AI & SOCIETY
https://doi.org/10.1007/s00146-023-01808-9

OPEN FORUM

# Trustworthy AI: AI made in Germany and Europe?

Hartmut Hirsch-Kreinsen[1] · Thorben Krokowski[2]

Received: 10 May 2023 / Accepted: 24 October 2023

**Abstract**
As the capabilities of artificial intelligence (AI) continue to expand, concerns are also growing about the ethical and social consequences of unregulated development and, above all, use of AI systems in a wide range of social areas. It is therefore indisputable that the application of AI requires social standardization and regulation. For years, innovation policy measures and the most diverse activities of European and German institutions have been directed toward this goal. Under the label "Trustworthy AI" (TAI), a promise is formulated, according to which AI can meet criteria of transparency, legality, privacy, non-discrimination, and reliability. In this article, we ask what significance and scope the politically initiated concepts of TAI occupy in the current process of AI dynamics and to what extent they can stand for an independent, unique European or German development path of this technology.

**Keywords** Artificial intelligence · AI Act · Trustworthy AI · AI ethics · AI Made in Europe · Human-centric AI

## 1 Introduction: discussion on AI regulation

As the possibilities and capabilities of artificial intelligence (AI) continue to expand, concerns are also growing about the ethical and social consequences of unregulated development and, above all, use of AI systems in a wide range of social areas. Ethically and socially problematic uses of AI in differing contexts, as well as assessments of its risks, such as in medicine, work processes and human resource management, education, law enforcement, and more, have long been the subject of intense socio-political debates.[1]

The widely shared assumption is that measures to regulate and standardize AI can reduce doubts and increase social acceptance, satisfy legal frameworks, and create legal certainty for AI applications, and moreover solve a wide range of ethical application problems (e.g., Tsamados 2022; Gervais 2023; Gill 2023; Finocchiaro 2023; Foffano et al. 2023). For example, the EU Commission pointedly states that the greatest risks associated with the use of AI concern the application of rules protecting fundamental rights, including data protection and privacy and non-discrimination, as well as security and liability issues (cf. EU 2020; Heesen et al. 2021). Especially since the voice system "ChatGPT" has been freely available to everyone and competitors are presenting comparable products, it has become clear to many stakeholders how necessary it is to regulate the new technology appropriately (cf. Lauer 2023; Schwartmann 2023; The Economist 2023). Only recently, a group of AI experts and well-known entrepreneurs, including X[2] and Tesla boss Elon Musk, who can only be classified as technology—and innovation—averse to a certain extent, published an open letter in which they warn of devastating AI risks and classify the out-of-control international AI race as a threat to humanity (see Vallance 2023). The authors proclaim, that current AI tools present "profound risks to society and humanity" (Metz and Schmidt 2023)—statements that, on the one hand, contribute to heightened public and media expectations and, moreover, seem to provide additional legitimacy to the narrative of dystopian AI futures.

✉ Hartmut Hirsch-Kreinsen
hartmut.hirsch-kreinsen@tu-dortmund.de

Thorben Krokowski
thorben.krokowski@tu-dortmund.de

[1] Faculty of Social Sciences, Social Research Centre Dortmund, TU Dortmund University, Dortmund, Germany

[2] Faculty of Social Sciences, Social Research Centre Dortmund, Research Training Group GRK 2193, Dortmund, Germany

[1] See e.g., the recent Special Issue of the journal AI&Society (Vol. 38, no. 2) on the problematic of a societal AI regulation AI&Society 2/2023) or Cambridge University Press' book publications: https://www.cambridge.org/core/series/artificialintelligence- for-social-good/4E5463903600210621 2C0A3812752C7C (accessed: May 5 2023).

[2] Formerly "Twitter".

Published online: 22 November 2023

On the other hand, the legitimacy of the call for regulatory control elements receives backing from a direction that tends to be unusual. Accordingly, an AI strategy has been pursued for some time in the EU as well as in Germany in terms of innovation policy, which can be explicitly summarized with the label "Trustworthy AI" (TAI). Following the definition of the European High-Level Expert Group on AI, this can be defined by three fundamental dimensions: "[L]awful, complying with all applicable laws and regulations; …ethical, ensuring adherence to ethical principles and values; and … robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm" (HLEG 2019, p. 5). What is meant by this is that a responsible and public welfare-oriented standardization and regulation of AI systems (cf. BMBF 2021, p. 13) is being strived for. With the label TAI, an AI promise is formulated according to which AI can meet the requirements of transparency, legality, accountability, privacy, non-discrimination, and reliability (cf. EU 2020; Heesen et al. 2020; HLEG 2019; 2020). This promise is also outlined with the formula "human-centric" or "human-centric approach to AI" (European Commission 2019, p. 9). Technologically, the focus is on development and application problems of the various machine learning and neural network methods.

With the promise of TAI, a special distinguishing feature and a competitive advantage of European solutions (cf. PLS 2020, p. 9) are aimed at in each case in several respects: First, it is about competitive advantages in the global technology race (cf. Bitkom and DFK 2017; Bundesregierung 2018; EU 2020; Van Roy et al. 2021). Second, however, the many critical socio-political objections in the AI discourse are to be taken up and implemented within the framework of a German and European development path for AI technology. Third, this is intended to expand Europe's position as a global center for Trustworthy AI and, in the process, establish an ecosystem for excellence and trust for AI (cf. European Commission 2021). Originally, the process of standardization and regulation of AI in the EU was supposed to enter a new, possibly decisive phase in spring 2023. Thus, almost complete draft laws on the risk-oriented use and liability of AI were available. The European Parliament wanted to decide on the "Artificial Intelligence Act" (AI Act)[3] in spring 2023 so that the legislative project could be completed in this legislative period. However, with the not-inconsiderable disruptive noise and new lines of development that have emerged in the course of the possibly unexpectedly radiant emergence of ChatGPT, there were then sometimes heated debates among numerous members of the European Parliament regarding, among other things, the content of the AI law and its compatibility with the comments on generative AI (cf. Riegert 2023). Above all, AI systems that serve general purposes and can perform a variety of tasks were not considered in the original proposal. This only happened with the emergence of ChatGPT, as a generative general-purpose AI model that takes text input and provides users with high-quality, contextual responses. Against this background, numerous compromised amendments have since been tabled, which could significantly prolong the negotiations and the time until the AI Act is actually passed (cf. Sharma 2023).

This article follows on from this debate with a specific focus on the European and German discourse. Specifically, the article discusses the following questions: First, which players are driving the discourse on TAI and what objectives are being pursued. Second, how can the impact and scope of this discourse and the corresponding measures be assessed? Third, what challenges and open problems can be identified. Fourth, to what extent can the concept of TAI be seen as the central feature of a European or German development path of AI? The structure of the paper is based on these questions.

## 2 Methodology

Methodologically, the study is based on a broad qualitative foundation. It is based on the results of the analysis of current scientific, political, and public discourses on AI, and here in particular on the review and evaluation of multiple "gray" documents, preprints, political pronouncements, websites, studies, and relevant specialist publications from national and international contexts. Interviews and statements by AI experts that were accessible on the Internet often proved to be particularly informative. Furthermore, the argumentation is based on the results of our own detailed expert interviews. Between October 2021 and March 2022, 19 interviews with 16 AI experts from Germany were conducted as part of the project. Ten basic and application-oriented scientists and six AI experts from AI-related scientific disciplines and application fields were interviewed. It should be emphasized that among them were four emeriti from the first generation of AI experts in western Germany. The interviews were conducted online and lasted at least one hour, often considerably longer. Some of the expert interviews included several rounds of feedback to clarify unanswered questions or to discuss initial theses. With three experts, individual questions could be deepened in a second detailed

[3] "The AI Act is a proposed European law on artificial intelligence (AI)—the first law on AI by a major regulator anywhere. The law assigns applications of AI to three risk categories. First, applications and systems that create an unacceptable risk, such as government-run social scoring of the type used in China, are banned. Second, high-risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated" (FLI 2023).

round of interviews. The interviews covered question complexes of AI development, its application problems, and its development perspectives. Especially, questions were asked about the special features and challenges of a specifically European and German AI development, about the possibilities and concepts of AI regulation and, above all, its practicability. The interview results were recorded, and the central statements of the conversations transcribed.

In addition to that, the analysis draws on the authors' own research findings on the social digitisation process of recent years. The results and findings from the authors' ongoing participation in relevant workshops and conferences, as well as informal discussions with academics, practitioners, and political representatives, have been incorporated into the argumentation.

The evaluation and interpretation of the analysis results was carried out according to the methodological principle of social science triangulation (e.g., Flick 2007). This social–scientific method of analysis made it possible to validate the different perspectives of different survey methods in a comparative way regarding the same facts. In the present study, the results of the literature research, the interviews, and the other more informal discussions were used. This made it possible to systematically analyze the subject area in its breadth and depth. The results of the analysis are presented in the following sections of the paper. The presentation of the intermediate methodological steps would go beyond the scope of this paper.[4]

## 3 Driving players and their objectives

The topic of "Trustworthy AI" is taken up in the context of a wide variety of bodies, commissions, and organizations and linked to questions of AI development and application. All in all, a very dynamic and difficult-to-understand field of actors and institutions is increasingly establishing itself, both in the national framework and internationally, to deal with standardization and regulation issues in AI (cf. Foffano et al. 2023, pp. 482). In addition to governmental and politically initiated activities and institutions, semi-governmental bodies, such as standardization and testing organizations, associations, such as trade unions and employers' associations, NGOs such as AlgorithmWatch or Future of Life, various institutes, and a broad variety of national and international companies are increasingly addressing this issue in different ways (cf. Kleine and Weber 2022).

The fact that the reference to ethical guidelines in the context of dealing with algorithms and artificial intelligence is considered politically expedient is justified in two ways (cf. BMI 2018):

- The regulatory sovereignty of private, non-European de facto regulators, the large platform companies, which mostly act without recourse to European values as well as applicable (data protection) law, is to be clearly countered.
- The lack of data sovereignty, as a result of Europe's high dependence on U.S. and Chinese data and analytics infrastructures, whose data protection regulations in part clearly run counter to the European understanding of (data) sovereignty, should be overcome (cf. Kagermann and Wilhelm 2020, p. 7).

For the EU Commission, this situation is reason enough to budget a sum of 9.2 billion euros in the EU funding program "Digital Europe"[5] for the purpose of triggering the digital transformation of the European economy and society. In the associated Coordinated Plan on Artificial Intelligence, a central activity is the work of a high-level expert group on AI (the so-called High-Level Expert Group, HLEG) with the task of developing ethics guidelines—"Ethics Guidelines for Trustworthy AI"—for the use of artificial intelligence (cf. HLEG 2019, 2020). The conceptualisation of a human-centered approach with recourse to ethics aspects in standardization for AI is thereby understood as a supporting pillar in the attempt to develop key AI technologies and technology-based innovations in Europe independently and to build up own production capacities for this (cf. BMBF 2021, p. 3). The label Ethics by Design is understood as an approach to consider ethical criteria already in the development process. It refers to "an approach to design that aims at the systematic inclusion of ethical values, principles, requirements, and procedures into design and development processes" (SIENNA 2021).

A current attempt at EU level to put a legal stop to the non-transparent or even discriminatory or abusive use of AI applications is embodied in the aforementioned AI Act. It represents the world's first cross-society effort to establish harmful effects protections for AI-based systems in the form of risk assessment categories. The law takes a holistic target group approach and affects not only EU citizens and European society as a whole, but also companies that provide or deploy AI systems within the EU (see AlgorithmWatch 2022; Urban 2023). Following the view of AlgorithmWatch,

[4] The detailed findings generated during the study are available from the corresponding author on reasonable request (will be completed after the review process).

[5] The Digital Europe (DIGITAL) program is a new EU funding program focused on delivering digital technologies to businesses, citizens, and public administrations (cf. European Commission 2022).

the provisions made apply whenever AI-based systems are used in the EU—regardless of where the systems are operated from—or whenever the operation of the systems has consequences in the EU—regardless of where they are operated (cf. AlgorithmWatch 2022). Thus, ChatGPT—or at least central functional applications of the tool—is/are also classified as a high-risk application within the framework of the AI Act (cf. Helberger and Diakopoulos 2023)—although it should be noted (see above) here that an officially adopted, final assessment of the scope of the restriction of ChatGPT is not yet available at the time of publication of this paper.

At the national level, for example, this discourse is being driven by the Bundestag's Commission of Inquiry, which discussed AI issues from a comprehensive socio-political perspective from 2018 to 2020 under the heading "Artificial Intelligence—Social Responsibility and Economic, Social and Ecological Potentials" and presented far-reaching socio-political recommendations for action on AI development in its voluminous final report (cf. Enquete Commission Artificial Intelligence 2020). The Data Ethics Commission set up by the government also has a major influence on the AI discourse. Its mandate is to propose a development framework for data policy, the handling of algorithms, AI and digital innovations, to make recommendations for action and to identify regulatory options. In this context, an action-guiding framework is to be developed above all by the criticality pyramid launched by the Data Ethics Commission (cf. Data Ethics Commission 2019, p. 177). Likewise, at the beginning of 2023, the German Ethics Council issued a fundamental statement regarding recent AI developments and applications as well as the associated ethical, legal, and social risks and multiple political and regulatory recommendations for action. The statement is to be understood as a reaction to the request formulated by the President of the German Bundestag in October 2020 to develop a multidisciplinary statement on the ethical issues of the relationship between "Human and Machine" (German Ethics Council 2023).

From an application-oriented perspective, a working group of the so-called "Learning Systems Platform" (PLS) set up by the German government is addressing the legal and ethical challenges of the new digital and technological systems (cf. Müller-Quade et al. 2019). This working group, entitled "IT-Security, Privacy, Law and Ethics", argues that learning systems are increasingly taking over tasks from humans, but that these systems themselves do not have legal personality. Therefore, new legal regulations, for example in liability law, would have to be found for AI use, or existing law would have to be adapted. In addition, the systems are not capable of making moral decisions on their own or of judging their decisions according to moral standards. Therefore, the ethical requirements would have to address the process of programming and using learning systems. The general objective is to lay the foundations for legal and ethical regulation and certification of AI systems and thus to create the preconditions for putting AI systems into use and exploiting their full potential benefits (cf. Heesen et al. 2020).

An essential prerequisite for the realization of this concept is to establish standards and norms for the implementation and application of AI systems. One starting point for this is the "AI Standardization Roadmap" launched in Germany. Its aim is to drive forward the development of a framework for the standardization of AI that supports the international competitiveness of the German economy and raises European value standards to the international level (cf. Wahlster and Winterhalter 2020). Linked to this is the ambition to develop compatible criteria and requirements for the development and use of AI in the various application areas at the national and European level in order to strengthen the confidence of users in AI. Based hereon, it is hoped that it will be possible to ensure the potential benefits for society and to exploit the potential provided by the use of AI in a way that is oriented toward the public welfare. The overarching goal of developing and enforcing standards is to strengthen the interests of German science and industry in the international competition for AI and also to create innovation-friendly conditions for sustainable and future-proof value creation in the field of AI technology (see BMWi 2020).

## 4 Limited impact of the politically driven discourse

Overall, the debate about a German or European, ethically oriented AI version has so far been primarily a politically driven discourse. One of the interviewed AI scientists takes a very critical view of this:

> In fact, this area is currently being filled with life in that people are sitting in Brussels and trying to cast these things into laws. But I don't think that's very effective. People are formulating laws about AI systems who don't know what an AI system is, because it's very difficult to define. When lawyers sit together, they come up with different ideas than AI people would. My impression is that this is negotiated by the lawyers alone, the legislation, without the subject matter being backed up by the AI experts (Anon., personal communication, 2021).

In other words, the discourse on TAI is mainly a top-down call in the name of public interest. Therefore, the discourse should be pushed toward more "participatory and inclusive processes" (Gill 2023) when developing and implementing AI systems. In addition, there is much to be said for the assessment that a discussion of ethical guidelines within the socio-technical field of AI technology has so far tended to be given only limited priority. “Though a number of groups are producing a range of qualitative or normative outputs in the AI ethics domain, the field generally lacks benchmarks” (Zhang et al. 2021, p. 127). Various reasons can be cited for this (cf. Beckert 2021):

### 4.1 Gap between AI strategies and ethical requirements

The intention of many developers and companies to offer an AI product as a first mover leads to the neglect of ethical criteria in a short-term oriented development and in the fastest possible deployment of the systems. An example of this are systems for staff assessment and recruitment, which are often placed in markets without consideration of inherent mechanisms of discrimination against certain groups of people. Notwithstanding their demonstrably discriminatory effect and manipulation-prone structure, pioneering companies have been able to secure a significant market advantage, despite all the criticism (cf. Noll 2019; Gupta et al. 2021). This example points to the fundamental problem of AI developers’ commercial strategies being at odds with the social and ethical demands placed on AI. As The Economist recently highlighted: “Microsoft recently disbanded one of its AI ethics team, for example. Indeed, some researchers think the true “alignment” problem is that AI firms, like polluting factories, are not aligned with the aims of society. They financially benefit from powerful models but do not internalize the costs borne by the world of releasing them prematurely” (The Economist 2023). An interviewed expert formulates this contradiction even more clearly:

> AI research is, by its very nature, innovation research with many start-ups and big players. They want to earn money, that’s clear. Of course, the [ethical] problems don’t come first. That comes from the outside with the ethics (Anon., personal communication, 2022).

### 4.2 Transparency deficits between TAI and application

Moreover, the relevance and the functional connection between TAI and concrete application are not always obvious. This is often unclear to operational users, for example in the industrial sector or in the research context, so that the recourse of ethical aspects is often considered superfluous—this is a problem that is inherent in almost all currently existing ethics frameworks. For example, a recent study showed that most frameworks almost exclusively do not specify target groups. Rather, most frameworks promise to present a one-size-fits-all solution. In practical application at the latest, however, this promise can hardly or not at all be fulfilled if, for example, the focus of the AI guidelines is strongly technical and can only be adequately understood by the AI system developers themselves (cf. Qiang 2023). It can be concluded that there is often a lack of appropriate know-how, specific planning, and additional resources to adequately meet the AI-genuine constitution, which consists of the ability to approximate any continuous function with arbitrary accuracy (cf. Kersting and Tresp 2019, p. 4), or to meet the requirements of TAI within workflows and testing processes with identical intensity (cf. Beckert 2021, p. 20).

### 4.3 Insufficient assessment of implementation challenges

According to interviewed experts, the last point in particular reveals a fundamental shortcoming since companies are already confronted with a variety of new areas of tension in the organizational and work context in view of the implementation of AI, which often cannot be properly assessed and result in new follow-up costs. It is true that there is a relatively high level of sensitivity among many stakeholders, with regard to the issue of TAI and the socio-political goals pursued with it. However, from an application-oriented perspective, this concept is seen as difficult to implement and also as primarily a political issue rather than one relevant in the discourse of innovation-related topics (cf. Hirsch-Kreinsen 2023a, b, p. 172).

### 4.4 Whitewashing of AI

In the more recent discussion, critical voices can also be found, which doubt the regulatory effects of this debate and the measures oriented to it. A central problem is that the field of AI applications is becoming increasingly unmanageable and cannot be grasped at all. Therefore, if one follows Jansen and Cath’s position, all activities of this kind aim at little more than a whitewashing of AI and its applications (e.g., Jansen and Cath 2021). Thus, the concept of TAI as a distinguishing feature in global competition does not always resonate particularly well even at some innovation policy levels. For example, the report of the German Commission of Experts for Research and Innovation (EFI) of 2022 is worth mentioning. On the one hand, the commission complains about the considerable backlog of the Federal Republic of Germany in many key technologies, in particular AI, in international comparison. On the other hand, however, apart from the topics of data protection and IT security, other terms of a European perspective on AI, such as trustworthy(ness) and ethical in connection with AI, are surprisingly not addressed anywhere in the comprehensive report (cf. EFI 2022).

### 4.5 Structural limits to the implementation of AI ethics

There is no doubt that there are also structural limits to the implementation of European-wide AI ethics and its criteria. The AI Ethics Impact Group (AIEI Group) sees such limitations primarily in non-negligible influencing factors such as cultural context dependencies, which have different logics of action and design. For example, different national institutional regulations that do not allow uniform implementation of standards throughout Europe must be taken into account. One example of this is the industrial relations in the individual EU countries, which in some cases diverge greatly, and which at the company level open up very different or in some cases no legal possibilities for intervention at all with regard to employee data protection. As is well known, in contrast to many other countries, work councils in Germany have comparatively far-reaching co-determination rights with regard to the use of personal data and associated control options in work processes.

### 4.6 TAI as an obstacle to innovation

There is also critical talk that legally and ethically oriented regulatory criteria could be a barrier to innovation for AI and that resistance to political activities is therefore emerging. This is formulated, for example, with regard to the introduction of the EU General Data Protection Regulation (GDPR). Not surprisingly, this criticism is formulated by the Federation of German Industries (BDI): “Under no circumstances should data protection law be allowed to become an obstacle to innovation and a location disadvantage” (BDI 2018). In addition, and this is an extremely interesting point, a group of German leading research institutes, scientists, and business representatives of the Large-Scale Artificial Intelligence Network (LAION e.V.)[6] recently wrote an open letter to the EU Parliament. They demand that the draft of the AI Act should be improved with a view to foreign policy security and economic competitiveness. In contrast to the genuine intention of the AI Act, namely to protect the freedom of citizens and to reduce the security risk, the authors accuse the AI Act of jeopardizing the freedom of research and digital resilience. The authors argue that one-size-fits-all regulation, which hinders open-source approaches in the field of generalized AI, stands in the way of transparency and security of AI systems. As a result of the adoption of the AI Act, Germany and Europe are facing a point of no return.

> European research and development would then be left behind in the long term, with significant consequences for business and research. It would also pose a political security risk. Companies would also become massively dependent as end users of US APIs, for example. Data and added value would flow out of the EU—and even if the servers were located within the EU, the local industry would be dependent on the goodwill of some foreign companies in other jurisdictions (Hahn 2023).

There are also quite a few critical voices pointing to the slowing effect that the AI Act, for example, could have on the domestic European economy. Fears of a scenario in which Germany or the EU develops from a high-tech location into an “industrial museum” (cf. Glauner 2023) due to overregulation are not uncommon. It is undisputed that Germany and the EU must walk a tightrope in this matter, which is primarily characterized by two decision options: On the one hand, the question can be raised as to whether it is the regulatory framework that is possibly inhibiting technological innovations and thus also AI innovations. On the other hand, there is no question that AI aspects, such as transparency, explainability, and traceability, should be subject to intensive discourse, especially in high-risk scenarios (cf. Pereira 2021). Besides, however, there may also be impetus for social innovations oriented toward and desired by society. It is reported, for example, that there are also voices from the companies themselves calling for clear framework conditions. Yet, these do not represent an obstacle to innovation, but rather a clear market structure with certain standards that provide impetus for new innovations. In addition, this is associated with the expectation that social acceptance problems and damage to the image of AI can be avoided. Finally, there are quite a few indications that some companies are hoping to exploit a new field of business for the purpose of testing and certification procedures that are becoming increasingly necessary in the wake of the emergence of ethical AI concepts (cf. Hirsch-Kreinsen 2023a, b, p. 174). For instance, quite a few consulting firms are already becoming active as first movers in the market for ethical AI in order to secure a promising place for themselves at an early stage with a view to the expected run on socially acceptable, TAI algorithms. For example, providing an “Ethics Compass for Data and Artificial Intelligence” for a fee is intended to support companies interested in integrating AI ethics into their governance as well as companies that have already established ethics governance and want to measure and increase the maturity of their AI ethics (cf. KPMG 2022).

[6] Large-Scale Artificial Intelligence Network (LAION e.V.): “LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models” (LAION 2023).

## 5 Unresolved challenges

Overall, the manifold and unresolved challenges of further development and enforcement of concepts of TAI are unmistakable. Many unanswered questions still exist regarding how legal and ethical criteria are precisely formulated and enforced in order to be able to influence the dynamics of AI. Thus, the current state of the debate on the criteria for legal and ethical regulation of AI development and application is characterized by a great deal of conceptual heterogeneity. A uniform canon of central normative criteria or a catalog of core principles isn't discernible so far. Rather, different actors are also emphasizing different normative rules (cf. Rudschies et al. 2021). In summary, the following should be highlighted here.

### 5.1 Problem of knowledge transfer

On the one hand, the problem of knowledge transfer between mutually isolated domains is problematised on the occasion of the feasibility of ethically oriented AI systems. Thus, according to Beckert, it is necessary to "bring together the two worlds of software development and ethics" (Beckert 2021, p. 21). An interviewed expert formulated this problem very aptly: "Software developers are not necessarily data protectionists" (Anon., personal communication, 2022). Therefore, the formation of interdisciplinary teams proves to be indispensable, but according to all the available findings, and especially in view of the lack of expertise in this field, it appears to have hardly been realized to date.

### 5.2 Ambiguities in standardization

On the other hand, it is not clear which areas—either data use or algorithmic processes—should be the primary focus of standardization. For example, the much-cited EU GDPR is viewed critically. It focuses exclusively on the quality of the data and neglects the question of how algorithmic AI processes arrive at certain results and decisions (cf. Hirsch-Kreinsen 2023a, b, p. 153). With regard to an evaluation or certification of algorithms, the different methods of AI have to be considered in different ways. Often, neural networks and their opacity are seen as a central regulatory problem (cf. Shrestha et al. 2019, p. 68; Steininger 2023). According to many involved experts in the discussion, it also remains unclear how to deal with other machine learning methods or symbolic AI methods. A solution to this problem that has been discussed for some time is to code AI systems according to normative or legal requirements i.e., AI machines are programmed to ensure that they will abide by a set of principles represented explicitly in AI system design and decision-making algorithms (cf. Gervais 2023, p. 399). As the discussion makes clear, however, this can only be a set of generally defined standards that are considered relevant. In concrete terms, these must then be continuously specified according to the field of application and the state of development of the technology.[7]

### 5.3 Inconsistent assessments of application contexts

In addition, there is skeptical discussion about how application risks can be assessed in different fields, how the concept of criticality is defined as a measure of potential dangers, and what normative consequences should be drawn from this. Divergent definitions are presented here, for example, by the Data Ethics Commission (2019) and the EU Commission with the AI Act (cf. Müller 2022). The unresolved question of how diverse forms of design and use of the systems should be captured and regulated also plays a role here. Typical of this is the different interpretation of AI systems, namely either as an assistance system that supports human action or as an autonomous system that is intended to make decisions on its own and aims to largely automate processes. It is therefore essential to take a differentiated view of AI: depending on the application context, one and the same system can be associated with a variety of risks or levels of criticality (cf. Heesen et al. 2021). Likewise, there are conditions of diverse areas of application, for which divergent needs and requirements for ethical AI guidelines are relevant in each case. While it is undisputed among experts that the concept of risk or criticality assessment chosen by the European Commission to ensure the quality of AI systems provides a good orientation function for evaluation and regulation, it should, if the experts are to be followed, be supplemented by a far more precise classification of criteria. These should cover the various application contexts as precisely as possible. At the same time, however, innovation potentials should not be hindered. It is emphasized that there is still a great need for research in this area (cf. Heesen et al. 2021).

### 5.4 The need to consider development and implementation processes

Besides, actionable frameworks for TAI algorithms would have to consider the complex development and implementation process, the socio-technical nature of AI systems, and the related responsibility of system development toward system users (cf. AIEI Group 2019, p. 10). On top of this, the opacity of AI dynamics must not be neglected in view of its multiplex application functions. Thus, the requirements for application and regulation are very differentiated in various

[7] (E.g., Gervais 2023, pp. 402) and the literature discussed there.

fields of application. If one follows a personnel expert, this can be described based on two extreme situations: On one side, there is a "Wild West" in regulatory terms, such as AI applications in the field of human resource management, where there are great risks of discrimination through system decisions. This is a "hot market" with a lot of "junk software", he said. Conversely, there is already a strong control apparatus in autonomous driving and especially in the medical field, "so you can't just do something with AI somehow" (quoted from Hirsch-Kreinsen 2023a, p. 154). The challenges that arise while regulating or certifying AI systems are vividly described by an expert from a certification body:

> Honestly, we'll have to work that out concretely in the coming years. But what I can say right now: One criterion will certainly be whether the data set with which such an AI is trained is free of bias. Has care been taken to ensure that minorities or minority opinions are not discriminated against in an unruly manner? There are statistical procedures that we can use to determine, for example, whether data in a particular category is sufficiently represented. Moreover, I think it would be wrong to look only at the result. Basically, we have to start in the development process of the AI –, and really from the beginning to the end, from the formulation of the problem to the selection of the data to the categorisation and preparation. Is the data sufficiently powerful? Are they representative? Then comes the whole AI training process, where a lot can go wrong. We simply need good test criteria there. This, by the way, is also because some AI decisions are not easy for humans to understand for another reason: We once had a pilot project here to detect damage to turbine blades. The AI was hooked up to a very high-resolution digital camera that can recognize 100 times more color values than a human eye. It found things that we didn't see at all. We don't know immediately whether this is a real error or whether the AI has made a mistake (Armbruster & Knop 2022).

### 5.5 Challenges of technological change and of workable standards

Additionally, it must be outlined that regulation and standardization approaches are fundamentally confronted with rapid technological change and must therefore constantly adapt to new challenges and be updated. It is also unclear which authorities would be sufficiently competent and legitimized for this. Above all, however, it would be worth discussing whether the aforementioned normative standards of coding (cf. Gervais 2023) would be a practicable starting point for subjecting dynamic development to regulation. Because a classification system of application risks like this of the EU may be overwrought and a principles-based approach would be more flexible (cf. The Economist 2023).

In summary, despite all obstacles and contradictions, it can be stated that the discourse on TAI and the diverse political approaches have significantly increased the general awareness of the legal and ethical challenges of AI. Apart from that, however, the actual consequences for the development and application of AI are largely unresolved. The discussion about this and furthermore the modes of implementation regarding corresponding criteria are quite obviously still in their infancy (cf. Heesen et al. 2021).

## 6 Trustworthy AI as a unique selling proposition?

In order to establish Germany and Europe as global centers of TAI and as an ecosystem for excellence and trust (cf. European Commission 2021), policymakers never tire of emphatically reiterating the promise of TAI: "Europe can distinguish itself from others by developing, deploying, using and scaling TAI, which we believe should become the only kind of AI in Europe, in a manner that can enhance both individual and societal well-being" (European Commission 2019, p. 6). Hardly compatible, even contradictory objectives, such as economic growth, security, sustainability, poverty, disease reduction, etc., are to be harmonized and made equally feasible here. Consequently, AI is seen as the magical key to solve national, European and global problems (cf. Ossewaarde and Gülenç 2020, p. 55). The rationale for a European AI path can be found in the following formulation by the EU Commission: "[B]y trying to innovate in AI just in the same way as the US or China, Europe has already lost the competitiveness race" (Pereira 2021). Following on from the foregoing, we will conclude by asking whether the concept of TAI with its ethical dimensions actually represents a unique European and German selling point in the global race for AI, or whether it has at least the potential to develop in this direction.

Clear doubts are warranted here, because: If one compares internationally AI-oriented pronouncements of a wide variety of governments and countries, one can find comparable statements and partly already implemented ideas regarding the conceptualisation, design and introduction of TAI-based criteria. Qiang et al. (2023) even go so far as to speak of a "bombardment of AI ethics frameworks […] published in the last decade". This is shown by authors of an international comparative study on the perspectives pursued with AI policies in the USA, UK, and EU: "In particular, transparency, accountability, and a 'positive impact' on the economy and society are among the key values indicative of the kind of view of a 'good AI society'" (Cath et al. 2017,

p. 523). Furthermore, since 2017 60 countries, territories and the EU have published over 700 AI policy documents to set out their visions on a societal oriented AI (cf. Foffano et al. 2023, p. 481). For example, Canada's "Pan Canadian AI Strategy", published in 2017 and considered the world's first official AI strategy document, already highlights the development of ethical implications for AI as one of its main concerns. In 2018, India published its "National Strategy on Artificial Intelligence: #AIforALL", in which it highlighted, among other things, the research relevance regarding data protection, privacy, and ethical, bias-preventing AI efforts. Brazil's 2020 "Brazilian Artificial Intelligence Strategy" also addresses this issue. In addition, the "AI Ethics Framework" published by the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) in 2019 deals in detail with the importance and necessity of regulatory and ethical frameworks and guidance while researching, developing and disseminating AI products and applications (cf. Zhang et al. 2021, p. 155). It is also surprising to note that even from China comes a "White Paper on Artificial Intelligence" (2021) published by the China Academy of Information and Communications Technology (CAICT). With key elements, such as trustworthiness, traceability, and fairness, the Chinese AI strategy clearly overlaps with the core dimensions addressed in the concept published by the EU.

One aspect that has fatally received little attention is manifested in the lack of concrete demonstrable impact and often suggested benefits of AI-based ethics frameworks on AI companies, products and projects, as recently identified by Qiang et al. (2023). The reason for this is largely to be found in a "principles-to-practices gap", which—if one follows the argumentation of Schiff et al. (2021)—ironically can be identified not least as the result of an existing oversupply of ethical AI strategies and principles. As a result of this oversupply, it is difficult for individuals to sort through and evaluate the usefulness of a particular tool or weigh the benefits of a tool against the many other tools available. This can lead to individuals and organizations missing out and organizations useful tools and methods already exist. From an international perspective, therefore, a critical assessment of European efforts is not surprising. For example, Jessica Newman, program manager for the AI Security Initiative (AISI) at UC Berkeley's Center for Long-Term Cybersecurity (CLTC), sees German and European efforts to tout ethical, trustworthy, standards-based, regulated AI applications as an in-house invention as a distortion of reality. It is also rather exaggerated to see the EU as the "technological guardian of the world" and thus the U.S. implicitly as the digital West (cf. Venkina 2021). Additionally, it can be noted that European data protection guidelines and laws are indeed stricter than those of the U.S. or China and threaten tougher sanctions. Nevertheless, this does not necessarily apply to AI-based technologies and, in the wake of the introduction of the criticality model, a "large part of AI use […] remains unregulated, and only voluntary guidelines are proposed to promote responsible use" (ibid.). The meaningfulness and intention of AI standards is certainly not diminished by this circumstance in any way. Nevertheless, with recourse to the underlying findings, it seems disputable whether trustworthy and ethical AI approaches can be regarded as an exclusive and original product of European provenance and whether the two main competing AI conceptualisations and their attested main intention—China (control) and the USA (commerce) (cf. Beckert 2021, p. 17)—can be denied an interest in TAI technology based on ethical values. For in both the USA and China, the ethical challenges of AI are accorded a certain status, at least at the level of discourse, although not necessarily equal in some areas. However, the European approach undoubtedly goes further. Despite criticism of it in detail, it offers interesting orientations and approaches for the development of effective regulatory measures in future. Given the current rapid development of AI, these will be indispensable, in whatever form. Qiang et al. (2023), in any case, conclude that a construct such as a "one-size-fits-all AI ethics framework", with which ethical problems and effects of AI can be adequately uncovered and controlled, does not exist and is in any case not purposeful for the development of AI. Following on from this, the dogged orientation toward the creation of a distinct form of ethical framework in the German or European context also proves to be unrealisable and runs counter to the exploitation of the potential that nevertheless exists.

**Acknowledgements** The paper summarizes specific findings of a research project ("Expertise") in the context of the framework project „KI—Mensch—Gesellschaft: Den Wandel des Mensch-Technik-Verhältnisses durch Künstliche Intelligenz ganzheitlich verstehen und bewerten (KI.Me.Ge.)" coordinated by the ISF Munich and funded by the German Federal Ministry of Education and Research. Thorben Krokowski's contribution to the paper was made within the framework of the Research Training Group (GRK 2193) "Adaption Intelligence of Factories in a Dynamic and Complex Environment", funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project No. 276879186.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data availability** Data will be made available on request.

## Declarations

**Conflict of interest** The authors have no conflict of interest to declare.

## References

AIEI Group (2019) From Principles to Practice: An interdisciplinary framework to operationalise AI ethics. Artificial Intelligence Ethics Impact Group. Bertelsmann Stiftung, Guetersloh

AlgorithmWatch (2022) Ein Leitfaden zum AI Act: Wie die EU KI regulieren will und was das für uns alle bedeutet. 26.09.2022, AlgorithWatch. https://algorithmwatch.org/de/ai-act-erklaert/. Accessed March 13 2023

Armbruster A, Knop C (2022) So stellt sich der TÜV die KI-Prüfung vor. 22.02.2022, FAZ. https://www.faz.net/aktuell/wirtschaft/so-stellt-sich-der-tuev-die-ki-pruefung-vor-17820701.html. Accessed April 8 2022

Beckert B (2021) Vertrauenswürdige Künstliche Intelligenz – Ausgewählte Praxisprojekte und Gründe für das Umsetzungsdefizit. TATuP 30:17–22

Bitkom & DFKI (2017) Künstliche Intelligenz: Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung. Positionspapier, 05.09.2017. https://www.bitkom.org/Bitkom/Publikationen/Entscheidungsunterstuetzung-mit-Kuenstlicher-Intelligenz.html. Accessed June 25 2021

Bundesministerium für Wirtschaft und Energie (BMWi) (2020) „KI – Made in Germany" etablieren. 30.11.2020, BMWi. https://www.bmwi.de/Redaktion/DE/Presse-mitteilungen/2020/11/20201130-ki-made-in-germany-etablieren.html. Accessed March 14 2022

Bundesministerium des Innern und für Heimat (BMI) (2018) Leitfragen der Bundesregierung an die Datenethikkommission, Berlin

Bundesministerium für Bildung und Forschung (BMBF) (2021) Technologisch souverän die Zukunft gestalten – BMBF-Impulspapier zur technologischen Souveränität. Herausgegeben durch Bundesministerium für Bildung und Forschung (BMBF), Berlin

Bundesregierung (2018) Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz. 18.07.2018. https://www.bmbf.de/files/180718%20Eckpunkte_KI-Strategie%20final%20Lay out.pdf. Accessed May 28 2021

Bundesverband der Deutschen Industrie e.V. (BDI) (2018) Datenschutz darf keinesfalls zum Innovationshemmnis und Standortnachteil werden. 20.05.2018, BDI. https://bdi.eu/artike l/news/datenschutz-darf-keinesfalls-zum-innovationshemmnis-und-standortnachteil-werden/. Accessed April 14 2022

Cath C, Wachter S, Mittelstadt B, Taddeo M, Floridi L (2017) Artificial Intelligence and the ›Good Society‹: The US, EU and UK approach. Sci Eng Ethics 24:505–528

Commission of Experts for Research and Innovation (EFI) (2022) Gutachten 2022. Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit, Berlin

Data Ethics Commission (2019) Gutachten der Datenethikkommission. Report, Herausgegeben durch Datenethikkommission der Bundesregierung und Bundesministerium des Innern, für Bau und Heimat (BMI), Berlin

European Commission (2021) Ein Europa für das digitale Zeitalter: Kommission schlägt neue Vorschriften und Maßnahmen für Exzellenz und Vertrauen im Bereich der künstlichen Intelligenz vor. Press release, European Commission, Brussels

Enquete Commission Artificial Intelligence (2020) Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Abschlussbericht, Drucksache 19/23700, Berlin

European Commission (2020) Weissbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. Whitepaper, 19.02.2020. https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_de. Accessed April 22 2022

European Commission (2022) Das Programm „Digitales Europa". 14.11.2022, European Commission. https://digital-strategy.ec.europa.eu/de/activities/digital-programme. Accessed March 13 2023

European Commission (2019) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building Trust in Human-Centric Artificial Intelligence. European Commission – Shaping Europe's Digital Future. 8.04.2019, European Commission. https://digital-strategy.ec.europa. eu/en/library/communication-building-trust-human-centric-artificial-intelligence. Accessed March 26 2023

Finocchiaro G (2023) The regulation of artificial intelligence. AI Soc. https://doi.org/10.1007/s00146-023-01650-z

FLI 2023 (Future of Life Institute) (2023): The Artificial Intelligence Act. FLI. artificial intelligenceact.eu/. Accessed March 13, 2023

Flick U (2007) Designing Qualitative Research. Sage Publications, London

Foffano F, Scantamburlo T, Cortés A (2023) Investing in AI for social good: an analysis of European national strategies. AI Soc 38:479–500

German Ethics Council (2023) Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Statement, Berlin

Gervais DJ (2023) Towards an effective transnational regulation of AI. AI Soc 38:391–410. https://doi.org/10.1007/s00146-021-01310-0

Gill SP (2023) Editorial: beyond regulatory ethics. AI Soc 38:437–438. https://doi.org/10.1007/s00146-023-01657-6

Glauner, P (2023) Kommentar zum AI Act: Es droht das Aus für ChatGPT in der EU. 21.02.2023, Heise Online. https://www.heise.de/meinung/Kommentar-zum-AI-Act-Es-droht-das-Aus-fuer-ChatGPT-in-der-EU-7522179.html. Accessed March 13 2023

Gupta MP, Carlos M, Dennehy D (2021) Questioning Racial and Gender Bias in AI-based Recommendations: Do Espoused National Cultural Values Matter? Inf Syst Front 24:1465–1481

Hahn S (2023) Top researchers on the AI Act: "Overregulation poses security risk for the EU". 30.04.2023, Heise Online. https://www.heise.de/news/Top-researchers-on-the-AI-Act-Overregulation-poses-security-risk-for-the-EU-8983825.html. Accessed May 7 2023

Heesen J et al. (2020) Ethik-Briefing – Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen. Whitepaper, Herausgegeben durch Plattform Lernende Systeme – Die Plattform für Künstliche Intelligenz, Berlin

Heesen J et al. (2021) Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten. Whitepaper, Herausgegeben durch Plattform Lernende Systeme – Die Plattform für Künstliche Intelligenz, Berlin

Helberger N, Diakopoulos N (2023) ChatGPT and the AI Act. Internet Policy Rev 12:1–6

Hirsch-Kreinsen H (2023a) Das Versprechen der Künstlichen Intelligenz – Gesellschaftliche Dynamik einer Schlüsseltechnologie. Campus Verlag, Frankfurt/New York

Hirsch-Kreinsen H (2023b) Artificial Intelligence: a "promising technology." AI Soc. https://doi.org/10.1007/s00146-023-01629-w

HLEG (High-Level Expert Group on Artificial Intelligence) (2019) Ethics Guidelines for Trustworthy AI. European Commission, Brussels

HLEG (High-Level Expert Group on Artificial Intelligence) (2020) The Assessment List for Trustworthy Intelligence (ALTAI) for Self-Assessment. European Commission, Brussels

Jansen F, Cath C (2021) Algorithmic registers and their limitations as a governance practice. In: Kaltheuner F (ed) Fake AI. Meatspace Press, Manchester, pp 183–192

Kagermann H, Wilhelm U (eds) (2020) European Public Sphere – Gestaltung der digitalen Souveränität Europas. acatech IMPULS, Munich

Kersting K, Tresp V (2019) Maschinelles und Tiefes Lernen – Der Motor für „KI made in Germany". Whitepaper, Herausgegeben durch Plattform Lernende Systeme – Die Plattform für Künstliche Intelligenz, Berlin

Kleine N, Weber K (2022) Formen und Möglichkeiten gesellschaftlicher Normierung von KI. Vortragspräsentation, Ergebnisworkshop Projekt KiMeGe, Munich

KPMG Deutschland (2022) Künstliche Intelligenz ohne Gewissen? KPMG unterstützt Unternehmen bei der Entwicklung gesellschaftsverträglicher Algorithmen. 2022, KPMG. https://klardenker.kpmg.de/digital-hub/kuenstliche-intelligenz-ohne-gewissen/. Accessed April 4/5 2022

LAION (2023) Large-scale Artificial Intelligence Open Network. https://laion.ai/. Accessed May 7 2023

Lauer C (2023) ChatGPT sorgt für Aufregung – was ihr über den KI-Chatbot wissen müsst und wofür ihr ihn einsetzen könnt. 03.02.2023, Business Insider. https://www.businessinsider.de/wissenschaft/chat-gpt-sorgt-fuer-aufregung-was-ihr-ueber-den-ki-chatbot-wissen-muesst-und-wofuer-ihr-ihn-einsetzen-koennt-h/. Accessed February 13 2023

Learning Systems Platform (PLS) (2020) Zukunftsfähigkeit mit KI sichern – Ansätze für mehr Resilienz und digitale Souveränität. Positionspapier, Lernende Systeme – Die Plattform für Künstliche Intelligenz, Berlin

Metz C, Schmidt G (2023) Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'. 29.03.2023, The New York Times. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html. Accessed May 7 2023

Müller A (2022) Der Artificial Intelligence Act der EU: Ein risikobasierter Ansatz zur Regulierung von Künstlicher Intelligenz. EuZ 1:1–25

Müller-Quade J et al. (2019) Künstliche Intelligenz und IT-Sicherheit – Bestandsaufnahme und Lösungsansätze. Whitepaper, Plattform Lernende Systeme, Munich

Noll, Andreas (2019) KI ersetzt Personaler – Software analysiert Bewerber. 01.10.2019, Deutschlandfunk Nova. https://www.deutschlandfunknova.de/beitrag/job-bewerbung-kuenstliche-intelligenz-entscheidet. Accessed March 17 2022

Ossewaarde R, Erdener G (2020) National Varieties of Artificial Intelligence Discourses: Myth, Utopianism, and Solutionism in West European Policy Expectations. Comp 53:53–61

Pereira D (2021) AI Ethics Sells…But Who's Buying? 19.04.2021, Towards Data Science. https://towardsdatascience.com/ai-ethics-sells-but-whos-buying-c050054ec44. Accessed April 6 2022

Qiang V, Rhim J, Moon AJ (2023) No such thing as one-size-fits-all in AI ethics frameworks: a comparative case study. AI Soc. https://doi.org/10.1007/s00146-023-01653-w

Riegert B (2023) EU: ChatGPT spurs debate about AI regulation. 15.04.2023, dw. https://www.dw. com/en/eu-chatgpt-spurs-debate-about-ai-regulation/a-65330099. Accessed May 7 2023

Rudschies C, Schneider I, Simon J (2021) Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities. IRIE 29:1–15

Schiff D, Rakova B, Ayesh A, Fanti A, Lennon M (2021) Explaining the Principles to Practices Gap in AI. IEEE Technol Soc Mag 40:81–94

Schwartmann R (2023) Welche Regeln für ChatGPT & Co. gelten – und was wir noch tun müssen. 16.02.2023, FAZ. https://www.faz.net/aktuell/wirtschaft/digitec/chatgpt-diese-probleme-wirft-kuenstliche-intelligenz-auf-18680994.html. Accessed March 9 2023

Sharma S (2023) EU closes in on AI Act with last-minute ChatGPT-related adjustments. 28.04.2023, Computerworld. https://www.computerworld.com/article/3695009/eu-closes-in-on-ai-act-with-last-minute-chatgpt-related-adjustments.html. Accessed May 4 2023

Shrestha YR, Ben-Menahem SM, Von Krogh G (2019) Organizational Decision-Making Structures in the Age of Artificial Intelligence. Calif Manag Rev 61:66–83

SIENNA (2021) Ethics by Design and Ethics of Use approaches for Artificial Intelligence, Robotics and Big Data. The SIENNA Project, Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact. 21.03.2021, SIENNA. https://sienna-project.eu/public-consultation/ai/ethics-by-design/#:~:text=Ethics%20by%20Design%20(EbD)%20is,into%20desig n%20and%20development%20processes. Accessed March 16 2023

Steininger T (2023) Künstliche neuronale Netze: Maschinelles Lernen – vom Gehirn inspiriert. 27.01. 2023, Computerwoche. https://www.computerwoche.de/a/maschinelles-lernen-vom-gehirn-inspiriert,3551313. Accessed March 13 2023

The Economist (2023) How generative models could go wrong. 19.04.2023, The Economist. https://www.economist.com/science-and-technology/2023/04/19/how-generative-models-could-go-wrong. Accessed April 22 2023

Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, Floridi L (2022) The ethics of algorithms: key problems and solutions. AI Soc 37:215–230. https://doi.org/10.1007/s00146-021-01154-8

Urban E (2023) EU AI Act: Schiebt er ChatGPT und Co. den Riegel vor? 16.02.2023, T3n - digital pioneers. https://t3n.de/news/eu-ai-act-chatgpt-hochrisiko-1535021/. Accessed March 13 2023

Vallance C (2023) Elon Musk among Experts urging a Halt to AI Training. 30.05.2023, BBC-News. https://www.bbc.com/news/technology-65110030. Accessed April 27 2023

Van Roy V, Rossetti F, Perset K, Galindo-Romero L (2021) AI Watch – National strategies on Artificial Intelligence: A European perspective. 2021 edn. EUR 30745 EN, Publications Office of the European Union, Luxembourg

Venkina E (2021) „Die Kluft zwischen den USA und der EU im Bereich KI ist übertrieben". 29.11.2021, Big Data Insider. https://www.bigdata-insider.de/die-kluft-zwischen-den-usa-und-der-eu-im-bereich-ki-ist-uebertrieben-a-1077581/. Accessed March 15 2022

Wahlster W, Winterhalter C (2020) Deutsche Normungsroadmap – Künstliche Intelligenz. DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE, Berlin

Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, Manyika J, Niebles JC, Sellitto M, Shoham Y, Clark J, Perrault R (2021) The AI Index 2021 Annual Report. Stanford University, California, AI Index Steering Committee, Human-Centered AI Institute

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.