Discover Artificial Intelligence

Discover

Research

Artificial intelligence by any other name: a brief history of the conceptualization of "trustworthy artificial intelligence"

Charlotte Stix[1]

Received: 29 July 2022 / Accepted: 28 November 2022
Published online: 21 December 2022

OPENPublished online: 21 December 2022

1 Introduction

Recent years have seen an increase in artificial intelligence (AI) capabilities and incidents. Correspondingly, there has been an influx of government strategies, panels, dialogues and policy papers, including efforts to regulate and standardize AI systems [12, 20, 37, 52]. A first step in most of these efforts is to delineate the scope of the resulting document, typically by either outlining a range of standard technical definitions of AI [76, 85] or referencing existing scholarly work [73]. After defining their scope, many policy documents published by governments delve deeper into the 'type' of AI they wish to solicit from industry players and deploy nationally or globally. This largely serves to ensure that the strategies, policy discussions and AI-related milestones sketched within these documents are guided by a 'north star', or overarching goal. The north star should be comprehensible to all who read and implement the document. Describing the north star allows a non-technical audience to follow and partake in the relevant policy discussions, though it does not replace technical definitions. Although more could be said as to why this is being done and whether it is sensible, such discussion is outside the scope of this paper. Instead, I focus on and contextualize some of these 'north star' definitions themselves. In particular, I explore one of the most prominent recent descriptions: the EU's concept of "trustworthy AI." I explain its background, its international effects and its drawbacks in more depth. What is in a name? What is in "trustworthy AI?".

2 Other terms

To provide proper context, this section describes a number of terminologies that political decision makers have used to reference the type of AI they desire to encourage. This serves two purposes. First, it primes the subsequent investigation of the term "trustworthy AI." Second, it highlights the difficulty of choosing and solidifying an appropriate term for use in governmental contexts. Below is a non-comprehensive selection of some of the most prominently used terms within the past couple of years. It should be noted that different groups are responsible for coining and/or advocating for each term. This is to say that these terms have not necessarily been originated by governments, though they have been picked up by governmental discourse. Moreover, the terms refer to varying objectives and measures as to how to achieve those objectives. A commonality across all these terms is that they look to describe and capture AI systems that will bring benefits to society—those that presumably will make the world a better place in the near future and for future generations.

✉ Charlotte Stix, c.stix@tue.nl | [1]Eindhoven University of Technology, Eindhoven, Netherlands.

Discover Artificial Intelligence (2022) 2:26 | https://doi.org/10.1007/s44163-022-00041-5

2.1 Ethical AI

Following a wave of AI ethics documents, charters and public AI ethics discussions [44, 74, 77], one increasingly popular term is "ethical AI."

The term "ethical AI" has been widely used to refer to AI systems that are in line with our moral values [15, 17, 62]. The term has had one of the earliest, strongest and most continuous influences on public and governmental discourse. The nascent field of AI ethics is increasingly crucial as we tackle the manifold potential harms society has suffered by AI systems—and as we work to preempt potential harms. For example, the following concerns have deeply impacted various groups in recent times: algorithmic bias [6, 8, 18],transparency and explainability [26, 40],the safety of autonomous vehicles [5, 69], privacy concerns in the face of widespread surveillance [14, 39], and the economic and political effects of technological unemployment [21, 38].

It should be noted that, while most interpret "ethical AI" to refer to AI that is in line with at least a subset of common ethical considerations,[1] some may (mis)interpret it as AI that exhibits 'ethical' behavior and, by extension, is a moral agent.[2] Having said that, it is a challenge to explore "ethical AI" as a concept per se. The term's ambiguity simultaneously minimizes the true nature of the field and maximizes the appeal of the term, without creating responsibility within the user to clearly define it. Accordingly, the term has indeed been co-opted, especially in media and speaker circuit discourse. The large and shifting scope of the term may explain why other, equally popular terms have arisen. Presumably, there has been a need for terms that more clearly delineate suggested concepts and goals.

2.2 AI for good

For example, another commonly used term, particularly from the earlier days of AI policy discourse, is "AI for good." This term has been particularly appealing to discourse within industry, though it has been equally popular with governmental actors. Whether the term is positive or negative in terms of strategic messaging and impact is outside the scope of this paper.

"AI for good" has become a hallmark for the United Nations International Telecommunications Unit (UN ITU) in particular. The UN has built a digital platform around the term, which is designed to encourage discussions and projects aimed at finding practical solutions for the UN's Sustainable Development Goals (SDGs) through AI [84]. The term "AI for good" then, in this context, refers to AI systems that help solve previously identified, complex global issues for society, thus benefiting humankind [78]. Those working to develop AI for good measure their success by their AI systems' ability to help society reach certain goals.

2.3 Beneficial AI

Another popular term, initially promoted more by the research community than by governments, is "Beneficial AI."

"Beneficial AI" was spearheaded through the Future of Life Institute's Asilomar Conference on Beneficial AI. This conference also yielded one of the very first sets of AI Principles, the Asilomar AI Principles[3] which were signed by 5720 people, including 1797 AI and robotics researchers. Given the principles described in the document, the term initially appears to have had a close connection to technical AI research. In particular, it referred to topics that fall under the research field of AI safety [4, 16, 48], as well as topics related to the long-term future.[4] Outside of this research space, the term has been co-opted to broadly reference AI systems that *benefit* society and the environment, while avoiding definable and undefinable harm.

[1] Discussions around finding agreement on which issues are the most important and universally agreed upon, as well as shortcomings of and complexities in that process, are outside the scope of this paper.

[2] This interpretation is explicitly not explored or understood to be referenced throughout this paper.

[3] See: https://futureoflife.org/2017/08/11/ai-principles.

[4] See: https://www.bbvaopenmind.com/en/articles/provably-beneficial-artificial-intelligence/.

2.4 Responsible AI

A fourth commonly used term is "responsible AI." While the previously mentioned terms refer mainly to AI systems, "responsible AI" more often refers to the actions, and actors, involved in developing and deploying those systems.

"Responsible AI" seems a more sophisticated term, perhaps due to the general connotations of "responsibility." It has been used to refer to many of the mechanisms or methods by which it could feasibly be achieved, such as responsible design and development for AI [7, 24]. It seems this term most often refers to processes that result in technical achievement and meet certain standards of responsibility. Some actors, particularly industry actors, likely feel that this term is more precise than terms that invoke ethics. It may be a preferable term from a communications perspective as it, similar to "trustworthiness," references a particular kind of behavior we regard as good when it is displayed by individuals or organizations.

To a degree, all of the aforementioned terms are open to interpretation, which may be welcome (or even intended) by some actors using these terms (cf., for example, discussions of ethics washing and ethics shopping [58, 59].

But what about "trustworthy AI?" The following section summarizes this term's history and the scope of its original definition, demonstrating that, in principle, it is very clearly defined. Subsequently, Sect. 4 critically examines some downsides of the term and places its impact on the international AI governance debate in context.

3 "Trustworthy AI:" the origin story

Following the publication of its AI Strategy [27], and the aim outlined therein to put forward an ethical and regulatory framework for AI, the European Commission established an independent expert group to fulfill part of this commitment. The expert group, the High-Level Expert Group on AI (henceforth: AI HLEG), was tasked with a number of projects. Most notably and relevant to this paper, their primary task under their initial mandate was to develop ethics guidelines for AI.

The AI HLEG, composed of 52 subject experts from various sectors and fields of expertise, began a comprehensive and iterative process to establish what ultimately became a building block for AI governance in the EU. Although a majority of the work was conducted internally, the AI HLEG shared their progress in meetings open to institutional observers. They solicited feedback on their first draft of the ethics guidelines half a year into the process via the AI Alliance [80, 81], a platform through which the public and institutional actors were able to interact with the AI HLEG. In that first draft [1], the conceptualization proposed by the AI HLEG was that AI should: (1) "respect fundamental rights, applicable regulation and core principles and values, ensuring an 'ethical purpose'" and (2) "be technically robust and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm." The framework outlined to achieve trustworthy AI was built around (i) ethical purpose based on values and principles as enshrined in human rights law and other relevant charters, (ii) realization of trustworthy AI through technical and non-technical methods and (iii) an assessment list with use cases for developers, deployers and users operationalizing trustworthy AI.

Following the implementation of public feedback, the AI HLEG presented their final Ethics Guidelines for Trustworthy AI in April 2019 [2].

The final Ethics Guidelines for Trustworthy AI proposed, for the first time, a complete conceptual understanding and agreement as to what type of AI should be encouraged within the EU. While the document is strongly anchored in EU values and fundamental rights, as enshrined in the Charter of Fundamental Rights of the European Union,[5] the core concept of trustworthy AI is novel.

Trustworthy AI, in its final form, is defined as being composed of three parts. In order for an AI system to count as "trustworthy," (1) it must be lawful; that is, adhering to all legal obligations which are binding and required at that time, (2) it should be ethical; that is, adhering to and fulfilling all ethical key requirements that have been put forward in the Ethics Guidelines for Trustworthy AI [2], and (3) it should be robust, both from a technical and a social perspective. The last means that it should be robust in functionality, accurate, reliable, resilient to attack and other cybersecurity and security considerations. Equally, it should be robust within society and the environment,it should support beneficial societal processes and encourage cohesion and a well-functioning society. This corresponds to pillar 2 in the draft Guidelines [1].

Given the depth and scope of these three pillars, it is clear that the conceptualization has to do a lot of heavy lifting. After all, the second pillar alone—the ethical component—is itself composed of seven key requirements that were

[5] See: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT.

spelled out by the AI HLEG in the very same document. In order for an AI system to adhere to the second pillar alone, it must meet standards in the following areas: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-Discrimination and Fairness; Societal and Environmental Well-Being; and Accountability. These key requirements, as they are referred to in the Ethics Guidelines for Trustworthy AI [2], were themselves distilled from four core principles the AI HLEG agreed upon in correspondence with feedback from the AI Alliance and other actors: Respect for Human Autonomy, Prevention of Harm, Fairness and Explicability.

The three pillars take a lifecycle approach, requiring monitoring and adherence from the research stage to deployment and long after, even as new standards and regulations are developed. They also impose a range of active prescriptive requirements on a range of different actors, such as technical researchers, governments and users.

This paper's aim is not to evaluate to what degree these pillars are sufficient or, indeed, the best ones to use as reference points (see further work on this by e.g. [61, 75]. Rather, it looks to demonstrate that there has been a proliferation of terms similar to "trustworthy AI" in government dialogue, and that the EU was the first (and so far, only) governmental actor that has delineated its preferred term in terms of clear, identifiable, and verifiable requirements.[6] Looking ahead, the conceptual clarity and coherence of "trustworthy AI" within the work of the EU, as well as its wide adoption, may provide some insight as to the EU's influence over international AI policy making.

4 "Trust in AI," "trusted AI" or "trustworthy AI:" promises, perils and problems

While the term "trustworthy AI" now refers to a clear enumeration of values, rights and technical specifications, the word "trustworthy" may itself be reason for some concern. "Trustworthy" is, arguably, extremely open to interpretation and, as such, likely to cause unwitting confusion, or to be misinterpreted or misused, intentionally or not. Unfortunately, this ambiguity could undermine the clear definition (and the crisp intention and requirements) it is supposed to provide, as investigated in 3.1.a. The risk of misinterpretation or misuse is even higher when "trustworthy AI" is reused, amended and adopted in international policy contexts without the original relation to the three pillars and their corresponding requirements, or with only partial reference to them. The wide adoption of the term runs the risk of diluting its meaning beyond recognition and backfiring on serious policy efforts, by way of losing both the original intent and the core content the terminology is meant to encourage after all.

4.1 Come hell or waters high: is this trustworthy AI?

This subsection highlights some initial concerns around the term "trustworthy AI." It proposes two overarching concerns. First, various meanings of the term, depending on actor and context, can easily get conflated, effectively creating different meanings and understandings which can mislead. Second, the term could easily be 'washed out' of its original meaning through repetition,[7] whether or not this is the underlying intent from a strategic and self-interested perspective of an agent. In short, one concern is with the word "trustworthy" itself, another with the repeated usage of the word and its associated definition. This creates the backdrop for the subsequent review in Sect. 4.2., which examines the dissemination of the term and how it has seemingly been re-used in various political contexts, as well as the results of that usage.

4.1.1 a. Conflation

The term "trustworthy AI" conflates at least five meanings:

- trust in the proper functioning and safety of the technology;
- the technology being worthy of the trust of the humans making use of it or encountering it otherwise;
- humans making use of it or encountering it seeing the technology as trustworthy;
- humans making use of it or encountering it experiencing the technology as trustworthy;

[6] It should be mentioned that, outside of governmental dialogue, various AI-relevant research communities have developed multiple frameworks and conceptualizations. Indeed, as the AI HLEG itself was composed of various experts from diverging research fields, the conceptualization of trustworthy AI was certainly inspired by, and building on, those research communities' work.

[7] Akin to the 'telephone' word repetition game in which the original sentence gets lost after sharing it too often along a chain of individuals.

- the technology that is worthy of trust to all.

What does "trustworthy" mean in different contexts and to different actors? It is questionable whether any technology can even be trustworthy, given that this is a concept predominantly applied to human–human interactions [13]. More specifically, the term is typically applied to interactions in which we cannot be certain of the intentions of another person, but we assume they are innocuous based on the person's past actions and other social signals [49]. When we step onto a plane or into a car, we do not necessarily describe these technologies as "trustworthy." Instead, we rely on the fact that they have been sufficiently tested and passed all necessary thresholds to be safe for us to use [71, 86]. If anything, we trust the humans that have been involved in ensuring these technologies are safe [12]. By extension, we trust the institutions, organizations and processes they have set up, are involved in and are accountable to.

Calling an AI system "trustworthy" seemingly personifies the technology. Just as we may have to resort to trusting a person when we cannot know their intentions, trusting an AI system seems to imply that it has inaccessible intentions that we have no control over, and that we are fine with that status quo. This implicitly assigns human-like properties to the AI and weighs them more heavily than is desirable. Moreover, calling an AI system "trustworthy" downgrades the expectations we might rightly have about being able to access a sufficient understanding of the full complexities, mechanisms and safety implications of the AI systems we encounter, by virtue of 'assumed trust' through the AI system's trustworthiness.

Moreover, those using the term "trustworthy AI" risk conflating the meanings of trusted and trustworthy. An individual such as the Tinder Swindler (*Tinder Swindler* [documentary], Netflix 2022) appears to have been (unfoundedly) trusted but he was not actually trustworthy. As humans trusting other humans we mostly deal with incomplete sets of information about the individual. We have to infer through past actions, behaviors, social networks and adjacent signals whether engaging with an individual is safe. While the concept of trustworthy AI is intuitively appealing, "trustworthy" does not suffice in a space where we need to have a reliably and sufficiently complete set of information—either directly or validated through experts—to ensure that indeed, the technology we engage with is sufficiently safe and desirable. In particular, there is a nagging worry that the term appeals to an intrinsically human concept and strongly overemphasizes the degree to which we should trust an AI system without expert supervision and access to empirical facts.

Even for some experts of the AI HLEG, the range of interpretations of "trustworthy" and its relation to various actors in the ecosystem remain manifold and in need of clarification. ALLAI, an organization founded by three members of the AI HLEG, recently provided feedback on the European Commission's proposal for a horizontal regulation for AI stating that [3]: "First of all, in the current wording the scoring should be aimed at evaluation or classification of trustworthiness of people. While we like the term trustworthiness for obvious reasons, in this context it is vague. What is considered the trustworthiness of a person?" (p. 11).[8] Indeed, what is the trustworthiness of a person? If we can't measure the trustworthiness of a person, why would we use this term to evaluate a technology, particularly when we seek to establish an evaluative framework that has clear, contextualized and verifiable metrics? The overall goal should be to encourage AI systems that function as expected, designed and deployed for, and are legal and robust. The goal is not to encourage systems that are trustworthy in the intuitive, colloquial sense. This would also match the actual work and suggestions of the AI HLEG.

Another tempting (mis)interpretation may be that an AI system is in a way responsible for its 'trustworthiness'. It should be noted that this is somewhat speculative and not reflective of any ongoing discussion in policy discourse. The speculative risk here, especially when it comes to using this term within a lay population, is that "trustworthy AI" subtly but distinctly deflects from who and what we want to trust. The terminology is capable of being interpreted, in a very subtle way, such that the responsibility to be trustworthy falls onto the AI system itself. As previously stated, presumably, what we do want to trust is that all the technical aspects of an AI system—everything that contributes to its functioning—have been adequately tested and developed. We want to trust that these technical aspects fulfill and pass all requirements necessary, that eventual misuses have been tested for and prevented to the best of the current state of the art of the technical research available. We want to trust that the AI system has been deployed in a manner that is in line with fundamental rights, the law and ethics. Presumably who we actually want to trust are the researchers, the individuals involved in the entire lifecycle of the AI system, those deploying it on the market and those testing the systems for benchmarks, certifications and standards. We want to ensure that they have good intentions and that their work is accessible for third parties to verify the accuracy and intent of it. And the way to do this is to verify our trust by outsourcing it to existing methods, be that audits, employment contracts, existing regulation or otherwise.

[8] The use of "trustworthiness" also appears in the proposal for a horizontal regulation for AI under 5 (1) c.

4.1.2 b. Rinse and repeat: 'washing out' any distinct meaning

Finally, there is a danger that the seductive graspability and familiarity of the term "trustworthy" could be weaponized by some actors to obfuscate the development process of their AI systems. While the EU has been incredibly clear as to what obligations an AI system's life cycle needs to fulfill for the system itself to be trustworthy, the meaning of the term is becoming more vague as it becomes increasingly popularized. Therefore, it can fall prey to being "green washed" [72] or "ethics washed" [9, 58, 59]. The term may be misused to reassure consumers about an AI system that may not actually adhere to any of the expected guardrails or checks. There is no law surrounding what can and cannot be called "trustworthy AI." In fact, the terminology might end up being used as a marketing tool more than to convey factual information about the AI system. In doing so, economic appeal may usurp ethical and legal rationales.

This is particularly salient as a consideration for the review undertaken in the next section, 4.2. In particular, it helps to place this concern within real-world circumstances, demonstrating that in fact the term already seems to be used as a marketing tool despite good intentions, and underlining why the high-level adoption of this term in international policy discourse, as discussed in Sect. 4.2.4d, may actually backfire on the original ambition of this term.

4.2 Are you for real? It all boils down to memetic appeal

Short of finding a new term, all of the points put forward under Sect. 4.1. suggest that it could be (at minimum) worth changing the term when used in its original sense (i.e. to indicate adherence to all three pillars as advocated by the AI HLEG). For example, we could adopt the term "trustworthy AI™." Such a division would allow us to effectively distinguish between what the original term conceptualizes versus what people may change it to mean. "Trustworthy AI™" in its original composition appears to have had significant appeal to policymakers across the globe as Sects. 4.2.2b–4.2.4d will highlight similar terms, ideas and concepts that have populated the policy discourse in recent times.

The following sections build on the aforementioned concerns about the conflation of meanings. In particular, they build the case that the term is prone to being 'washed clear' of its original meaning. In doing so, it investigates the shift "trustworthy AI™" has undergone in international policy discourse and how the term has been (mis)appropriated and changed throughout the course of its adaptation into different contexts.

In the following sections, I establish the salience and timeliness of this discussion. First, I briefly introduce two regulatory developments to highlight the rapidly changing strategic and political landscape within which this term is being used. I will then highlight a number of relevant international policy efforts that seem to make use of versions of "trustworthy AI™" and are adjacent in intent to the EU's original ambition. This will provide a better understanding of (1) how the EU's conceptualization may have shaped international conversation and (2) how the term has been (re)used to hold multiple meanings and was consequently watered down to a suitcase word [57], void of specific meaning, content and subsequently, actionable intent.[9]

4.2.1 a. Political shifts: salience and timeliness

Recent research indicates that the number of bills passed into law containing references to "AI" rose from 1 in 2016 to 18 in 2021 across 25 countries [22]. This includes approaches to tackle ethical concerns related to AI. Over the past years, we have seen an increase in concern about the use of AI-based technology in hiring decisions due to their opaqueness and associated ethical issues such as a lack of ability to challenge the algorithm, potential bias, etc. One legal example that tackles this is the Artificial Intelligence Video Interview Act (820 [89]).[10] It is the first US state law regulating the use of AI during the evaluation of prospective employees' interviews. Depending on the algorithm and the intellectual property (IP), it is often difficult or impossible to gain a full understanding of the reasoning that leads to an applicant's

[9] Without actionable intent it will be difficult, if not impossible, to hold governments and industry accountable if they fail to live up to their promises. They will always be able to minimize what the term means. It should be noted that this is happening with multiple AI-related terms and not only with the term that is the focus of the paper. Other terms that come to mind are "explainability" and "accountability," which can denote quite different things depending on context and audience and are often used without any clear reference point, leading to confusion for the reader.

[10] It should be noted that the law does not define 'artificial intelligence' — creating difficulties to clearly delineate which systems the law applies to — and that the law solely addresses artificial intelligence-based technology used in videos recorded of the interview by the employer.

final outcome, rank or score. This is particularly relevant to requirement II of the Artificial Intelligence Video Interview Act, denoting that each job applicant shall be informed of how the AI-based system works and what characteristics it uses to evaluate candidates. In short, the law requires employers to: (i) notify applicants in a written format that AI may be used to analyze their video interview, (ii) give the applicants information about the workings of the AI and the characteristics it uses for evaluation of the interview, and (iii) obtain the applicant's consent to use the aforementioned artificial intelligence. Applicants can request the deletion of their video file and employers are prohibited to share the video file beyond those actors necessary to evaluate it.

More recently, the EU has become the first governmental actor to put forward a horizontal regulatory framework for high-risk AI systems and a ban for certain AI systems with the AI Act [31]. In this regulatory proposal, a number of proposed legal obligations have been outlined which a high-risk AI system must fulfill through a conformity assessment before it can be deployed on the EU market.[11] An AI system is considered "high-risk" if it falls under certain categories outlined in Annex III of the AI Act (such as certain areas of access to education or employment).[12] This regulatory proposal is currently discussed in the European Parliament, the Council of the European Union and the European Commission. Amendments have already been proposed by various member states holding the presidency of the Council of the EU, ranging from tackling general purpose AI systems to real world testing. It has also received over 3000 tabled amendments in the European Parliament following the first published report [32] on the AI Act by the two lead committees on the file, the Committee on Internal Market and Consumer Protection and the Committee on Civil Liberties, Justice and Home Affairs.

Developments such as the aforementioned corroborate the overall sense that there is increased attention and concern about AI and its impacts on society from governments, policymakers and legislators alike. With the increased discussion, it is increasingly important to ensure that concepts, terminology, methods and measures are accurate and coherent.

4.2.2 b. EU-adjacent efforts

Many EU member states have adopted the EU's concept for use within their own AI strategies and policy documents. These member states include Czechia (Czech [19], Luxembourg [50], Malta [53–56]), and the Netherlands [63]. This is unsurprising given (a) the novelty of the concept at a time when many countries did not yet have fully fleshed-out AI strategy and (b) the fact that the EU mandated a similar and cooperative approach to AI governance to combat fragmentation, as outlined in the Declaration on Cooperation [29] and the EU's Coordinated Plan [28].

Relatedly, the recent AI Act [31] indicates that standards could play a crucial role in the conformity assessment procedures of high-risk AI systems. It leaves scope for standards or technical specifications to replace matching aspects in the conformity assessment, which high-risk AI systems would need to undergo otherwise. Currently, no matching standards exist. However, in light of the overall discussion in this paper, many of the legal obligations in the AI Act's [31] conformity assessment match the areas under "trustworthy AI™". This means that the original conceptualization has been highly influential beyond ethical considerations, feeding into regulatory and standardization efforts. It is noteworthy then, that standardization committees have now started working on trustworthy AI as a topic.

One group that has adopted the term at a high level is the ISO Committee IEC TR 24,028:2020[13] on "Information technology—Artificial intelligence—Overview of trustworthiness in artificial intelligence." This working group predominantly focuses on areas related to the legal obligations for the conformity assessment outlined in the AI Act under Art. 11, "Annex IV Technical Documentation," and Art. 15, "Accuracy, Robustness and Cybersecurity." This matches at least one of the ethical key requirements outlined in the Ethics Guidelines for Trustworthy AI [2] and demonstrates how the term's original conceptualization has affected concrete policy action. Moreover, it indicates that standards committees have adopted some of the terminology and thinking of the EU to, at the very least, preempt upcoming regulation such as the AI Act [31].

Other standardization efforts, such as the US Senate Bill "S.1849—Leadership in Global Tech Standards Act of 2021" [88] do not make any reference to trustworthy AI.

[11] Interestingly, the legal obligations in the AI Act closely match the seven key requirements outlined in the Ethics Guidelines for Trustworthy AI proposed by the AI HLEG, presented earlier in this paper.

[12] Interestingly, the legal obligations in the AI Act closely match the seven key requirements outlined in the Ethics Guidelines for Trustworthy AI proposed by the AI HLEG, presented earlier in this paper.

[13] See: https://www.iso.org/standard/77608.html.

4.2.3 c. International partnerships, agreements and cooperation pipelines

Many international partnerships have used the term since its inception, often diluting and modifying the original meaning.

The OECD Recommendations on AI [70], signed by over 35 countries, recognize that "the trustworthiness of AI systems," (p. 6) is a key factor of AI diffusion and consider it vital to foster the "adoption of trustworthy AI in society and to turning AI trustworthiness into a competitive parameter in the global marketplace," (p. 6) Not only is the adoption of the original term evident, it is, moreover, clear that the trustworthiness of AI systems is seen as a competitive advantage. This matches policy documents from the EU, such as the Communication on Building Trust in Human Centric AI [30]. There, the EU envisions its approach to trustworthy and human-centric AI as one that strengthens its reputation for safe, reliable and ethical products. At the same time, although the OECD Recommendations on AI [70] match many of the original requirements outlined in the Ethics Guidelines for Trustworthy AI [2], it does not match the full scope of "trustworthy AI™."

The 2019 G20 Ministerial Statement on Trade and Digital Economy [42] reflects the EU's vision as well. In particular, the section on the G20 AI Principles [42] focuses on the "responsible stewardship of trustworthy AI," (p. 1) and "national policies and international co-operation for trustworthy AI," (p. 3). These Principles refer back to the OECD Recommendations on AI [70], and do not define trustworthy AI. As such, they have completely detached themselves from the original meaning of "trustworthy AI™," adopting what appears to be an intermediary use of the term without its original context.

The 2022 UNESCO Recommendations on the Ethics of Artificial Intelligence [82] also mention that the "the trustworthiness and integrity of the life cycle of AI systems is essential to ensure that AI technologies will work for the good of humanity, individuals, societies and the environment and ecosystems, and embody the values and principles set out in this Recommendation," (p. 18).

Most recently, the term has permeated international cooperation discourse, as evidenced by the EU-US Trade and Technology Council's inaugural Pittsburgh statement [33]. This states that "The European Union and the United States affirm their willingness and intention to develop and implement trustworthy AI and their commitment to a human-centered approach that reinforces shared democratic values and respects universal human rights, which they have already demonstrated by endorsing the OECD Recommendation on AI." (p.11). It is interesting that despite the EU being one of the two main leads in this discourse, the document itself refers back to the OECD Recommendations on AI [70] and not to the Ethics Guidelines for Trustworthy AI [2] when referring to trustworthy AI.

4.2.4 d. International actors

Beyond international fora, AI strategies of international powers such as the United States have equally explored versions of trustworthy AI. There appears to have been a distinct uptake in the use of the term in US policy documents subsequent to the publication of the draft version of the final Ethics Guidelines for Trustworthy AI [1].[14]

A preliminary review compared and contrasted documents published before and after the draft Ethics Guidelines for Trustworthy AI [1], which first mentioned the concept of trustworthy AI. The documents covered were the following: Preparing for the Future of Artificial Intelligence [67], The National Artificial Intelligence Research and Development Strategic Plan [68], Artificial Intelligence, Automation, and the Economy [34], The FUTURE of Artificial Intelligence Act of 2017 [83], the Algorithmic Accountability Act of 2019 [46]; Supporting the development of guidelines for ethical development of Artificial Intelligence [45]; notes from the Office of Science and Technology's Select Committee on Artificial Intelligence's inaugural meeting June 27, 2018; AI Principles[15]; Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Defense [23], National Security Strategy of the United States of America [87], Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity [25], and the request for comments on a Draft Memorandum to the Heads of Executive Departments and Agencies on the subject of 'Guidance for Regulation of Artificial Intelligence Applications'.[16]

[14] However, given the number of similar sounding terminologies that have been in use, including aspects discussed in Sect. 2 of this paper, it is difficult to say with certainty whether the US policy space's frame of reference has indeed been shaped by the EU or whether they decided to use this term independently.

[15] See: https://epic.org/wp-content/uploads/privacy/ai/WH-AI-Select-Committee-First-Meeting.pdf.

[16] See: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

It should be noted that none of these documents published prior to the draft Ethics Guidelines for Trustworthy AI [1] mention a variation of 'trustworthy AI', this includes two reviewed documents put forward after the publication of the Ethics Guidelines for Trustworthy AI [2].

In publications under the Trump administration, trustworthy AI as a concept notably begins featuring in the 2019 Interim report of the National Security Commission on Artificial Intelligence [66], the Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government [36]; the National Artificial Intelligence Research and Development Strategic Plan: 2019 Update [65], and in the NIST's U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools [64]. The Executive Order on Maintaining American Leadership in Artificial Intelligence [35], which was published prior to the final Ethics Guidelines for Trustworthy AI [2] but subsequent to their draft version [1] does not use the term "trustworthy AI." Yet, it comes close by referencing the "development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies" (p. 3970).

Similar to other documents introduced under Sect. 4.2. many international policy documents ended up using "trustworthy AI" without properly defining it or referencing the EU's definition. For example, the Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government [36] has "trustworthy AI" in its title, but lacks concrete conceptualization throughout the the text. On the other hand, the NIST's U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools [64] mentions "reliable, robust, and trustworthy AI technology development" (p. 4, p. 22). While it never concretizes the term, "reliability" and "robustness" match a subset of the original "trustworthy AI's™" three pillars (i.e. pillar three).

Looking at AI governance efforts in China, according to CSET's translation of the 2021 Ethical Norms for New Generation Artificial Intelligence[17] (original text source: [60]) published by the PRC's Ministry of Science and Technology, this policy document references trustworthiness at large. In particular, Art. IV of the document outlines norms for "Assurance of Controllability and Trustworthiness." However, it does not use the term "trustworthy AI" or refer to the EU's efforts in that space.

4.3 Lessons

There appear to be two intertwined lessons we can draw from the preceding sections. First, given the rapid rise in AI strategies, there was a significant policy vacuum that raised new considerations. This contributed to a situation in which governments and policymakers were under significant time pressure to develop relevant discourse. This, in turn, made it more appealing to align work with that of others who faced similar issues and to handle them in a similar manner, be that in concept or content. One area where this 'copy-paste' discourse was especially evident was in the development of Ethics Guidelines. Because all actors arrived at similar conclusions and were inconspicuously inspired by similar texts, almost all Ethics Guidelines ended up developing a similar subset of areas [43, 47].[18] Generally speaking, it makes sense that well thought-out and researched ideas would proliferate similar discourse and inspire adjacent strategies and documents. In many cases, this may even be a good thing, particularly if the original concept or idea is robust and translated into many different documents by virtue of reference or adaptation, without it being watered down or amended. A similar process applied to the conceptualization of what overarching class of AI systems governments wish to encourage, and thus "trustworthy AI™" inspired a class of efforts to use a variation of the term in their own documents and processes, for better or worse.

Second, the preceding sections demonstrate that the EU has an opportunity to shape the AI policy space. Section 4 supports the idea that the EU has been wielding some degree of soft power with "trustworthy AI™" (albeit only in description and not in content). This may be indicative of the vacuum with regards to good AI policy approaches, as described above, as well as of the memetic appeal of the original term.

The belief that the EU has some capability of influencing non-EU actors to adopt its laws and policies is often referred to as the "Brussels effect" (coined by Anu Bradford [11] and similar to the California effect[19] in the US).[20] Extrapolating from the memetic effect the term "trustworthy AI" has had, it will be interesting to see the degree to which policy proposals or

[17] See: https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/

[18] It should be noted, however, that it is difficult to capture an entire field and that this coherence may also be due to a simplification and accessibility concern for audiences.

[19] See: https://en.wikipedia.org/wiki/California_effect.

[20] One of the most commonly cited examples of the Brussels Effect is the European General Data Protection Regulation which California emulated when it passed the CCPA: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375.

regulatory proposals—such as the AI Act [31]—will shape international policy discourse and development. For example, with regard to the AI Act [31], there is already some concrete evidence that supports the Brussels effect. In particular, the Canadian government recently published their Artificial Intelligence and Data Act [10], a draft act to regulate AI. This draft act clearly orients itself on the AI Act [31] by taking a risk-based approach to regulating AI and proposing requirements that a select group of AI systems would need to adhere to (described as "high impact", in the AI Act these would be described as "high-risk"). Most recently, the Brussels effect has been explored in scholarly discourse evaluating how the EU may shape strategic regulatory interventions internationally against a specific set of criteria [79].

It should be noted for completeness that the AI Act itself builds on previous formulations of risk-based governance approaches, such as the German national AI strategy [41, 51].

5 Conclusion

This paper provides the background of the development, conceptualization, and proliferation of the term "trustworthy AI," as advanced by the EU in government discourse. It elaborates on similar terminologies that arose during the same time frame and puts forward a select number of concerns with regard to the term. In reviewing various international efforts, both collaborative and individual, this paper illustrates that the term has had broad appeal to policymakers across and outside of the EU. While it is too early to tell, it is likely that the memetic impact of this term, and the EU's first mover advantage in defining it, will be replicated in the EU's more consequential policy and regulatory efforts, most notably the AI Act [31], suggesting a possible Brussels effect.

Author contributions The author wrote the entire manuscript and conducted the associated research. All authors read and approved the final manuscript.

Declarations

Competing interests The authors declare no competing interests.

References

1. AI HLEG. Draft ethics guidelines for trustworthy artificial intelligence. Independent high-level expert group on artificial intelligence. 2018. https://www.euractiv.com/wp-content/uploads/sites/2/2018/12/AIHLEGDraftAIEthicsGuidelinespdf.pdf.
2. AI HLEG. Ethics guidelines for trustworthy artificial intelligence. Independent high-level expert group on artificial intelligence. 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
3. ALLAI. EU proposal for artificial intelligence act, analysis and recommendation. 2021. https://allai.nl/wp-content/uploads/2021/08/EU-Proposal-for-Artificial-Intelligence-Act-Analysis-and-Recommendations.pdf.
4. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv [cs.AI]. arXiv. 2016. http://arxiv.org/abs/1606.06565.
5. Anderson JM, Kalra N, Stanley K, Sorensen P, Samaras C, Oluwatola TA. Autonomous vehicle technology: a guide for policymakers. RAND Corporation. 2016. https://www.rand.org/pubs/research_reports/RR443-2.html.
6. Barocas S, Selbst AD. Big data's disparate impact. California Law Review. 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.
7. Arrieta B, Alejandro N-R, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Int J Inform Fusion. 2020;58(June):82–115.
8. Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: the state of the art. Sociological Methods & Research, July. 2018; 004912411878253.
9. Bietti E. From ethics washing to ethics bashing: a moral philosophy view on tech ethics. J Soc Comput. 2021;2(3):266–83.
10. Bill C-27. An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts. House of Commons of Canada. Minister of Innovation, Science and Industry. 2022. https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading.

11. Bradford A. The Brussels effect: how the European union rules the world. Oxford: Oxford University Press; 2020.
12. Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv [cs.CY]. arXiv. 2020. http://arxiv.org/abs/2004.07213.
13. Bryson. No one should trust artificial intelligence. Sci Technol Innov Govern. 2018; 11: 14
14. Calo R. Peeping HALs: making sense of artificial intelligence and privacy. EJLS Eur J Legal Stud. 2010;2(3):168–92.
15. Christian B. The alignment problem: how can machines learn human values? London: Atlantic Books; 2021.
16. Christiano L, Brown. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences.
17. Coeckelbergh M. AI ethics. Cambridge: MIT Press; 2020.
18. Crawford K, Dobbe R, Dryer T, Fried G, Green B, Kaziunas E, Kak A, et al. AI now 2019 report. AI Now Institute. 2019. https://ainowinstitute.org/AI_Now_2019_Report.pdf.
19. Czech Republic. National Artificial Intelligence Strategy of the Czech Republic. Ministry of Industry and Trade. 2019. https://www.mpo.cz/assets/en/guidepost/for-the-media/press-releases/2019/5/NAIS_eng_web.pdf.
20. Dafoe. AI governance: opportunity and theory of impact. Effective altruism forum, 17th September. 2020.
21. Danaher J. Automation and utopia: human flourishing in a world without work. Cambridge: Harvard University Press; 2019.
22. Daniel Z, Maslej N, Brynjolfsson E, Etchemendy J, Lyons T, Manyika J, Ngo H, Niebles JC, Sellitto M, Sakhaee E, Shoham Y, Clark J, Perrault R. The AI Index 2022 annual report. AI Index Steering Committee. Stanford University. Human-Centered Artificial Intelligence. 2022. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf.
23. DIB. AI principles: recommendations on the ethical use of artificial intelligence by the Department of Defense. Defense Innovation Board. 2019. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.
24. Dignum V. Responsible artificial intelligence: how to develop and use ai in a responsible way. Berlin: Springer Nature; 2019.
25. DoD. Summary of the 2018 Department of Defense Artificial Intelligence Strategy: harnessing AI to advance our security and prosperity. United States Department of Defense. 2018. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.
26. Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. arXiv:1710.00794 [cs], October. 2017. http://arxiv.org/abs/1710.00794.
27. European Commission. Communication from the Commission—Artificial Intelligence for Europe (COM(2018) 237 final). Brussels: European Commission, 2018a. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.
28. European Commission. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions—Coordinated Plan on Artificial Intelligence (COM/2018/795 final). Brussels: European Commission, 2018b, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:795:FIN.
29. European Commission. (Digital Day) Declaration on Cooperation on Artificial Intelligence. European Commission website—JRC Science Hub—Communities. 2018c. https://ec.europa.eu/jrc/communities/en/community/digitranscope/document/eu-declaration-cooperation-artificial-intelligence.
30. European Commission. 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 (COM/2019/168 final). Brussels: European Commission. 2019. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168.
31. European Commission. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM/2021/206 final). Brussels: European Commission. 2021. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.
32. European Parliament. Draft report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9–0146/2021 – 2021/0106(COD)). Committee on the Internal Market and Consumer Protection Committee on Civil Liberties, Justice and Home Affairs. 2022. https://www.europarl.europa.eu/doceo/document/CJ40-PR-731563_EN.pdf.
33. EU-US Trade and Technology Council. Pittsburgh Statement. Joint Declaration. 2021. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_21_4951.
34. Executive Office of the President. Artificial intelligence, automation, and the economy. 2016. https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF.
35. Executive Office of the President. E.O. 13859. Executive Order on Maintaining American Leadership in Artificial Intelligence. 2019. https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.
36. Executive Office of the President. E.O. 13960. Executive order on promoting the use of trustworthy artificial intelligence in the Federal Government. 2020. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.
37. Fischer S-C, Leung J, Anderljung M, O'keefe C, Torges S, Khan SM, Garfinkel B, et al. AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment." Accessed June 1, 2022. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-%E2%80%93-Fischer-et-al.pdf.
38. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? Technol Forecast Soc Chang. 2017;114:254–80.
39. Gasser U. Recoding privacy law: reflections on the future relationship among law, technology, and privacy. 2016. https://harvardlawreview.org/2016/12/recoding-privacy-law-reflections-on-the-future-relationship-among-law-technology-and-privacy/.
40. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, Crawford K. Datasheets for datasets. Commun ACM. 2021;64(12):86–92.
41. German Federal Government. Artificial Intelligence Strategy of the German Federal Government. 2020. https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf.
42. G20. Ministerial Statement on Trade and Digital Economy. 2019. https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf.

43. Hagendorff T. The ethics of Ai ethics: an evaluation of guidelines. Minds Mach. 2020; 1–22.
44. Hagerty A, Rubinov I. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv [cs.CY]. arXiv. 2019. http://arxiv.org/abs/1907.07892.
45. H.Res.153. Supporting the development of guidelines for ethical development of artificial intelligence. House resolution 153. 2019. https://www.congress.gov/bill/116th-congress/house-resolution/153/text.
46. H.R.2231. Algorithmic Accountability Act of 2019. House Bill 2231. 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231/text.
47. Jobin A, Ienca M, Vayena E. Artificial intelligence: the global landscape of ethics guidelines. arXiv [cs.CY]. arXiv. 2019. http://arxiv.org/abs/1906.11668.
48. Leike J, Martic M, Krakovna V, Ortega PA, Everitt T, Lefrancq A, Orseau L, Legg S. AI safety gridworlds. arXiv [cs.LG]. arXiv. 2017. http://arxiv.org/abs/1711.09883.
49. Lockey S, Gillespie N, Holm D, Someh IA. A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. In Hawaii International Conference on System Sciences 2021 (HICSS-54). 2021. https://aisel.aisnet.org/hicss-54/os/trust/2/.
50. Luxembourg. Artificial Intelligence: a strategic vision for Luxembourg. The Government of the Grand Duchy of Luxembourg. 2019. https://digital-luxembourg.public.lu/sites/default/files/2019-05/AI_EN.pdf.
51. Lütge C, Hohma E, Boch A, Poszler F, Corrigan C. White paper—on a risk-based assessment approach to AI Ethics Governance. IEAI, 2022. https://www.ieai.sot.tum.de/wp-content/uploads/2022/06/IEAI-White-Paper-on-Risk-Management-Approach_2022-FINAL.pdf.
52. Maas MM. Aligning AI Regulation to Sociotechnical Change. 2021. https://doi.org/10.2139/ssrn.3871635.
53. Malta. Malta the ultimate AI Launchpad: a strategy and vision for Artificial Intelligence in Malta 2030. 2019a. https://malta.ai/wp-content/uploads/2019/11/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf.
54. Malta. Malta towards an AI strategy: high-level document for public consultation. 2019b. https://malta.ai/wp-content/uploads/2019/04/Draft_Policy_document_-_online_version.pdf.
55. Malta. MCAST: Artificial Intelligence strategy: roadmap 2025. Malta College of Arts, Science & Technology. 2019c. https://www.mcast.edu.mt/wp-content/uploads/AI-Strategy_Final.pdf.
56. Malta. Malta towards trustworthy AI: Malta's Ethical AI Framework. 2019d. https://malta.ai/wp-content/uploads/2019/10/Malta_Towards_Ethical_and_Trustworthy_AI_vFINAL.pdf.
57. Minsky M. The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind. New York: Simon and Schuster; 2007.
58. Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L. Ethics as a service: a pragmatic operationalisation of AI ethics. Mind Mach. 2021;31(2):239–56.
59. Morley J, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L. Operationalising AI ethics: barriers, enablers and next steps. AI Soc. 2021. https://doi.org/10.1007/s00146-021-01308-8.
60. MOST. Ethical Norms for New Generation Artificial Intelligence. The National New Generation Artificial Intelligence Governance Specialist Committee (国家新一代人工智能治理专业委员会). PRC Ministry of Science and Technology (MOST; 科学技术部; 科技部). 2019. http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.
61. Mökander J, Juneja P, Watson DS, et al. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?. Minds Mach. 2022. https://doi.org/10.1007/s11023-022-09612-y.
62. Müller. Ethics of Artificial Intelligence 1. The Routledge Social Science Handbook of AI. https://doi.org/10.4324/9780429198533-9/ethics-artificial-intelligence-1-vincent-müller.
63. Netherlands. Strategisch Actieplan voor Artificiële Intelligentie. Ministerie van Economische Zaken en Klimaat. 2019. https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voorartificiele-intelligentie/Rapport+SAPAI.pdf.
64. NIST. National Institute of Standards and Technology, US Department of Commerce. U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools. 2019. https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.
65. NITRD. National Artificial Intelligence Research and Development Strategic Plan: 2019 update. A report by the Select Committee on Artificial intelligence of the National Science and Technology Council. 2019. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.
66. NSCAI. Interim report of the National Security Commission on Artificial Intelligence, National Security Commission on Artificial Intelligence. 2019. https://epic.org/wp-content/uploads/foia/epic-v-ai-commission/AI-Commission-Interim-Report-Nov-2019.pdf.
67. NSTC. Preparing for the Future of Artificial Intelligence. Executive Office of the President. National Science and Technology Council. Subcommittee on Machine Learning and Artificial Intelligence. 2016a.https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
68. NSTC. The National Artificial Intelligence Research and Development Strategic Plan. Executive Office of the President. National Science and Technology Council. Networking and Information Technology Research and Development Committee. 2016b. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf.
69. Nyholm S, Smids J. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? Ethic Theory Moral Pract Int Forum. 2016;19(5):1275–89.
70. OECD. Recommendation of the Council on Artificial Intelligence. C(2019)34, C/MIN(2019)3/FINAL, OECD/LEGAL/0449. 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
71. ÓhÉigeartaigh SS, Whittlestone J, Liu Y, Zeng Yi, Liu Z. Overcoming barriers to cross-cultural cooperation in AI Ethics and governance. Philos Technol. 2020;33(4):571–93.
72. Ramus CA, Montiel I. When are corporate environmental policies a form of greenwashing? Bus Soc. 2005;44(4):377–414.
73. Russel S, Norvig P. A modern approach. Prentice Hall Upper Saddle River, NJ. https://www.sti-innsbruck.at/sites/default/files/Knowledge-Representation-Search-and-Rules/Russel-&-Norvig-Inference-and-Logic-Sections-7.pdf.

74. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc. 2020;19(1):61–86.
75. Salo-Pöntinen H, Saariluoma P. Reflections on the human role in AI policy formulations: how do national AI strategies view people?. Discov Artif Intell. 2, Article 3. 2022. https://doi.org/10.1007/s44163-022-00019-3.
76. Samoili S, Cobo ML, Gomez E, De Prato G, Martinez-Plumed F, Delipetrev B. AI Watch. Defining Artificial Intelligence. Towards an Operational Definition and Taxonomy of Artificial Intelligence. 2020. https://eprints.ugd.edu.mk/28047/.
77. Schiff D, Biddle J, Borenstein J, Laas K. What's next for AI ethics, policy, and governance? A global overview. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 153–58. AIES '20. New York, NY, USA: Association for Computing Machinery; 2020.
78. SDG. Sustainable development goals. 2015. https://sdgs.un.org/goals.
79. Siegmann C, Anderljung M. The Brussels effect and artificial intelligence: how EU regulation will impact the global AI market. 2022. https://uploads-ssl.webflow.com/614b70a71b9f71c9c240c7a7/62fbe1c37eff7d304f0803ac_Brussels_Effect_GovAI.pdf.
80. Stix C. Actionable principles for artificial intelligence policy: three pathways. Sci Eng Ethics. 2021a;27(1):15.
81. Stix C. The Ghost of AI Governance Past, Present and Future: AI Governance in the European Union. arXiv [cs.CY]. arXiv. 2021b. http://arxiv.org/abs/2107.14099.
82. UNESCO. Recommendations on the Ethics of Artificial Intelligence. Adopted November 2021. 2022. https://unesdoc.unesco.org/ark:/48223/pf0000381137.
83. U.S. Senate. The FUTURE of Artificial Intelligence Act of 2017. BAG17H16. 2017. https://www.cantwell.senate.gov/imo/media/doc/The%20FUTURE%20of%20AI%20Act%20Introduction%20Text.pdf.
84. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, Felländer A, Langhans SD, Tegmark M, Nerini FF. The role of artificial intelligence in achieving the sustainable development goals. Nat Commun. 2020;11(1):1–10.
85. Wang P. On defining artificial intelligence. J Artif Gen Intell. 2019. https://sciendo.com/downloadpdf/journals/jagi/10/2/article-p1.pdf.
86. Winter PM, Eder S, Weissenböck J, Schwald C, Doms T, Vogt T, Hochreiter S, Nessler B. Trusted artificial intelligence: towards certification of machine learning applications. 2021. arXiv [stat.ML]. arXiv. http://arxiv.org/abs/2103.16910.
87. White House. National Security Strategy of the United States of America. 2017. https://trumpwhitehouse.archives.gov/wp-content/uploads/2017/12/NSS-Final-12-18-2017-0905.pdf.
88. 117th Congress. S.1849—Leadership in Global Tech Standards Act of 2021. 2021–2022. https://www.congress.gov/bill/117th-congress/senate-bill/1849.
89. 820 ILCS 42/. Artificial intelligence video interview act. 2020. https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68.