy and security issues in deep learning: a survey. IEEE Access **9**, 4566–4593 (2021). https://doi.org/10.1109/ACCESS.2020.3045078
301. Véliz, C.: Privacy Is Power. Melville House (2021)
302. Curzon, J., Kosa, T.A., Akalu, R., El-Khatib, K.: Privacy and Artificial Intelligence. IEEE Trans. Artificial Intell. **2**(2), 96–108 (2021). https://doi.org/10.1109/TAI.2021.3088084
303. Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Computing Surveys **55**(2), (2022). https://doi.org/10.1145/3491209
304. Wing, J.M.: Trustworthy AI. Commun. ACM **64**(10), 64–71 (2021). https://doi.org/10.1145/3448248
305. Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., Yeung, K.: Trustworthy AI. In: Braunschweig, B., Ghallab, M. (eds.) Reflections on Artificial Intelligence for Humanity, pp. 13–39. Springer (2021). https://doi.org/10.1007/978-3-030-69128-8_2
306. Durán, J.M., Formanek, N.: Grounds for trust: essential epistemic opacity and computational reliabilism. Minds Mach. **28**(4), 645–666 (2018). https://doi.org/10.1007/s11023-018-9481-6
307. Floridi, L.: Establishing the rules for building trustworthy AI. Nat. Mach. Intell. **1**(6), 261–262 (2019). https://doi.org/10.1038/s42256-019-0055-y
308. Krüger, S., Wilson, C.: The problem with trust: on the discursive commodification of trust in AI. AI & Society, 1753–1761 (2023). https://doi.org/10.1007/s00146-022-01401-6
309. Yazdanpanah, V., Gerding, E.H., Stein, S., Dastani, M., Jonker, C.M., Norman, T.J., Ramchurn, S.D.: Reasoning about responsibility in autonomous systems: challenges and opportunities. AI Soc. **38**(4), 1453–1464 (2023). https://doi.org/10.1007/s00146-022-01607-8
310. Johansen, J., Pedersen, T., Johansen, C.: Studying human-to-computer bias transference. AI Soc. **38**, 1659–1683 (2023). https://doi.org/10.1007/s00146-021-01328-4

311. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: From principles to practices. ACM Computing Surveys **55**(9), (2023). https://doi.org/10.1145/3555803
312. Coeckelbergh, M.: AI Ethics. MIT Press (2020)
313. Spiekermann, S.: Digitale Ethik: Ein Wertesystem Für Das 21. Jahrhundert, Droemer (2019)
314. Dubber, M.D., Pasquale, F., Das, S.: The Oxford Handbook of Ethics of AI. Oxford University Press (2020). https://doi.org/10.1093/oxfordhb/9780190067397.001.0001
315. Véliz, C. (ed.): The Oxford Handbook of Digital Ethics. Oxford University Press (2023). https://doi.org/10.1093/oxfordhb/9780198857815.001.0001
316. Glüge, S., Amirian, M., Flumini, D., Stadelmann, T.: How (not) to measure bias in face recognition networks. In: Schilling, F.-P., Stadelmann, T. (eds.) Artificial Neural Networks in Pattern Recognition, pp. 125–137. Springer (2020). https://doi.org/10.1007/978-3-030-58309-5_10
317. Loi, M., Heitz, C., Ferrario, A., Schmid, A., Christen, M.: Towards an ethical code for data-based business. In: 6th Swiss Conference on Data Science (SDS), pp. 6–12 (2019). https://doi.org/10.1109/SDS.2019.00-15
318. Baumann, J., Heitz, C.: Group fairness in prediction-based decision making: From moral assessment to implementation. In: 2022 9th Swiss Conference on Data Science (SDS), pp. 19–25 (2022). IEEE
319. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.O., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Preprint (2018). https://doi.org/10.48550/arXiv.1802.07228
320. Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A.S., Maharaj, T., Sherwin, E.D., Mukkavilli, S.K., Kording, K.P., Gomes, C.P., Ng, A.Y., Hassabis, D., Platt, J.C., Creutzig, F., Chayes, J., Bengio, Y.: Tackling climate change with machine learning. ACM Comput. Surv. **55**(2) (2022). https://doi.org/10.1145/3485128
321. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. Preprint (2019). https://doi.org/10.48550/arXiv.1906.02243
322. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 610–623. Association for Computing Machinery (2021). https://doi.org/10.1145/3442188.3445922
323. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al.: An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
324. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin Books (2019)
325. Weinberg, A.M.: Can technology replace social engineering? Bull. Atomic Sci. **22**(10), 4–8 (1966). https://doi.org/10.1080/00963402.1966.11454993
326. Morozov, E.: To save everything, click here. J. Inf. Policy (2014)
327. Baum, S.D.: Reconciliation between factions focused on near-term and long-term Artificial Intelligence. AI Soc. **33**(4), 565–572 (2018). https://doi.org/10.1007/s00146-017-0734-3
328. Schaeffer, R., Miranda, B., Koyejo, S.: Are Emergent Abilities of Large Language Models a Mirage? Preprint (2023). https://doi.org/10.48550/arXiv.2304.15004
329. Bengio, Y., et al.: Pause giant AI experiments: an open letter. Future of Life Institute Open Letter, https://futureoflife.org/open-letter/pause-giant-ai-experiments (2023)
330. Prabhakaran, V., Mitchell, M., Gebru, T., Gabriel, I.: A Human rights-based approach to responsible AI. Preprint (2022). https://doi.org/10.48550/arXiv.2210.02667
331. Gill, K.S.: Seeing beyond the lens of platonic embodiment. AI Soc. **38**(4), 1261–1266 (2023). https://doi.org/10.1007/s00146-023-01711-3
332. Bostrom, N.: Existential risk prevention as global priority. Global Policy **4**(1), 15–31 (2013)
333. Greaves, H., MacAskill, W.: The case for strong longtermism. Technical report, Global Priorities Institute, University of Oxford (2021)
334. Grunwald, A., Nordmann, A., Sand, M. (eds.): Hermeneutics, History, and Technology: The Call of the Future. Routledge (2023). https://doi.org/10.4324/9781003322290
335. Sotala, K., Gloor, L.: Superintelligence as a cause or cure for risks of astronomical suffering. Informatica **41**(4), (2017)
336. Spaemann, R.: Personen. Klett-Cotta (2006)
337. Taylor, C.: The Language Animal: The Full Shape of the Human Linguistic Capacity. Harvard University Press (2016)
338. Piantadosi, S.T., Hill, F.: Meaning without reference in large language models. Preprint (2022). https://doi.org/10.48550/arXiv.2208.02957
339. Brodie, M.L.: What is data science? In: Braschler, M., Stadelmann, T., Stockingers, K. (eds.) Applied Data Science: Lessons Learned for the Data-Driven Business, pp. 101–130. Springer (2019). https://doi.org/10.1007/978-3-030-11821-1_8
340. Reutlinger, A., Saatsi, J. (eds.): Explanation Beyond Causation: Philosophical Perspectives on Non-causal Explanations. Oxford University Press (2018)
341. Goldman, A., Beddor, B.: Reliabilist Epistemology. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Summer, 2021st edn. Stanford University, Metaphysics Research Lab (2021)
342. Eisenstein, M., et al.: Artificial Intelligence powers protein-folding predictions. Nature **599**(7886), 706–708 (2021). https://doi.org/10.1038/d41586-021-03499-y
343. Grimm, S.: Understanding. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Summer, 2021st edn. Stanford University, Metaphys. Res. Lab (2021)
344. Heidegger, M.: Being and Time. Suny Press (1996 [1926])
345. Dreyfus, H.L., Wrathall, M.A.: Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action. Oxford University Press (2014). https://doi.org/10.1093/acprof:oso/9780199654703.001.0001
346. Sellars, W.S.: Philosophy and the scientific image of man. In: Colodny, R. (ed.) Science, Perception, and Reality, pp. 35–78. Humanities Press (1962)
347. Rouse, J.: Articulating the World: Conceptual Understanding and the Scientific Image. University of Chicago Press (2019)
348. Odling-Smee, F.J., Lala, K.N., Feldman, M.: Niche Construction: The Neglected Process in Evolution. Princeton University Press (2003)
349. Wagner, B.: Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. Policy Internet **11**(1), 104–122 (2019). https://doi.org/10.1002/poi3.198
350. Kaun, A.: Suing the algorithm: the mundanization of automated decision-making in public services through litigation. Inform. Commun. Soc. **25**(14), 2046–2062 (2022). https://doi.org/10.1080/1369118X.2021.1924827
351. Calvo, R.A., Peters, D., Vold, K., Ryan, R.M.: 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, pp. 31–54. Springer (2020). https://doi.org/10.1007/978-3-030-50585-1_2

352. Stiegler, B.: Automatic Society, Volume 1: The Future of Work. John Wiley & Sons (2018)
353. Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press (2016). https://doi.org/10.1093/acprof:oso/9780190498511.001.0001
354. Kanner, A.D.: Technological wisdom. ReVision **20**(4), 45–46 (1998)
355. Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N., Trench, M.: Artificial Intelligence: the next digital frontier? McKinsey Global Institute (2017)
356. Stadelmann, T.: Wie maschinelles Lernen den Markt verändert. In: Haupt, R., Schmitz, S. (eds.) Digitalisierung: Datenhype Mit Werteverlust?: Ethische Perspektiven Für Eine Schlüsseltechnologie, pp. 67–79. SCM Hänssler (2019)
357. Tricot, R.: Venture capital investments in Artificial Intelligence. OECD Digital Economy Papers (319), (2021). https://doi.org/10.1787/f97beae7-en
358. Zuboff, S.: The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power. Public Affairs (2019)
359. Zuboff, S.: The age of surveillance capitalism. In: Longhofer, W., Winchester, D. (eds.) Social Theory Re-Wired, pp. 203–213. Routledge (2023)
360. Bruni, L., Sugden, R.: Reclaiming virtue ethics for economics. J. Econ. Perspect. **27**(4), 141–164 (2013). https://doi.org/10.1257/jep.27.4.141
361. Bruni, L., Héjj, T.: The economy of communion. In: Handbook of Spirituality and Business, pp. 378–386. Springer (2011). https://doi.org/10.1057/9780230321458_45
362. Keilty, P.: Desire by design: pornography as technology industry. Porn Stud. **5**(3), 338–342 (2018). https://doi.org/10.1080/23268743.2018.1483208
363. Kergel, D., Paulsen, M., Garsdal, J., Heidkamp-Kergel, B. (eds.): Bildung in the Digital Age. Routledge (2022)
364. Coeckelbergh, M.: The Political Philosophy of AI: An Introduction. Wiley (2022)
365. Sattarov, F.: Power and Technology: A Philosophical and Ethical Analysis. Rowman & Littlefield (2019)
366. Lewis, C.S.: The Abolition of Man. Oxford University Press (1943)
367. Crawford, K., Paglen, T.: Excavating AI: the politics of images in machine learning training sets. AI Soc. **36**, 1399 (2021). https://doi.org/10.1007/s00146-021-01301-1
368. Kane, T.B.: Artificial Intelligence in politics: establishing ethics. IEEE Technol. Soc. Mag. **38**(1), 72–80 (2019). https://doi.org/10.1109/MTS.2019.2894474
369. Sætra, H.S.: A typology of AI applications in politics. In: Visvizi, A., Bodziany, M. (eds.) Artificial Intelligence and Its Contexts: Security, Business and Governance, pp. 27–43. Springer (2021). https://doi.org/10.1007/978-3-030-88972-2_3
370. Marwala, T.: Artificial Intelligence in politics. In: Artificial Intelligence, Game Theory and Mechanism Design in Politics, pp. 41–58. Springer (2023). https://doi.org/10.1007/978-981-99-5103-1_4
371. Ienca, M.: On Artificial Intelligence and manipulation. Topoi **42**, 833–842 (2023). https://doi.org/10.1007/s11245-023-09940-3
372. Bishop, J.: Elster, j.: "sour grapes: Studies in the subversion of rationality". Australasian J. Philosophy **63**, 245 (1985)
373. Fogg, B.J.: Persuasive technology: Using computers to change what we think and do. Ubiquity (2002). https://doi.org/10.1145/764008.763957
374. Wilson, D.G.: The ethics of automated behavioral microtargeting. AI Matters **3**(3), 56–64 (2017). https://doi.org/10.1145/3137574.3139451
375. Zuiderveen Borgesius, F.J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., de Vreese, C.: Online political microtargeting: Promises and threats for democracy. Utrecht Law Review (2018). https://doi.org/10.18352/ulr.420
376. Susser, D.: Invisible influence: Artificial Intelligence and the ethics of adaptive choice architectures. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES '19, pp. 403–408. Association for Computing Machinery (2019). https://doi.org/10.1145/3306618.3314286
377. Milano, S., Taddeo, M., Floridi, L.: Recommender systems and their ethical challenges. AI Soc. **35**, 957–967 (2020). https://doi.org/10.1007/s00146-020-00950-y
378. Susser, D., Roessler, B., Nissenbaum, H.: Technology, autonomy, and manipulation. Internet Policy Rev. **8**(2), (2019). https://doi.org/10.14763/2019.2.1410
379. Mele, C., Russo Spena, T., Kaartemo, V., Marzullo, M.L.: Smart nudging: How cognitive technologies enable choice architectures for value co-creation. J. Business Res. **129**, 949–960 (2021). https://doi.org/10.1016/j.jbusres.2020.09.004
380. Ashton, H., Franklin, M.: The problem of behaviour and preference manipulation in AI systems. In: Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022), vol. 3087 (2022). CEUR Workshop Proceedings. https://discovery.ucl.ac.uk/id/eprint/10146136
381. Simchon, A., Edwards, M., Lewandowsky, S.: The persuasive effects of political microtargeting in the age of generative AI. Preprint (2023). https://doi.org/10.31234/osf.io/62kxq
382. Smith, J., de Villiers-Botha, T.: Hey, google, leave those kids alone: against hypernudging children in the age of big data. AI Soc. **38**, 1639–1649 (2023). https://doi.org/10.1007/s00146-021-01314-w
383. Carroll, M., Chan, A., Ashton, H., Krueger, D.: Characterizing Manipulation from AI Systems. Preprint (2023). https://doi.org/10.48550/arXiv.2303.09387
384. Berghel, H.: Malice domestic: The cambridge analytica dystopia. Computer **51**(5), 84–89 (2018). https://doi.org/10.1109/MC.2018.2381135
385. Geller, A.: Social Scoring durch Staaten. PhD thesis, Ludwig-Maximilians-Universität, München (2022)
386. Heinrichs, B., Heinrichs, J.-H., Rüther, M.: Künstliche Intelligenz. De Gruyter (2022). https://doi.org/10.1515/9783110746433
387. Berk, R.A.: Artificial Intelligence, predictive policing, and risk assessment for law enforcement. Ann. Rev. Criminol. **4**(1), 209–237 (2021). https://doi.org/10.1146/annurev-criminol-051520-012342
388. Awotunde, J.B., Misra, S., Ayeni, F., Maskeliunas, R., Damasevicius, R.: Artificial Intelligence based system for bank loan fraud prediction. In: Abraham, A., Siarry, P., Piuri, V., Gandhi, N., Casalino, G., Castillo, O., Hung, P. (eds.) Hybrid Intelligent Systems, pp. 463–472. Springer (2022). https://doi.org/10.1007/978-3-030-96305-7_43
389. Turiel, J., Aste, T.: Peer-to-peer loan acceptance and default prediction with Artificial Intelligence. R. Soc. Open Sci. **7**(6), 191649 (2020). https://doi.org/10.1098/rsos.191649
390. Rong, G., Mendez, A., Bou Assi, E., Zhao, B., Sawan, M.: Artificial Intelligence in healthcare: review and prediction case studies. Engineering **6**(3), 291–301 (2020). https://doi.org/10.1016/j.eng.2019.08.015
391. Yang, C.C.: Explainable Artificial Intelligence for predictive modeling in healthcare. J. Healthcare Inform. Res. **6**(2), 228–239 (2022). https://doi.org/10.1007/s41666-022-00114-1
392. Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., Biancone, P.: The role of Artificial Intelligence in healthcare: a

structured literature review. BMC Med. Inform. Decision Making **21**, 125 (2021). https://doi.org/10.1186/s12911-021-01488-9

393. Vallès-Peris, N., Domènech, M.: Caring in the in-between: a proposal to introduce responsible AI and robotics to healthcare. AI Soc. **38**(4), 1685–1695 (2023). https://doi.org/10.1007/s00146-021-01330-w
394. Ford, K.M., Hayes, P.J., Glymour, C., Allen, J.: Cognitive orthoses: toward human-centered AI. AI Mag. **36**(4), 5–8 (2015). https://doi.org/10.1609/aimag.v36i4.2629
395. Uhl, A.: Extended intelligence: Awareness-based interventions into the ecology of autonomous and intelligent systems. PhD thesis, Harvard University Graduate School of Arts and Sciences (2021). https://dash.harvard.edu/handle/1/37368514
396. Karachalios, K., Ito, J.: Human intelligence and autonomy in the era of 'extended intelligence'. Council on Extended Intelligence (2018)
397. Council on Extended Intelligence: Our Vision (2021). https://globalcxi.org/vision/
398. Clark, A., Chalmers, D.: The extended mind. Analysis **58**(1), 7–19 (1998). Accessed 2023-08-24
399. Ito, J.: Resisting Reduction: A Manifesto. Journal of Design and Science (2017)
400. Aurum, A., Biffl, S., Boehm, B., Erdogmus, H., Grünbacher, P.: Value-Based Software Engineering. Springer (2005)
401. Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A.: Value sensitive design and information systems. In: Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E. (eds.) Early Engagement and New Technologies: Opening up the Laboratory, pp. 55–95. Springer (2013). https://doi.org/10.1007/978-94-007-7844-3_4
402. Spiekermann, S.: Ethical It Innovation: A Value-Based System Design Approach. CRC Press (2015)
403. Spiekermann, S., Winkler, T.: Value-based engineering with IEEE 7000. IEEE Technol. Soc. Mag. **41**(3), 71–80 (2022). https://doi.org/10.1109/MTS.2022.3197116
404. Shneiderman, B.: Human-Centered AI. Oxford University Press (2022)
405. Herrmann, T., Pfeiffer, S.: Keeping the organization in the loop: a socio-technical extension of human-centered Artificial Intelligence. AI Soc. **38**(4), 1523–1542 (2023). https://doi.org/10.1007/s00146-022-01391-5