A guide to AI: Understanding the technology, applying it successfully, and shaping a positive future

Thilo Stadelmann

Abstract If artificial intelligence were a road, what would the signs look like? 'No Speed Limit' / 'Caution, Danger Ahead' / 'Toll Station Ahead' / 'Start of Multi-Lane Construction'? Such narratives are indeed told about AI, sometimes contradicting each other. Which ones you trust is crucial for the future – in business, society, and private life. Therefore, we take an analytical look at the state of this road and clear up the forest of signs. We first explore the scientific, philosophical, business, and societal levels of the metaphor 'artificial intelligence' to get closer to its core. We examine the underlying technology and look at primary fields of application and frequently mentioned risks. From this, we derive concrete options for dealing with and implementing AI, and provide an outlook on our future.

1 A roadmap for the high-tech landscape

Many narratives about artificial intelligence (AI) are currently circulating: A utility revolution for the economy [1] with market potential in the trillions [2]; or potential extinction-level threats to humanity through a sudden emergence of hypothetical "artificial general intelligence" (AGI) [3] and other dystopian scenarios [4] – you name it [5]. Like too many signs in a highway construction site, this can be confusing. Which ones should you pay attention to?

This article understands AI as two things: a *powerful tool* on one hand, which can actually be understood, scientifically grasped, and advantageously used; and an *empty phrase* on the other hand, which can be filled with any content and thus fuel hype and, depending on one's worldview, stir up fears or create utopias. Both variants

Prof. Thilo Stadelmann, Dr. rer. nat.
ZHAW Centre for Artificial Intelligence, Technikumstrasse 71, CH-8400 Winterthur, Switzerland
AlpineAI AG, Obere Strasse 22b, CH-7270, Davos, Switzerland
e-mail: stdm@zhaw.ch

of meaning must be considered as they influence our economic, private, and social reality.

How did we get here? The roots of the scientific field of AI go back to the 1950s. The name "artificial intelligence" for the new discipline was chosen for monetary considerations: the founders wanted to attract substantial funding for the new science of "complex computer applications" [6]. This choice continues to provoke more extreme emotions than usual in science, and thus more extreme hype and disappointment, not the least because intelligence and everything associated with it deeply affects humans. For example, AI only emerged from hibernation in the mid-2010s, where it had been sent by public and professional punishment through neglect due to certain unfillfilled promises of the 1990s. As late as 2014, according to a study in Switzerland, only a few research groups were dedicated to the topic [7] (today there are hundreds); and the introduction of an AI curriculum [8] in a computer science program raised major concerns that same year ("Old hat, do we need this?"). At the same time, machine learning was already finding its way into businesses due to its utility [9]. This usefulness reflects the continuous progress in a field that inherently operates in an application-oriented way and thus produces useful tools [10]. The reservations, hypes, and hibernations, on the other hand, are expressions of the extreme expectations projected onto the technology by humans and society through its anthropomorphization.

In the following, we'll explore the foundations of both meanings of "AI" (Chapter 2) and explain how it works. We go on to exemplarily span the arc of today's application possibilities and illuminate risks (Chapter 3). Thus equipped with a solid basic understanding of the technology's possibilities and limitations, we will finally peer into the future (Chapter 4) —- which we as humanity will shape ourselves.

2 Background: What is artificial intelligence?

2.1 Scientific foundation

Artificial intelligence is the scientific field concerned with the *mimicking of intelligent behavior using computers*. As such, it belongs to the discipline of computer science. While human capabilities often serve as the benchmark for the quality of results, AI is neither about simulating the underlying biological processes, nor does it methodologically rely on (or contain) a unified theory of intelligence. Rather, it is a methodological *toolbox* of diverse techniques that exploit the advantages of modern computers (large memory for data on which fast, simple computational operations are performed) [11]. These methods can simulate specific, more or less isolated aspects of intelligent behavior. The definitive textbook on the complete field of AI comes from Russell & Norvig [12].

Historically, the "AI toolbox" has two major compartments. In the first one are methods often described as *knowledge-based*, which ultimately aim to process a collection of facts using incorruptible logic to arrive at new statements about the

world. The attempt to reduce all intelligent behavior to logic and thus build a world model can be considered failed [13]. Nevertheless, the underlying methods are used millions of times daily: Fast search across all combinations of individual steps to arrive at a complex goal with a suitable step sequence not only helped IBM's AI system "Deep Blue" defeat reigning chess world champion Garri Kasparov in 1997. It also enables pathfinding in our navigation systems today.

However, it is the second major compartment in the AI toolbox that is responsible for the current boom around intelligent systems. Presumably every currently widely discussed, astonishing AI system, including generative AI, is based on its methods of *machine learning* (ML). These are procedures for generating behavior that we could not satisfactorily express in rules (or program code) manually, but can describe exemplarily through examples. Take the challenge of distinguishing cats from dogs in photos as an example: It is unclear what set of rules would uniquely and correctly describe this separation. For every rule about fur or head shape, exceptions could be found. However, it is simple to compile a *dataset* of dog and cat images from which a human would easily learn the relationship (or *target function*, $f(x) = y$). Here, x would symbolize the input, i.e., an image, y would correspond to the output, say 1 for dog and -1 for cat, and $f()$ would represent the function that maps images x to so-called *labels* y.

In ML, the computer takes over this *training*: It receives a dataset of images $\mathbf{x}$ in a suitable encoding (e.g., as a vector that arranges all pixels as values in sequence, image row by image row) and a suitably adaptable function $f()$ (for example, one that can represent many different curved surfaces through choice of *parameters*). Starting from an initial (e.g., random) configuration, it successively adjusts the parameters of $f()$ to *minimize* the *error* between *predicted* y' and actual y for a given $\mathbf{x}$, and this for all $(\mathbf{x}, y)$ pairs in the dataset.

For simplicity, let's imagine this with images represented by just two pixels $\mathbf{x} = (x_1, x_2)$—making them merely points in a two-dimensional coordinate system that we can still easily visualize (see Figure 1). We can assume that dog images share some similarity among themselves and thus form a cluster in the 2D space, somewhat distant from all cat images. Therefore, we could try to separate the two clusters with a straight line—a linear equation. The target function $f()$ would then be that a dog image ends up above the line defined by $(\theta_0 + \theta_1 x_0 + \theta_2 x_1 = 0)$ and a cat image below it, thus:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \theta_0 + \theta_1 x_0 + \theta_2 x_1 \geq 0 \\ -1 & \text{if } \theta_0 + \theta_1 x_0 + \theta_2 x_1 < 0 \end{cases}$$

Here, the three parameters θ_1–θ_3 are more intuitively known as the slope ($m = \frac{\theta_1}{\theta_2}$) of the line as well as its intersection point with the vertical axis ($b = \frac{\theta_0}{\theta_2}$), and the linear equation itself as $x_2 = mx_1 + b$. By varying the (two or three) parameters alone, all possible lines in two dimensions can be realized. The goal of machine learning in this exampe is thus to find a configuration of m and b (or θ_1–θ_3) (i.e., to place the line in space by rotating and shifting it) so that ideally all $\mathbf{x}$ representing dogs ($y = 1$) end up above the line, while all cat-$\mathbf{x}$ ($y = -1$) land below. For a new image $\mathbf{x}'$,

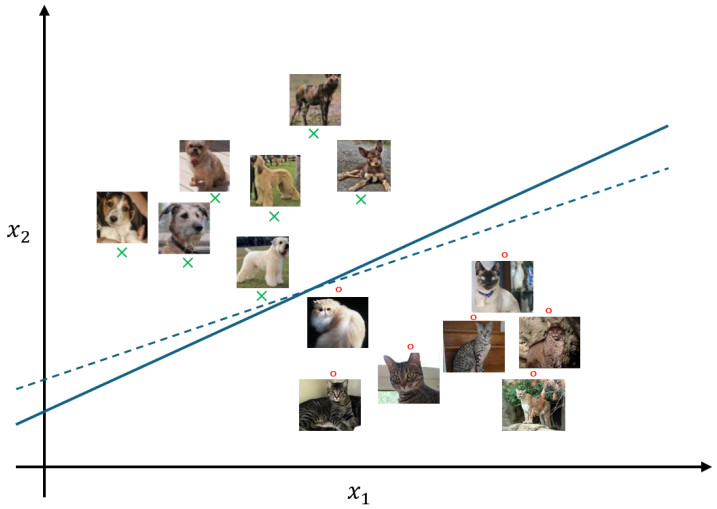

Fig. 1 Images of dogs and cats (source of individual images: ImageNet [14]), represented by only 2 coordinates (red ∘, green ×). These are perfectly separated into the two categories by the line $\theta_0+\theta_1 x_0+\theta_2 x_1 = 0$ (blue, solid line): All dog images are found above the line ($\theta_0+\theta_1 x_0+\theta_2 x_1 > 0$), the cats lie below ($\theta_0 + \theta_1 x_0 + \theta_2 x_1 < 0$). Alternative separation lines are also conceivable (such as the dashed, blue line). ML methods find the optimal parameters $\boldsymbol{\theta}$ for given encodings $\mathbf{x}$ of the data points and a predetermined function class $f()$ (such as the class of linear functions).

the trained *model* $f()$ (with ideal parameters found using the training dataset) now directly outputs whether it is predicted to be a cat ($f(\mathbf{x}') < 0$) or dog ($f(\mathbf{x}') \geq 0$).

The example above shows how a specific use case (distinguishing cats from dogs) must first be formalized for the computer (suitable encoding of input x and output y), a training dataset of (x, y) pairs must be compiled, and a suitable function type $f()$ must be chosen. The computer then takes over the systematic adjustment of the parameters of $f()$, so that the deviation between calculated output and actual label is minimized across all examples. If the training set was large enough and representative of future expected inputs, there is good reason to assume that the learned model *generalizes*. In real applications, where inputs are high-dimensional (such as images, videos, text, sensor recordings) and the relationship with the desired output is complex (nonlinear, like mapping visual appearance to biological species), one would choose a more adaptable target function than a 2D line—such as a so-called neural network, which leads to the *deep learning* variety of ML. Its up to billions of parameters ultimately provide for arbitrary "curviness" of the separation surface from above and thus for high adaptability to the data and the relationship to be modeled. However, the learning principle remains the same. For a brief introduction for non-technical readers, it is recommended to read Chapter 2 from Segessenmann et al.'s article [15]; the definitive textbook on deep learning was recently presented by Prince [16].

What properties does such a ML model have? It was learned directly from the data without much explicit prior knowledge ("data-centric" [17, 18]). This means what

wasn't in the data isn't in the model. The model further maintains a purely statistical view of the world, because the above-described type of target function optimization, "maximum likelihood," indeed implicitly estimates how well the probability distribution of predictions matches the observed distribution in the training data. This statistical view works for many applications on average across many predictions (sometimes significantly) better than what a human could achieve manually. This applies to the above dog-cat classifier just as much as to Large Language Models (LLMs) like ChatGPT, which predict the most probable next word from a context of words. However, the result in individual cases can be completely wrong, as statements about statistical plausibility—to stick with the example of language—are not statements about truth. This should be kept in mind when the next impressive AI demonstration leaves one amazed.

2.2 Philosophical and narrative context

Since technology and particularly science fiction literature have existed, humans have fantasized (in the best sense of the word) about "artificial intelligence"—not as real existing technology or science, but as a container for futuristic ideas. That such "artificial intelligence in Hollywood's science fiction sense" and AI have nothing in common except the same 23 characters in identical configuration should ideally be neither special nor concerning. For instance, the flight recorder also known as "black box" shares its name with entirely unrelated phenomena ("black boxes": things that are considered completely incomprehensible due to their complexity), yet this does not lead to serious misunderstandings. However, AI is—economically and thus socially speaking—currently a game with extremely high stakes, so that mixing the two levels of meaning in public statements becomes problematic: The projection of events from science fiction onto unrelated present-day technologies can evoke fear and uncertainty.

As an example of such a statement, take the open letter from the Future of Life Institute calling for an AI moratorium [19]: The background of the numerous signing researchers and entrepreneurs alone lends technical-scientific credibility to the concern. However, the mentioned risks and proposed solutions are based on extrapolations that are philosophically-ideologically founded and are described as "hypothetical" and "imaginary" by influential parts of the research community [20].

That such narratives around self-improving, general or superintelligent AI exist took its origin at least with the thought experiment known as "Laplace's demon". It constitutes that all true statements about the world could be derived using complete data and perfect models [21]. If (and only if) these two conditions of completeness and perfection would be fulfilled, the resulting demon could operate as a flawless predictor of the future and answer any question by logical deduction. Interestingly, Laplace's motivation in introducing the demon was merely to reduce the associated idea of complete predictability to absurdity—it is infeasible precisely because the two conditions are thwarted by the inherent uncertainty of things. Regardless, this

feasibility fantasy lives on today and finds new philosophical forms, which Émile Torres and Timnit Gebru summarize under the acronym *TESTCREAL* [22]: Materialism thus leads to belief in the singularity (the "S" in TESCREAL) and thereby to the possibility that AI could ruin humanity's distant (longtermism, "L") but glorious future in space (extropianism "E" and cosmism "C") as transhuman beings ("T"), since the leap to realizing quasi-human properties in machines could happen at any time and make the technology uncontrollable. Accordingly, rationalism and effective altruism ("R" and "EA") suggest drastic measures in the here and now. Torres ultimately speaks of these as the "eugenics of the 21st century"—which makes it clear even to non-philosophers that science fiction-fueled AI anxiety should ideally have no broad ideological basis. How widespread it nevertheless is in the companies responsible for a large part of today's AI progress is exemplarily (and drastically) shown by Leopold Aschenbrenner's blog [23]: An entire community here pays homage to a questionable idea of the future and influences policy-making worldwide on a belief basis (no scientific or technical arguments involved!) that from an outside perspective is also described as "ridiculous" [24, 25].

So what is artificial intelligence? A collection of methods on one hand, with which certain intelligent behavior can be mimicked using computers today (and better in the future); on the other hand, a term often filled with ideologically shaped content inspired by science fiction, which can generate much attention and fear. Both levels of meaning are partially intertwined in public discourse as well as in product promises.

3 Application: What can artificial intelligence do?

3.1 Business Value

AI is on everyone's lips not because of a technological revolution (if anything, it would currently be 7 years old: [26]), but due to a utility revolution [1]. For generative AI, particularly text and image generation, the potential applications in both professional and private contexts become obvious to many people after just brief experimentation. Five minutes are enough to get a first impression of the potential, and the entry barriers are incredibly low thanks to simple online access and freemium services. Of course, it is important to note (and absolutely necessary in professional use) to operate in a data-protection-compliant way even during testing, to avoid making internal information public through the prompt history. As a first approximation, only paid versions of corresponding services should be used professionally, where the terms of service usually exclude the provider from freely reusing the information contained in the prompts and replies. For enterprise-wide deployment, specialized companies like AlpineAI[1] offer more control. Alternatively, local instances can be operated [27].

[1] `https://alpineai.swiss/`

If one is lacking ideas about where generative AI could bring the greatest business value, the 100 use cases in the Harvard Business Review guide offer a good starting point [28]. Generally, it is recommended to strictly align any AI deployment (like any other technology) with real business cases: Starting from the lucrative but unsolved problems in each business area is the best way to plan where deployment would be worthwhile in principle. Through experimentation or exchange with specialists, feasibility can then be efficiently clarified in the second step[2] [29].

With all the benefits of generative AI, it is important not to neglect that AI's methodological toolbox has more to offer: For instance, as part of a data science pipeline for analyzing data points and streams [30] and thus in conjunction with databases and digitalization projects [31, 32, 33]. Here, classical machine learning methods play just as much a role [29] as deep learning, depending on the data context and task [34]. To explore the possibilities in an exemplary way, the following compilation of applications may serve, which teams from the ZHAW Centre for AI have realized together with companies in recent years: In document processing, for example, the segmentation of newspaper pages into individual articles [35], scanning sheet music into machine-readable form [36], or making technical documentation accessible [37, 38]. In medical image processing and the healthcare sector, for instance, improving image-supported diagnosis through data pooling across hospital boundaries [39], accelerating cancer diagnoses through increased automation [40], reducing artifacts in CT images that may result from involutary patient movement during imageing to acquire better images using less harmful radiation [41], or monitoring intensive care patients to avoid false alarms in nursing care [42]. In the industrial sector, work on production planning for optimized complexity management [43], production parameter estimation for better performance of photovoltaic modules [44], automatic quality control of rotating machinery [9], and process monitoring of plastic injection molding processes [45]. Additional industrial applications for image and time series analysis are discussed in [46, 47].

All mentioned AI systems were developed in direct collaboration with practice partners for everyday use within one to two years, as there were no ready-made methods or products on the market to enable the underlying business case. In particular, none of these applications could so far be replaced by ChatGPT and similar tools, although some of the systems have been in use since 2017. Custom development in partnership between SMEs and AI-experienced research partners thus was not only economically possible but also sensible[3]. However, if no research gap needs to be filled for the intended solution, results are even possible within a few weeks to months (sometimes even just days) with ROIs in the triple-digit range [51]. This is good news for any organization that has success-relevant questions for which *better*

[2] `https://data-innovation.org/innovation/` refers to an exemplary innovation process.

[3] However, research also includes a risk for failure, including for technical reasons. For example, it has been predicted [48] and later claimed [49] that moderately improved machine learning techniques would make voice recognition applications much more human-like and reliable. However, it turned out on closer inspection [50] that respective AI systems *"cheated"*: They were not able to *practically* learn from data what they could model in *principle* because current learning algorithms greedily take a path of least resistance instead building up understanding in a human-like fashion.

predictions in the broadest sense could be helpful (AI is "prediction technology": The next word given the context; product quality given a measurement; document content given the scan; object type given a photo; etc.). A basic understanding of the technology helps to conduct the initial screening of possible use cases for feasibility oneself, and to proceed to detailed clarifications with partners on the selection with good prognosis.

What kind of skills and business strategy does an organization need on top of this to become successfull with AI? Actually, not much to get started. A *fearless* but *cautious* and *determined* attitude helps. Cautious, because risks exist (see below: technical, and otherwise), but determined, since the risk of not acting (and hence missing out, also catastrophically) is real as well. Building up the mentioned basic technical understanding in-house is a very good starting point. A small team that recruits its members from an *alliance of the willing* and spans technical, application, and human factor perspectives, volunteering to start as a track group, is a proven pattern. *Strategically*, a dual approach is advisable: *top-down*, by thinking the business forward: How could our markets / environments change? Where does AI affect what we do, not just how we do it (not limited to, but including thinking about data products [52] and a suitable organizational structure [32])? What will others do—are new players from unusual directions likely? Where do we *want* to go? Which key use cases hold the largest value for being transformed using AI? LLMs can be useful in working throuh these questions as they can be prompted to act like different personas, thus offering the basis to quantifiable scenario analysis (where a population responding to a poll is simulated by such pre-prompted LLMs).

In addition, another strategic task is to create the framework conditions for the second part of the strategy—the bottom-up part. These are mainly the technical and regulatory conditions to enable every member of the organization to explore the use and potential of AI in their own sphere of influence. This second pillar is important since the sum of all the small potential gains in endless processes where only the people working on them see how to improve them hold an accumulated potential that may exceed the value of the biggest strategic business cases.

For this dual top-down/bottom-up strategy to work, certain cultural values in the organization are definitely an asset, foremost *trust* amongst individuals across organizational levels. As tasks will be transformed and roles might shift, this will also help mitigate potential ensuing volatility in individual work environments. Competitive advantage, after all, comes from the humans involved, their ingenuity and dedication, given the proper use of available tools like AI. AI officers, AI stragies, or availability of AI tools *per se* is more overhead than advantage unless the steps above are implemented (use cases identified, staff engaged, etc.).

3.2 Societal challenges

The opportunities outlined above are accompanied by risks. Many of these are serious yet manageable, particularly those based on shortcomings of current AI

systems with *impact on individuals*, stemming from technical inadequacies of the employed methodology. This includes, for example, the problem of AI bias, meaning that subgroups of society are disadvantaged through the use of an AI system, perhaps because their faces are less well recognized by facial recognition methods (this is disadvantageous when it leads to poorer treatment by police, for instance). Or the problem of adversarial attacks on image classifiers (meaning that through intentional invisible modification of images, the system's results can be manipulated almost arbitrarily). Both problems are well understood [53, 54, 55] and it is absolutely within the power of any organization using such systems to prevent their negative effects using methods of algorithmic fairness. However, the associated effort must be carried out responsibly [56].

Similarly, mitigability is true for risks that arise from the tool-like nature of AI systems in general (beyond their effect on an individual): As a tool, AI can be used for various purposes, good ones (like fighting cancer, see above) and bad ones. The latter include generating fake content (e.g., fake news, deep fakes). When considering this risk, it helps to bring to mind—and this has methodological character for similar questions as well—what is actually new about the risk and what remains constant since pre-AI times: Humans have been faking content since originals existed. What is changing is the scale. Creating arbitrarily professional-looking content is now easy and cost-effective for anyone. There are concerns that a flood of fake content could pose a serious problem for democratic processes and social cohesion. But humans are quick learners and very intolerant of feeling deceived. It is conceivable that they will quickly learn to no longer trust anonymous internet sources in the usual way, and instead help restore more importance to real, trusted relationships. As a supporting measure, AI is of course also being used to detect and filter fake content, which defuses part of the flood. This does not mean that these challenges would not pose a problem. But it puts them in perspective with other societal challenges, which, while serious, are manageable.

Many are concerned about AI's potentially negative impacts on the job market. However, experts often doubt predictions of massive unemployment: AI as a tool does not automate jobs, but individual tasks. Years ago, for example, the profession of radiologist was predicted to decline after computer vision systems achieved excellent results in X-ray diagnostics. Today, there is even a shortage of radiologists in some places, as image analysis makes up only a (small) part of a radiologist's job [57]. The way AI is successfully deployed is in team with humans, as a cognitive orthosis. It extends human capabilities, but like glasses in a pocket, it is quite worthless on its own [58].

The metaphor of AI as a tool also helps in contextualizing specific challenges in education. First, better tools are a positive development. They help us get further in life. However, fundamental skills that enable us to use tools successfully in the first place are learned *without* them. For example, we must learn to calculate to build a solid sense for quantities; and to write, as it trains thinking. The current canon of teaching material thus keeps its relevancy. Language ability, in particular, becomes increasingly important: It now even forms the user interface to computers. Added to this is the need for a basic understanding of technology for everyone, to avoid seeing

the constantly reported breakthroughs as magical, but rather to realistically assess possibilities and limitations and make them usable. As fake becomes increasingly difficult to distinguish from handwork, performance assessments are better conducted orally.

However, the question arises: How will learners motivate themselves in the future to invest the necessary time and effort when the AI solution (always tempting, but not helping the learning objective) is only a few keystrokes away? Learning is associated with pain, and we achieve a good part of the *Good Life* only through voluntarily accepting certain pain. This seems to be the greatest challenge related to AI for society: If AI ideally offers enormous new conveniences and relief, but we only achieve the life we desire if we remain active ourselves (in learning, in relationships, through meaningful activity)—how do we pull ourselves together? How do we remain human (Neil Lawrence sees the core of being human also in its limitedness, which is to be embraced as a strength [59]), when superpowers beckon [60], which are helpful when used specifically and dehumanizing when overstrained? The metaphors for AI could be pointedly supplemented with that of its duality as medicine and drug. One part of the answer will lie in AI being used as tutors, serving a teachers material in my individual learning style, and in human teachers to become master motivators, next to curators of meaning.

4 Future: How do we want to live?

AI will continue to develop in the coming years, unlikely slower than before. At the same time, our society will continue to transform. This follows from the usefulness of already the current methods, leading to them further spreading throughout the economy, and the associated shifts in money that will increasingly be gained with the help of AI (shifting value chains to other departments, different suppliers, operating on other ressources) [29]. This is even independent of any major technological progress, as even the current utility of this technology has not been fully exploited yet. If we had a choice, what should be the goal of this transformation?

4.1 Individual recommendations for action

At the individual and organizational level, it makes sense to engage with AI and integrate corresponding systems into daily life and work where sensible and appropriate. *Try it out*! Since development moves quickly, this should happen periodically: What is not good enough today might look different in a quarter.

Looking ahead, which in AI can only reasonably be warranted for at most five years, one can expect AI systems to reach the utility of capable personal assistants. This does not require AGI or superintelligence, but is rather a plausible extrapolation. It stems from the capabilities of today's LLMs, enhanced with the ability for systems

to prompt themselves for a while before returning feedback to humans (so-called "agentic workflows" [61]). Combined with insights from neuroscience, it is also conceivable that the exorbitant hunger for data and electricity, which the current AI development paradigm requires (namely, "scaling" ML systems), could be reduced by orders of magnitude [62, 63]. Corresponding ML methods could also lead to better "world models" that understand their environment less purely statistically and more in terms of cause and effect—thus developing a kind of "machine common sense" that is less susceptible to obviously nonsensical outputs. Again: without AGI.

This raises the question: What actually defines being human and having value [15, 59]? The winners of the future are certainly those people who do not let themselves be intimidated by the capabilities of a fast computer or have their identity questioned. Instead, they know what is good for them as social relational beings, and understand the importance of meaningful life content. This is supported by individually strengthening and engaging with *humanistic* thought that deals with the value and dignity of humans as such, instead of technomorphizing them [15]. This strenghening of human worth and value is something we as humans need to actively persue—and something we should also put our technology in service of. Future technological development should be guided by asking "does the intended use strengthen what makes us human at the core" (e.g., our thriving in relationships, our limitedness, our need for having resonsibility over something). In that sense, the spam filter (shielding us from ever more information overflow) becomes a much more ethical AI application than the next productivity tool (that promises superpowers and might deliver burnout).

4.2 Setting the course for society

On a societal level, as with any major change, the question of regulation arises, and rightly so. Current approaches, however, often focus on compliance and liability. The EU AI Act, for example, sets out only a few specific guidelines on use, but extensively regulates what needs to be documented and how. Returning to the traffic metaphor introduced at the beginning, it is evident that the laws surrounding the emergence of modern mobility also conveyed clear signals about the kind of world their designers wanted to live in: high-speed travel only on highways, traffic calming in residential areas, mandatory vehicle permits and driver licensing, and so on. The aim was evidently to avoid unrestricted high-speed vehicles, anonymity, or untrained drivers. A similar intention in shaping AI could mean regulating not just technology or application domains, but specifically the associated business models. This approach might help avoid collateral damage, as has been seen with social media [64].

Above, we identified the greatest risk of AI as the possibility that, with such "superpowers" on offer, many people might *opt out* of the essential struggles of being human. This would hinder their maturity, making them ill-equipped to thrive in a highly technological world, ultimately cutting them off from the Good Life they desire. Furthermore, a large number of people opting out would become a burden

on a society that depends on cooperation. To counter this, we need societal efforts in *character formation*: it takes the best within us, our best selves, to handle the most powerful tools responsibly. This has implications for curricula (in addition to industry-focused competence orientation, a renewed focus on ethics, philosophy, and spiritual resources that contain deep insights into human worth and value). Perhaps even AI could play a role, serving as a coach to support each of us in the necessary overcoming of oneself. The future, however, is shaped by ourselves. We are the ones having agency.

References

[1] Pascal Kaufmann, Thilo Stadelmann, and Benjamin F Grewe. *ChatGPT läutet Tech-Revolution ein*. Finanzen und Wirtschaft, https://www.fuw.ch/chatgpt-laeutet-tech-revolution-ein-303487897856. 2023.

[2] Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, and Rodney Zemmel. *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company, available online (19.11.2024): https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier. 2023.

[3] Ruth Fulterer. *Künstliche Intelligenz contra Menschheit: Diesen Kampf gibt es nicht. Er ist nur eine rhetorische Strategie*. NZZ, https://www.nzz.ch/meinung/kuenstliche-intelligenz-vs-menschheit-diesen-kampf-gibt-es-nicht-er-ist-nur-eine-rhetorische-strategie-ld.1732360. 2023.

[4] Yuval Noah Harari. *Homo deus: A brief history of tomorrow*. Harper, 2017.

[5] Kai-Fu Lee and Chen Qiufan. *AI 2041: Ten visions for our future*. Crown Currency, 2021.

[6] Thilo Stadelmann, Martin Braschler, and Kurt Stockinger. "Introduction to applied data science". In: *Applied data science: lessons learned for the data-driven business*. Springer, 2019, pp. 3–16.

[7] Jean-Daniel Dessimoz, Jana Koehler, and Thilo Stadelmann. "Artificial intelligence research in Switzerland". In: *AI Magazine* 36.2 (2015), pp. 102–105.

[8] Thilo Stadelmann, Julian Keuzenkamp, Helmut Grabner, and Christoph Würsch. "The AI-atlas: didactics for teaching AI and machine learning on-site, online, and hybrid". In: *Education Sciences* 11.7 (2021), p. 318.

[9] Thilo Stadelmann, Vasily Tolkachev, Beate Sick, Jan Stampfli, and Oliver Dürr. "Beyond ImageNet: deep learning in industrial practice". In: *Applied data science: lessons learned for the data-driven business* (2019), pp. 205–232.

[10] Thilo Stadelmann. "KI als Chance für die angewandten Wissenschaften im Wettbewerb der Hochschulen". In: *Bürgenstock-Konferenz der Schweizer*

Fachhochschulen und Pädagogischen Hochschulen, Luzern, Schweiz, 20.-21. Januar 2023. 2023.

[11] Rich Sutton. *The Bitter Lesson*. Incomplete Ideas, available online (01.10.2024): http://www.incompleteideas.net/IncIdeas/BitterLesson.html. 2019.

[12] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach, 4^{th} edition*. Pearson, 2022.

[13] Stephen Wolfram. *Remembering Doug Lenat (1950–2023) and His Quest to Capture the World with Logic*. Stephen Wolfram Writings, available online (01.10.2024): https://writings.stephenwolfram.com/2023/09/remembering-doug-lenat-1950-2023-and-his-quest-to-capture-the-world-with-logic. 2023.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[15] Jan Segessenmann, Thilo Stadelmann, Andrew Davison, and Oliver Dürr. "Assessing deep learning: a work program for the humanities in the age of artificial intelligence". In: *AI and Ethics* (2023), pp. 1–32.

[16] Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

[17] Paul-Philipp Luley, Jan M Deriu, Peng Yan, Gerrit A Schatte, and Thilo Stadelmann. "From concept to implementation: the data-centric development process for AI in industry". In: *2023 10th IEEE Swiss Conference on Data Science (SDS)*. IEEE. 2023, pp. 73–76.

[18] Thilo Stadelmann, Tino Klamt, and Philipp H Merkt. "Data centrism and the core of Data Science as a scientific discipline". In: *Archives of Data Science, Series A* 8.2 (2022).

[19] Future of Life Institute. *Pause Giant AI Experiments: An Open Letter*. FLI Open Letters, available online (02.10.2024): https://futureoflife.org/open-letter/pause-giant-ai-experiments/. 2023.

[20] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. *Statement from the listed authors of Stochastic Parrots on the "AI pause" letter*. DAIR Institute Blog, available online (02.10.2024): https://www.dair-institute.org/blog/letter-statement-March2023/. 2023.

[21] Pierre-Simon Laplace. *Essai philosophique sur les probabilités*. Courcier, 1814.

[22] Émile P Torres. *The Acronym Behind Our Wildest AI Dreams and Nightmares*. Truthdig, available online (01.10.2024): https://www.truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares. 2023.

[23] Leopold Aschenbrenner. *The Decade Ahead*. Situational Awareness, available online (01.10.2024): https://situational-awareness.ai/. 2024.

[24] Melissa Heikkilä. *Meta's AI leaders want you to know fears over AI existential risk are "ridiculous"*. MIT Technology Review, available online (03.10.2024): https://www.technologyreview.com/2023/06/20/1075075/metas-ai-leaders-want-you-to-know-fears-over-ai-existential-risk-are-ridiculous/. 2023.

[25] Andrew Ng. *A Victory for Innovation and Open Source*. The Batch Letters, available online (03.10.2024): https://www.deeplearning.ai/the-batch/a-victory-for-innovation-and-open-source/. 2024.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[27] Lukas Tuggener, Pascal Sager, Yassine Taoudi-Benchekroun, Benjamin F Grewe, and Thilo Stadelmann. "So you want your private LLM at home?: a survey and benchmark of methods for efficient GPTs". In: *11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 30-31 May 2024*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. 2024.

[28] Marc Zao-Sanders. *How People Are Really Using GenAI*. Harvard Business Review, available online (02.10.2024): https://hbr.org/2024/03/how-people-are-really-using-genai. 2024.

[29] Thilo Stadelmann. "Wie maschinelles Lernen den Markt verändert". In: *Digitalisierung: Datenhype mit Werteverlust? Ethische Perspektiven für eine Schlüsseltechnologie*. SCM Hänssler, 2019, pp. 67–79.

[30] Martin Braschler, Thilo Stadelmann, and Kurt Stockinger. "Data science". In: *Applied Data Science: Lessons Learned for the Data-Driven Business*. Springer, 2019, pp. 17–29.

[31] Thilo Stadelmann, Kurt Stockinger, Martin Braschler, Mark Cieliebak, Gerold Baudinot, Oliver Dürr, and Andreas Ruckstuhl. "Applied data science in Europe: Challenges for academia in keeping up with a highly demanded topic". In: *9th European Computer Science Summit, Amsterdam, Niederlande, 8-9 October 2013*. 2013.

[32] Thilo Stadelmann, Kurt Stockinger, Gundula Heinatz Bürki, and Martin Braschler. "Data scientists". In: *Applied Data Science: Lessons Learned for the Data-Driven Business*. Springer, 2019, pp. 31–45.

[33] Kurt Stockinger, Martin Braschler, and Thilo Stadelmann. "Lessons learned from challenging data science case studies". In: *Applied Data Science: Lessons Learned for the Data-Driven Business*. Springer, 2019, pp. 447–465.

[34] Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, et al. "Deep learning in the wild". In: *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*. Springer. 2018, pp. 17–38.

[35] Benjamin Meier, Thilo Stadelmann, Jan Stampfli, Marek Arnold, and Mark Cieliebak. "Fully convolutional neural networks for newspaper article segmentation". In: *2017 14th IAPR International conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 414–419.

[36] Lukas Tuggener, Raphael Emberger, Adhiraj Ghosh, Pascal Sager, Yvan Putra Satyawan, Javier Montoya, Simon Goldschagg, Florian Seibold, Urs Gut, Philipp Ackermann, et al. "Real world music object recognition". In: *Transactions of the International Society for Music Information Retrieval* 7.1 (2024), pp. 1–14.

[37] Felix M Schmitt-Koopmann, Elaine M Huang, Hans-Peter Hutter, Thilo Stadelmann, and Alireza Darvishy. "FormulaNet: A benchmark dataset for mathematical formula detection". In: *IEEE Access* 10 (2022), pp. 91588–91596.

[38] Felix M Schmitt-Koopmann, Elaine M Huang, Hans-Peter Hutter, Thilo Stadelmann, and Alireza Darvishy. "MathNet: A Data-Centric Approach for Printed Mathematical Expression Recognition". In: *IEEE Access* (2024).

[39] Pascal Sager, Sebastian Salzmann, Felice Burn, and Thilo Stadelmann. "Unsupervised domain adaptation for vertebrae detection and identification in 3D CT volumes using a domain sanity loss". In: *Journal of Imaging* 8.8 (2022), p. 222.

[40] Peter R Jermain, Martin Oswald, Tenzin Langdun, Santana Wright, Ashraf Khan, Thilo Stadelmann, Ahmed Abdulkadir, and Anna N Yaroslavsky. "Deep learning-based cell segmentation for rapid optical cytopathology of thyroid cancer". In: *Scientific Reports* 14.1 (2024), p. 16389.

[41] Mohammadreza Amirian, Javier A Montoya-Zegarra, Ivo Herzig, Peter Eggenberger Hotz, Lukas Lichtensteiger, Marco Morf, Alexander Züst, Pascal Paysan, Igor Peterlik, Stefan Scheib, et al. "Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks". In: *Medical Physics* 50.10 (2023), pp. 6228–6242.

[42] Raphael Emberger, Jens Michael Boss, Daniel Baumann, Marko Seric, Shufan Huo, Lukas Tuggener, Emanuela Keller, and Thilo Stadelmann. "Video object detection for privacy-preserving patient monitoring in intensive care". In: *2023 10th IEEE Swiss Conference on Data Science (SDS)*. IEEE. 2023, pp. 85–88.

[43] Lukas Hollenstein, Lukas Lichtensteiger, Thilo Stadelmann, Mohammadreza Amirian, Lukas Budde, Jürg Meierhofer, Rudolf M Füchslin, and Thomas Friedli. "Unsupervised learning and simulation for complexity management in business operations". In: *Applied data science: lessons learned for the data-driven business* (2019), pp. 313–331.

[44] Mattia Battaglia, Ennio Comi, Thilo Stadelmann, Roman Hiestand, Beat Ruhstaller, and Evelyne Knapp. "Deep ensemble inverse model for image-based estimation of solar cell parameters". In: *APL Machine Learning* 1.3 (2023).

[45] Peng Yan, Ahmed Abdulkadir, Giulia Aguzzi, Gerrit Schatte, Benjamin F Grewe, and Thilo Stadelmann. "Automated process monitoring in injection molding via representation learning and setpoint regression". In: *11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 30-31 May 2024*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. 2024.

[46] Niclas Simmler, Pascal Sager, Philipp Andermatt, Ricardo Chavarriaga, Frank-Peter Schilling, Matthias Rosenthal, and Thilo Stadelmann. "A survey

of un-, weakly-, and semi-supervised learning methods for noisy, missing and partial labels in industrial vision applications". In: *2021 8th Swiss Conference on Data Science (SDS)*. IEEE. 2021, pp. 26–31.

[47] Peng Yan, Ahmed Abdulkadir, Paul-Philipp Luley, Matthias Rosenthal, Gerrit A Schatte, Benjamin F Grewe, and Thilo Stadelmann. "A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions". In: *IEEE Access* (2024).

[48] Thilo Stadelmann and Bernd Freisleben. "Unfolding speaker clustering potential: a biomimetic approach". In: *Proceedings of the 17th ACM international conference on Multimedia*. 2009, pp. 185–194.

[49] Thilo Stadelmann, Sebastian Glinski-Haefeli, Patrick Gerber, and Oliver Dürr. "Capturing suprasegmental features of a voice with RNNs for improved speaker clustering". In: *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*. Springer. 2018, pp. 333–345.

[50] Daniel Neururer, Volker Dellwo, and Thilo Stadelmann. "Deep neural networks for automatic speaker recognition do not learn supra-segmental temporal features". In: *Pattern Recognition Letters* 181 (2024), pp. 64–69.

[51] Dorian Selz. *Where's the ROI for AI?* LinkedIn Post, available online (03.10.2024): `https://www.linkedin.com/posts/dselz\_wheres-the-roi-for-ai-lets-get-real-activity-7247176825972387840-rLWN`. 2024.

[52] Jürg Meierhofer, Thilo Stadelmann, and Mark Cieliebak. "Data products". In: Springer, 2019, pp. 47–61.

[53] Mohammadreza Amirian, Friedhelm Schwenker, and Thilo Stadelmann. "Trace and detect adversarial attacks on CNNs using feature response maps". In: *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*. Springer. 2018, pp. 346–358.

[54] Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. "How (not) to measure bias in face recognition networks". In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer. 2020, pp. 125–137.

[55] Samuel Wehrli, Corinna Hertweck, Mohammadreza Amirian, Stefan Glüge, and Thilo Stadelmann. "Bias, awareness, and ignorance in deep-learning-based face recognition". In: *AI and Ethics* 2.3 (2022), pp. 509–522.

[56] Eleonora Viganò, Corinna Hertweck, Christoph Heitz, and Michele Loi. "People are not coins: Morally distinct types of predictions necessitate different fairness constraints". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 2293–2301.

[57] Saurabh Jha and Eric J Topol. "Upending the model of AI adoption". In: *The Lancet* 401.10392 (2023), p. 1920.

[58] Kenneth M Ford, Patrick J Hayes, Clark Glymour, and James Allen. "Cognitive Orthoses: Toward Human-Centered AI". In: *AI Magazine* 36.4 (2015), pp. 5–8.

[59] Neil D Lawrence. *The atomic human: Understanding ourselves in the age of AI*. Allen Lane, 2024.

[60] Andy Crouch. *The Life We're Looking for: Reclaiming Relationship in a Technological World*. Convergent Books, 2022.

[61] Andrew Ng. *Agentic Design Patterns Part 1*. The Batch Letters, available online (03.10.2024): https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/. 2024.

[62] Christoph von der Malsburg, Thilo Stadelmann, and Benjamin F Grewe. "A theory of natural intelligence". In: *arXiv preprint arXiv:2205.00002* (2022).

[63] Pascal J Sager, Jan M Deriu, Benjamin F Grewe, Thilo Stadelmann, and Christoph von der Malsburg. "The Dynamic Net Architecture: Learning Robust and Holistic Visual Representations Through Self-Organizing Networks". In: *arXiv preprint arXiv:2407.05650* (2024).

[64] Jeff Orlowski. *The social dilemma*. Documentary, available as a Netflix original, see also (03.10.2024): https://www.thesocialdilemma.com/. 2020.