8 Appendix

Document 1: Artificial Intelligence for Europe

No.	Page	Citation	Category
1	2	Building on this strong political endorsement, it is time to make significant efforts to ensure that: […] No one is left behind in the digital transformation. […] New technologies are based on values.	E1
2	2	This is where the EU's sustainable approach to technologies creates a competitive edge, by embracing change on the basis of the Union's values [5]. As with any transformative technology, some AI applications may raise new ethical and legal questions, for example related to liability or potentially biased decision-making. The EU must therefore ensure that AI is developed and applied in an appropriate framework which promotes innovation and respects the Union's values and fundamental rights as well as ethical principles such as accountability and transparency.	E1, E2
3	3	Ensure an appropriate ethical and legal framework, based on the Union's values and in line with the Charter of Fundamental Rights of the EU. This includes forthcoming guidance on existing product liability rules, a detailed analysis of emerging challenges, and cooperation with stakeholders, through a European AI Alliance, for the development of AI ethics guidelines	E1, E3
4	12	To manage the AI transformation, workers whose jobs are changing or may disappear due to automation must have every opportunity to acquire the skills and knowledge they need, to master new technology and be supported during	E1

		labour market transitions. This anticipatory approach and focus on investing in people is a cornerstone of a human-centric, inclusive approach to AI, and will require a significant investment.	
5	12	More women and people of diverse backgrounds, including people with disabilities, need to be involved in the development of AI, starting from inclusive AI education and training, in order to ensure that AI is non-discriminatory and inclusive.	E1
6	12	The importance of ethics in the development and use of new technologies should also be featured in programmes and courses.	E1
7	13	An environment of trust and accountability around the development and use of AI is needed.	E1
8	13, 14	The values set out in Article 2 of the Treaty on European Union constitute the foundation of the rights enjoyed by those living in the Union. In addition, the EU Charter of Fundamental Rights brings together all the personal, civic, political, economic and social rights enjoyed by people within the EU in a single text.	E1, E4
9	14	The General Data Protection Regulation ensures a high standard of personal data protection, including the principles of data protection by design and by default. […] The Commission will closely follow the Regulation's application in the context of AI and calls on the national data protection authorities and the European Data Protection Board to do the same.	E1
10	14	This is essential as citizens and businesses alike need to be able to trust the technology they interact with, have a	E1

		predictable legal environment and rely on effective safeguards protecting fundamental rights and freedoms.	
11	14	To further strengthen trust, people also need to understand how the technology works, hence the importance of research into the explainability of AI systems. Indeed, in order to increase transparency and minimise the risk of bias or error, AI systems should be developed in a manner which allows humans to understand (the basis of) their actions.	E1
12	14	Like every technology or tool, AI can be used to positive but also to malicious ends. Whilst AI clearly generates new opportunities, it also poses challenges and risks, for example in the areas of safety and liability, security (criminal use or attacks), bias51 and discrimination.	E1, E2
13	14	As a first step to address ethical concerns, draft AI ethics guidelines will be developed by the end of the year, with due regard to the Charter of Fundamental Rights of the European Union.	E1, E3
14	15	The draft guidelines will address issues such as the future of work, fairness, safety, security, social inclusion and algorithmic transparency. More broadly, they will look at the impact on fundamental rights, including privacy, dignity, consumer protection and non-discrimination.	E1, E3
15	15	The emergence of AI, in particular the complex enabling ecosystem and the feature of autonomous decision-making, requires a reflection about the suitability of some established rules on safety and civil law questions on liability.	E1, E3
16	1	Amid fierce global competition, a solid European framework is needed.	E5

17	2	The EU can lead the way in developing and using AI for good and for all, building on its values and its strengths.	E5
18	2	The EU is also well placed to lead this debate on the global stage.	E5
19	2	This is how the EU can make a difference – and be the champion of an approach to AI that benefits people and society as a whole.	E5
20	14	The EU has a strong and balanced regulatory framework to build on, which can set the global standard for a sustainable approach to this technology.	E5
21	18	The EU will continue to encourage discussions on AI and its various dimensions – including research and innovation cooperation as well as competitiveness – in such fora. It will promote the use of AI, and technologies in general, to help solve global challenges, support the implementation of the Paris Climate agreement and achieve the United Nations Sustainable Development Goals.	E5
22	18	The EU can make a unique contribution to the worldwide debate on AI based on its values and fundamental rights.	E5
23	19	The main ingredients are there for the EU to become a leader in the AI revolution, in its own way and based on its values.	E5

Document 2: Coordinated Plan on Artificial Intelligence

No.	Page	Citation	Category
24	1	The Commission proposed an approach that places people at the centre of the development of AI (human-centric AI) and encourages the use of this powerful technology to help solve the world's biggest challenges: from curing diseases to fighting climate change and anticipating natural disasters, to	E1, E3

		making transport safer and fighting crime and improving cybersecurity.	
25	4	Strengthening excellence in trustworthy AI technologies and broad diffusion	
26	6	Given the disruptive nature of many of the technological advances, policy-makers will develop strategies to deal with employment changes in order to ensure inclusiveness, as the pace with which some jobs will disappear and others appear is likely to accelerate, while business models and the way tasks or jobs are performed will change.	E1, E2
27	6	Further developments in AI require a well-functioning data ecosystem built on trust, data availability and infrastructure31.	E1
28	6	The General Data Protection Regulation (GDPR) [32] is the anchor of trust in the single market for data. It has established a new global standard with a strong focus on the rights of individuals, reflecting European values, and is an important element of ensuring trust in AI. This trust is especially important when it comes to the processing of healthcare data for applications driven by AI. The Commission would like to encourage the European Data Protection Board to develop guidelines on the issue of the processing of personal data in the context of research.	E1, E3
29	7	The work will meet all necessary regulatory, security, and ethical requirements.	E1
30	7	To gain trust, which is necessary for societies to accept and use AI, the technology should be predictable, responsible, verifiable, respect fundamental rights and follow ethical rules.	E1, E4

31	8	Crucially, humans should understand how AI makes decisions.	E1
32	8	To anchor such principles more firmly in the development and use of AI, the Commission appointed an independent AI high-level expert group with the task of developing draft AI ethics guidelines. A first version will be published by the end of 2018 and the experts will present their final version of the guidelines to the Commission in March 2019 after wide consultation through the European AI Alliance41.	E1, E3
33	8	Further developments in AI also require a regulatory framework that is flexible enough to promote innovation while ensuring high levels of protection and safety.	E1, E3
34	8	The increasing potential and sensitivity of AI applications in many areas of the digital economy and society, such as autonomous mobility or avoiding power blackouts, means it is highly relevant to establish cybersecurity requirements for AI.	E1, E3
35	8	Europe can become a global leader in developing and using AI for good and promoting a human-centric approach and ethics-by-design principles.	E5
36	8	The ambition is then to bring Europe's ethical approach to the global stage. The Commission is opening up cooperation to all non-EU countries that are willing to share the same values.	E5
37	8	The Union will continue to stress that international law, including International Humanitarian Law and Human Rights Law, applies fully to all weapons systems, including autonomous weapons systems, and that States remain	E5

		responsible and accountable for their development and use in armed conflict.	
38	9	For Europe to become a leading player in AI, it needs to build on its strengths and support the development of an ethical, secure and cutting-edge AI made in Europe.	E5

Document 3: Ethics Guidelines for Trustworthy AI

No.	Page	Citation	Category
38	4	To support the implementation of this vision, the Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG), an independent group mandated with the drafting of two deliverables: (1) AI Ethics Guidelines and (2) Policy and Investment Recommendations.	E1, E3
39	4	To do this, AI systems8 need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately.	E1, E2
40	4	We also want producers of AI systems to get a competitive advantage by embedding Trustworthy AI in their products and services. This entails seeking to maximise the benefits of AI systems while at the same time preventing and minimising their risks.	E1, E3
41	4	In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies and sustainable development. We therefore identify Trustworthy AI as our foundational	E1, E3

		ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place.	
42	4	It is through Trustworthy AI that we, as European citizens, will seek to reap its benefits in a way that is aligned with our foundational values of respect for human rights, democracy and the rule of law.	E1
43	4, 5	Trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems. Without AI systems – and the human beings behind them – being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered, preventing the realisation of the potentially vast social and economic benefits that they can bring. To help Europe realize those benefits, our vision is to ensure and scale Trustworthy AI.	E1, E3
44	5	Striving towards Trustworthy AI hence concerns not only the trustworthiness of the AI system itself but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle.	E1
45	5	Trustworthy AI has three components, which should be met throughout the system's entire life cycle: 1. it should be lawful, complying with all applicable laws and regulations; 2. it should be ethical, ensuring adherence to ethical principles and values; and 3. it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.	E1

46	5	Each of these three components is necessary but not sufficient in itself to achieve Trustworthy AI [10]. Ideally, all three work in harmony and overlap in their operation. In practice, however, there may be tensions between these elements (e.g. at times the scope and content of existing law might be out of step with ethical norms). It is our individual and collective responsibility as a society to work towards ensuring that all three components help to secure Trustworthy AI.	E1, E3
47	5	These Guidelines are intended to foster responsible and sustainable AI innovation in Europe. They seek to make ethics a core pillar for developing a unique approach to AI, one that aims to benefit, empower and protect both individual human flourishing and the common good of society.	E1, E3
48	6	These Guidelines articulate a framework for achieving Trustworthy AI based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law.	E1
49	6, 7	Laws are not always up to speed with technological developments, can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms.	E1
50	7	Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should perform in a safe, secure and reliable manner, and	E1

		safeguards should be foreseen to prevent any unintended adverse impacts. It is therefore important to ensure that AI systems are robust.	
51	9	AI ethics is a sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment, and use of AI. Its central concern is to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life or human autonomy and freedom necessary for a democratic society.	E1, E4
52	9	Ethical reflection on AI technology can serve multiple purposes. First, it can stimulate reflection on the need to protect individuals and groups at the most basic level. Second, it can stimulate new kinds of innovations that seek to foster ethical values, such as those helping to achieve the UN Sustainable Development Goals [13], which are firmly embedded in the forthcoming EU Agenda 2030.	E1
53	9	Trustworthy AI can improve individual flourishing and collective wellbeing by generating prosperity, value creation, and wealth maximization. It can contribute to achieving a fair society, by helping to increase citizens' health and well-being in ways that foster equality in the distribution of economic, social, and political opportunity.	E1
54	9	As with any powerful technology, the use of AI systems in our society raises several ethical challenges, for instance relating to their impact on people and society, decision-making capabilities and safety. If we are increasingly going to use the assistance of or delegate decisions to AI systems, we need to make sure these systems are fair in their impact on people's lives, that they are in line with values that	E1

		should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this.	
55	9	With this document, we intend to contribute to this effort by introducing the notion of Trustworthy AI, which we believe is the right way to build a future with AI.	E1
56	9	A domain-specific ethics code – however consistent, developed and fine-grained future versions of it may be – can never function as a substitute for ethical reasoning itself, which must always remain sensitive to contextual details that cannot be captured in general Guidelines. Beyond developing a set of rules, ensuring Trustworthy AI requires us to build and maintain an ethical culture and mind-set through public debate, education, and practical learning.	E1, E4
57	9	We believe in an approach to AI ethics based on the fundamental rights enshrined in the EU Treaties, [15] the EU Charter, and international human rights law. Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalized in the context of AI.	E1
58	9, 10	These rights are described in the EU Charter by reference to dignity, freedoms, equality, and solidarity, citizens' rights and justice. The common foundation that unites these rights can be understood as rooted in respect for human dignity – thereby reflecting what we describe as a "human-centric approach" in which the human being enjoys a unique and inalienable moral status of primacy in the civil, political, economic, and social fields.	E1

59	10	Understood as the rights of everyone, rooted in the inherent moral status of human beings, they also underpin the second component of Trustworthy AI (ethical AI), dealing with ethical norms that are not necessarily legally binding yet crucial to ensure trustworthiness.	E1
60	10	AI systems should hence be developed in a manner that respects, serves, and protects humans' physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs.	E1
61	10	In an AI context, freedom of the individual for instance requires mitigation of (in)direct illegitimate coercion, threats to mental autonomy and mental health, unjustified surveillance, deception, and unfair manipulation.	E1
62	11	Respect for democracy, justice, and the rule of law. All governmental power in constitutional democracies must be legally authorised and limited by law. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation, or democratic voting systems. AI systems must also embed a commitment to ensure that they do not operate in ways that undermine the foundational commitments upon which the rule of law is founded, mandatory laws and regulation, and to ensure due process and equality before the law.	E1
63	11	In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs (e.g. the data used to train AI systems should be as inclusive as possible, representing different population groups). This also requires adequate respect for potentially vulnerable	E1

		persons and groups, such as workers, women, persons with disabilities, ethnic minorities, children, consumers or others at risk of exclusion.	
64	11	AI systems offer substantial potential to improve the scale and efficiency of government in the provision of public goods and services to society. At the same time, citizens' rights could also be negatively impacted by AI systems and should be safeguarded.	E1
65	11	This section lists four ethical principles, rooted in fundamental rights, which must be respected in order to ensure that AI systems are developed, deployed and used in a trustworthy manner.	E1
66	12	These are the principles of: (i) Respect for human autonomy (ii) Prevention of harm (iii) Fairness (iv) Explicability	E1
67	12	The principle of respect for human autonomy: The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight [28] over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support	E1, E4

		humans in the working environment, and aim for the creation of meaningful work.	
68	12, 13	The principle of prevention of harm: AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. They must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment and use of AI systems. Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings.	E1, E4
69	13	The principle of fairness: The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity in terms of access to education, goods, services and	E1, E4

		technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives [31]. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them [32]. In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.	
70	13	The principle of explicability: Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity	E1, E3, E4

		of the consequences if that output is erroneous or otherwise inaccurate.	
71	14	The principles outlined in Chapter I must be translated into concrete requirements to achieve Trustworthy AI. These requirements are applicable to different stakeholders partaking in AI systems' life cycle: developers, deployers and end-users, as well as the broader society.	E1, E3
72	14	The below list of requirements is non-exhaustive. It includes systemic, individual and societal aspects: 1) Human agency and oversight, Including fundamental rights, human agency and human oversight. 2) Technical robustness and safety, Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility. 3) Privacy and data governance, Including respect for privacy, quality and integrity of data, and access to data. 4) Transparency, Including traceability, explainability and communication. 5) Diversity, non-discrimination and fairness, Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation. 6) Societal and environmental wellbeing, Including sustainability and environmental friendliness, social impact, society and democracy. 7) Accountability, Including auditability, minimisation and reporting of negative impact, trade-offs and redress.	E1, E2
73	17	Privacy and data governance: Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its	E1, E2

		relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.	
74	17	Privacy and data protection. AI systems must guarantee privacy and data protection throughout a system's entire lifecycle [41]. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations). Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.	E1, E2
75	17	Quality and integrity of data. The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set. In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems. Processes and data sets used must be tested and documented at each step such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere.	E1, E2, E3

76	17	Access to data. In any given organisation that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so.	E1, E3
77	18	Transparency: This requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.	E1
78	18	Traceability. The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.	E1, E3
79	18	Explainability. Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant	E1, E2, E3

		impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).	
80	18	Communication. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.	E1, E2, E3
81	18	Diversity, non-discrimination and fairness: In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of fairness.	E1

82	18	Avoidance of unfair bias. Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination [42] against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non-transparent market [43]. Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner. Moreover, hiring from diverse backgrounds, cultures and disciplines can ensure diversity of opinions and should be encouraged.	E1, E2, E3
83	18, 19	Accessibility and universal design. Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design44 principles addressing the widest possible range of users, following relevant accessibility standards [45]. This	E1, E3

		will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.	
84	19	Stakeholder Participation. In order to develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.	E1, E3
85	19	Societal and environmental well-being: In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations.	E1
86	19	Sustainable and environmentally friendly AI. AI systems promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly way possible. The system's development, deployment and use process, as well as its entire supply chain, should be assessed in this regard, e.g. via a critical examination of the resource usage and energy	E1, E3

		consumption during training, opting for less harmful choices. Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.	
87	19	Social impact. Ubiquitous exposure to social AI systems [47] in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or impact our social relationships and attachment. While AI systems can be used to enhance social skills, [48] they can equally contribute to their deterioration. This could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered.	E1, E2
88	19	Society and Democracy. Beyond assessing the impact of an AI system's development, deployment and use on individuals, this impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts.	E1, E2
89	19	Accountability: The requirement of accountability complements the above requirements, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.	E1

90	19, 20	Auditability. Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.	E1, E3
91	20	Minimisation and reporting of negative impacts. Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system. The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact. These assessments must be proportionate to the risk that the AI systems pose	E1, E2, E3

92	20	Trade-offs. When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights. In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form. Any decision about which trade-off to make should be reasoned and properly documented. The decision-maker must be accountable for the manner in which the appropriate trade-off is being made, and should continually review the appropriateness of the resulting decision to ensure that necessary changes can be made to the system where needed	E1, E2, E3
93	20	Redress. When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress [50]. Knowing that redress is possible when things go wrong is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.	E1, E3
94	21	Methods to ensure values-by-design provide precise and explicit links between the abstract principles which the system is required to respect and the specific implementation decisions. The idea that compliance with norms can be implemented into the design of the AI system is key to this method. Companies are responsible for identifying the impact of their AI systems from the very	E1

		start, as well as the norms their AI system ought to comply with to avert negative impacts.	
95	35	However, we are equally concerned to ensure that the risks and other adverse impacts with which these technologies are associated are properly and proportionately handled.	E2, E1
96	35	In this context, it is important to build AI systems that are worthy of trust, since human beings will only be able to confidently and fully reap its benefits when the technology, including the processes and people behind the technology, are trustworthy.	E1
97	35	Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations, (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since to ensure that, even with good intentions, AI systems do not cause any unintentional harm. Each component is necessary but not sufficient to achieve Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. Where tensions arise, we should endeavour to align them.	E1
98	37	Ethical AI: In this document, ethical AI is used to indicate the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values. It is the second of the three core elements necessary for achieving Trustworthy AI.	E1, E4

99	37	Human-Centric AI: The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoy a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come.	E1, E4
100	37	Robust AI: Robustness of an AI system encompasses both its technical robustness (appropriate in a given context, such as the application domain or life cycle phase) and as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. Robustness is the third of the three components necessary for achieving Trustworthy AI.	E1, E4
101	4	This is the path that we believe Europe should follow to become the home and leader of cutting-edge and ethical technology.	E5
102	5	We believe that this will enable Europe to position itself as a global leader in cutting-edge AI worthy of our individual and collective trust	E5

103	5	Just as the use of AI systems does not stop at national borders, neither does their impact. Global solutions are therefore required for the global opportunities and challenges that AI systems bring forth. We therefore encourage all stakeholders to work towards a global framework for Trustworthy AI, building international consensus while promoting and upholding our fundamental rights-based approach.	E5
104	35	The current document forms part of a vision that promotes Trustworthy AI which we believe should be the foundation upon which Europe can build leadership in innovative, cutting-edge AI systems.	E5

Document 4: Policy and Investment Recommendations for Trustworthy AI

No.	Page	Citation	Category
105	6	In our first deliverable, the Ethics Guidelines for Trustworthy AI [3] published on 8 April 2019 (Ethics Guidelines), we stated that AI systems need to be human-centric, with the goal of improving individual and societal well-being, and worthy of our trust. In order to be deemed trustworthy, we put forward that AI systems – including all actors and processes involved therein – should be lawful, ethical and robust. Those Guidelines therefore constituted a first important step in identifying the type of AI that we want and do not want for Europe, but that is not enough to ensure that Europe can also materialise the beneficial impact that Trustworthy AI can bring.	E1

106	6	Taking the next step, this document contains our proposed Policy and Investment Recommendations for Trustworthy AI, addressed to EU institutions and Member States.	E3
107	6	Building on our first deliverable, we put forward 33 recommendations that can guide Trustworthy AI towards sustainability, growth and competitiveness, as well as inclusion – while empowering, benefiting and protecting human beings.	E1, E3
108	6	This may necessitate specific and targeted governance measures that provide appropriate safeguards to protect individuals and society. In this report, we make recommendations to position Europe so that it can maximise the extent to which it can benefit from the opportunities presented by AI, while simultaneously ensuring that these benefits are felt throughout the entire European society, and that any risks are prevented or minimised.	E2, E3
109	8	As already stated in our Ethics Guidelines, in building a future with AI, our point of departure is human-centricity. By placing the human at the centre of our thinking, we underscore the fact that AI is not an end in itself, but a means to enhance human well-being and freedom. All policy recommendations that we put forward in this document have this as their direct or indirect goal. Human-centricity, however, not only implies attention to individuals, but also to the well-being of society at large and the environment that humans live in. Europe should champion the use of AI towards sustainable development in line with the Agenda 2030.	E1, E5

110	10	carries risks for humans and societies, which need to be identified and addressed. Hence, we need to foster AI solutions that can empower human beings, and monitor the impacts they create, ensuring that this happens in a way that protects our rights and values. It is therefore essential that individuals gain awareness, knowledge and understanding of the capabilities, challenges and limitations of AI systems, and of their rights related thereto.	E2, E3
111	10	Encourage Member States to increase digital literacy through courses (e.g. MOOCs) across Europe providing elementary AI training. This includes fostering the understanding of AI systems more generally (including a basic understanding of machine learning and reasoning), but also raising awareness of data protection rights, an understanding of how (personal) data can be used, the implications of digital tracking, and the importance of issues such as fairness, explainability, transparency, robustness of AI systems, and knowledge of these topics. Efforts need to be made to ensure that such courses are accessible to all, taking due account of the digital divide and paying particular attention to the lower skilled and disadvantaged.	E3
112	11	Institutionalise a dialogue between policy-makers, developers and users of AI technology, for instance through the European AI Alliance, on the ethical and legal limits of AI and examine how the policy and regulatory framework needs to be further developed in order to guarantee legal certainty and foster beneficial innovation while ensuring due respect for human rights, democracy and the rule of law.	E3
113	11	However, if not applied in a trustworthy manner, AI systems could cause adverse impacts to individuals, society and the	E2, E3

		environment, such as unjust discrimination or bias, privacy infringement, social or economic exclusion or environmental decline. Adequate protection should be put in place to counter such impacts.	
114	11	Refrain from disproportionate and mass surveillance of individuals. While there may be a strong temptation for governments to "secure society" by building a pervasive surveillance system based on AI systems, this would be extremely dangerous if pushed to extreme levels. Governments should commit not to engage in mass surveillance of individuals and to deploy and procure only Trustworthy AI systems, designed to be respectful of the law and fundamental rights, aligned with ethical principles and socio-technically robust.	E2, E3
115	12	Introduce a mandatory self-identification of AI systems. In situations where an interaction takes place between a human and an AI system, and whenever there is a reasonable likelihood that end users could be led to believe that they are interacting with a human, deployers of AI systems should be attributed a general responsibility to disclose that in reality the system is non-human. This goes hand-in-hand with ensuring the transparency of AI systems.	E3
116	13	Introduce a duty of care for developers of consumer-oriented AI systems to ensure that these can be used by all intended users, fostering a universal design approach, and do not lead to the exclusion of users with disabilities, particularly when used in public services.	E1, E3
117	14	AI should be developed with due regard to all grounds that are protected from discrimination in EU law, which also	E1

		includes – as well as some of the grounds listed above – the prohibition of discrimination on the ground of sex.	
118	15	Foster the availability of legal and technical support to implement Trustworthy AI solutions that comply with the Ethics Guidelines.	E3
119	16	Such innovation should be incentivised, for instance by establishing competitions, creating recognised standards and encourage open access on FRAND terms (fair, reasonable and non-discriminatory) to facilitate technology transfer.	E3
120	16	In B2C segments, such competitions can also be steered towards applications ensuring a universal design approach and accessibility, and the development of AI products and services for creating social good.	E3
121	17	Europe has a strong public sector that can play a significant role when it comes to the uptake and scaling of Trustworthy AI and establishing a Single Market for Trustworthy AI in Europe.	E5
122	18	This should not lead to a lower quality of human relationships within public services or a reduction of such services; the very purpose of the contribution of AI systems in the public sector is to be human-centric, and lies in the facilitation of the tasks of civil servants to ensure better services to individuals.	E1, E3
123	18	For instance, the development and deployment of those systems should occur in a transparent and accountable manner, to ensure that they operate in ways that are consistent with the principles of good administration, respect for fundamental rights, democracy and the rule of	E1, E3

		law. More generally, governments have the crucial task to safeguard individuals' fundamental rights, to protect them from harmful uses of AI, and to protect the integrity of public institutions.	
124	18	Consider adopting a proactive model for the delivery of public services for particular contexts and services in which they might enhance the effectiveness and quality of public services whilst ensuring due respect for fundamental rights and the rule of law.	E3
125	18	Where an AI-based service does not run properly or when an individual so requests, he or she should be able to interact with a human interlocutor, when there is a significant impact on the individual.	E3
126	19	Public services should invest in conversational user interfaces that can meet the needs of individuals 24/7, serving them in a more agile, accessible and faster way, from a single point of contact. This could for instance be done through the use of chatbots or natural language interfaces with multilingual support, that can help individuals by redirecting them to the information or service that they seek, and that could also simplify the filling in of forms in a conversational manner. Feedback mechanisms that allow users to share their comments on the interfaces and thus help improving their AI models should be developed. Moreover, it must be ensured that such AI-enabled services are trustworthy, i.e. legal, ethical and robust.	E3
127	19	Develop tools to ensure that public services can be deployed for all, and in a manner that safeguards individuals' fundamental rights, democracy and the rule of law.	E1, E3

128	19	In case it concerns personal data, it should be ensured this happens in a manner that complies with privacy, data protection rules and other fundamental rights.	E1
129	19	Create European large annotated public non-personal databases for high quality AI that are reliable and trustworthy.	E3
130	20	Introduce clear eligibility and selection criteria that in the procurement rules and processes of EU institutions, agencies and Member States that require AI systems to be trustworthy (lawful, ethical and robust), ensuring that they effectively protect people's personal data, privacy and autonomy. The Ethics Guidelines' assessment list can provide a helpful means to operationalise such requirement.	E3
131	20	Methods should be created to validate whether the government's decisions that rely on data-driven systems were biased against individuals compared to other similar decisions, given that access to one's own personal data is not enough to ensure the analysis of fair and just decisions that are in accordance with legal standards.	E2, E3
132	20	Make available to any individual who is subject to an AI-informed governmental decision that produces legal effects or similarly significantly affects that individual, information on the logic of the algorithms and how data is used to inform such decisions, enabling the affected individual to understand, evaluate and potentially challenge the decision.	E3
133	20	Fund and facilitate the development of AI tools that can assist in detecting biases and undue prejudice in governmental decision-making.	E3

134	20	Ban AI-enabled mass scale scoring of individuals as defined in our Ethics Guidelines, and set very clear and strict rules for surveillance for national security purposes and other purposes claimed to be in the public or national interest in line with EU regulation and case law. Develop trustworthy ways to do this where legal, necessary and proportionate, and ensure that this is not used in ways to suppress or undermine (political) opposition or democratic processes.	E2, E3
135	21	In particular, research and innovation on AI that address complementarity between AI systems and humans, that foster Trustworthy AI solutions and that address societal challenges should be promoted.	E1, E3
136	21	The roadmap should in particular foster research that can help ensuring AI solutions that meet the Trustworthy AI principles and requirements, enabling for instance requirements such as human oversight, privacy-by-design, robustness, non-discrimination and transparency (including the traceability and explainability of AI systems).	E1, E3
137	26	Europe takes pride in its sound regulatory environment that enables and stimulates competition and innovation while safeguarding fundamental rights and protection from unacceptable risk or harm. Yet, the new challenges raised by AI require reflection on an appropriate governance framework and a review of the adequacy of the current regulatory regime, pursuant to a comprehensive mapping of relevant EU regulations and potential legal gaps to both maximise AI's benefits and prevent and minimise its risks. Such a review should generally be based on a risk-based approach to AI policy-making, and take into account both individual and societal risks. For unacceptable risks, the	E2, E3

		revision of existing rules or the introduction of new regulation should be considered.	
138	28	A fundamental rights-based personal data infrastructure as put forward in the GDPR should be fostered and its enforcement should be ensured.	E1
139	29	Develop mechanisms for the protection of personal data, and individuals to control and be empowered by their data, thereby addressing some aspects of the requirements of trustworthy AI. Tools should be developed to provide a technological implementation of the GDPR and develop privacy preserving/privacy by design technical methods to explain criteria, causality in personal data processing of AI systems (such as federated machine learning).	E1, E3
140	29	Consider the introduction of a data access regime on FRAND terms, namely fair, reasonable, and non-discriminatory.	E3
141	31	At the same time, the future workforce will have to be equipped with a new – human centric – set of skills that empowers them on a cognitive and a socio-cultural level face the challenges ahead.	E1
142	37	Ensuring Trustworthy AI necessitates an appropriate governance and regulatory framework. By appropriate, we mean a framework that promotes socially valuable AI development and deployment, ensures and respects fundamental rights, the rule of law and democracy, while safeguarding individuals and society from unacceptable harm. On 8 April 2019, we published our Ethics Guidelines that set out three components for Trustworthy AI: (1) lawful AI, (2) ethical AI and (3) robust AI. The Ethics Guidelines only deal with the two latter components, yet the first is equally crucial. Many of the principles set out in the	E1, E2, E4

		Guidelines reflect existing EU law. This section complements the Guidelines by providing guidance on appropriate governance and regulatory approaches beyond voluntary guidance.	
143	37;38	Adopt a risk-based approach to regulation. The character, intensity and timing of regulatory intervention should be a function of the type of risk created by an AI system. In line with an approach based on the proportionality and precautionary principle, various risk classes should be distinguished as not all risks are equal.52 The higher the impact and/or probability of an AI-created risk, the stronger the appropriate regulatory response should be. 'Risk' for this purpose is broadly defined to encompass adverse impacts of all kinds, both individual and societal.53	E3, E2
144	38	For specific AI applications that generate "unacceptable" risks or pose threats of harm that are substantial, a precautionary principle-based approach should be adopted instead.54 Regulatory authorities should adopt precautionary measures when scientific evidence about an environmental, human health hazard or other serious societal threat (such as threats to the democratic process), and the stakes are high. Questions about the kinds of risks deemed unacceptable must be deliberated and decided upon by the community at large through open, transparent and accountable deliberation, taking into account the EU's legal framework and obligations under the Charter of Fundamental Rights.	E1, E2
145	38	Give due consideration to the level of autonomy in AI-based decision-making (e.g. is it an information source only, a support function, or a fully autonomous system without	E3

		human involvement) and the autonomy in learning when developing and updating policy measures for AI systems.	
146	38	Foster a principle-based approach to regulation. Unnecessarily prescriptive regulation should be avoided. In contexts characterised by rapid technological change, it is often preferable to adopt a principled-based approach, as well as outcome-based policies, subject to appropriate monitoring and enforcement. The European Commission should ground its policy measures on AI in EU values, as discussed and presented in our Ethics Guidelines, and should translate our aspirational goal of Trustworthy AI into a concrete set of indicators that can be used for monitoring the convergence of the European market towards the desired policy goals.	E1, E3
147	38	Consider the adoption of a segment-specific methodology when further developing the regulatory framework for AI. Both the necessary measures to protect individuals against adverse effects and the market environment of AI products and services developed and deployed in the B2C, B2B and P2C contexts differ from each other and merit a tailored approach.	E3
148	39	For civil liability55 and accountability rules: in the context of laws in areas significantly affecting individuals, consider whether for safety-critical and fundamental rights-critical applications it is necessary or desirable to introduce traceability and reporting requirements for AI applications to facilitate their auditability, ex-ante external oversight before AI systems can be deployed, systematic monitoring and oversight by competent authorities on an ongoing basis, and the obligation for meaningful human intervention and	E3, E2

		oversight when using AI decision in specific sectors (e.g. a human doctor to check a medical treatment decision). Finally, civil liability rules must be able to ensure adequate compensation in case of harm and/or rights violations (either through strict or tort liability), and may need to be complemented with mandatory insurance provisions.	
149	39	For criminal law provisions: consider the need to ensure that criminal responsibility and liability can be attributed in line with the fundamental principles of criminal law.	E3
150	39	For consumer protection rules: consider the extent to which existing laws have the capacity to safeguard against illegal, unfair, deceptive, exploitative and manipulative practices made possible by AI applications (for instance in the context of chatbots, include misleading individuals on the objective, purpose and capacity of an AI system) and whether a mandatory consumer protection impact assessment is necessary or desirable.	E3, E2
151	39	For data protection rules: consider whether existing laws allow sufficient access to public data and data for legitimate research purposes whilst preserving privacy and personal data protection, the appropriate scope of intellectual property rights protection, and whether the GDPR mandated transparency and explainability offers sufficient protection in light of the limitation of its scope to the processing of personal data and the fact that automated decision-making processes can also significantly affect individuals when the system is not fully automated or based on non-personal data.	E3, E2
152	39	For non-discrimination provisions: consider the extent to which laws prohibiting unlawful discrimination require the	E3, E2

		explicitation of obligations upon AI developers to verify the absence of unjust bias in AI systems' decisions, and the adequacy of enforcement mechanisms against discriminatory outcomes.	
153	39	For cyber-security rules: consider the extent to which the current cybersecurity regime provides sufficient protection against cybersecurity risks posed by AI systems.	E3, E2
154	39	For competition rules: consider the volume of data or incumbency data advantages – the building block of many AI systems – in the assessment of market power for the purposes of applying rules on anti-competitive behaviour, abuse of dominance or (algorithmic) collusion, and when evaluating mergers.	E3, E2
155	40	Examine the need for new regulation to address the critical concerns listed in our Ethics Guidelines for Trustworthy AI. More generally, it should continuously be evaluated whether AI systems generate risks that are not adequately addressed by existing legislation. In particular, individuals should not be subject to unjustified personal, physical or mental tracking or identification, profiling and nudging through AI powered methods of biometric recognition such as: emotional tracking, empathic media, DNA, iris and behavioural identification, affect recognition, voice and facial recognition and the recognition of micro-expressions. Exceptional use of such technologies, such as for national security purposes, must be evidence based, necessary and proportionate, as well as respectful of fundamental rights.	E1, E2
156	40	Monitor and restrict the development of automated lethal weapons, considering not only actual weapons, but also cyber attack tools that can have lethal consequences if	E1, E2

		deployed. With respect to offensive LAWS [56], advocate to the Member States to actively participate in the ongoing international debate, involve internationally recognised, non-military funded scientists and academics, experts in artificial intelligence, and propose to international partners the adoption of a moratorium on the development of offensive LAWS.	
157	40	Monitor the development of personalised AI systems built on children's profiles and ensure their alignment with fundamental rights, democracy and the rule of law. Consider introducing a legal age at which children receive a "clean data slate" of any public or private storage of data related to them as children [57].	E1, E2
158	40	For AI systems deployed by the private sector58 that have the potential to have a significant impact on human lives, for example by interfering with an individual's fundamental rights at any stage of the AI system's life cycle [59] and for safety-critical applications, consider the need to introduce: a mandatory obligation to conduct a trustworthy AI assessment (including a fundamental rights impact assessment which also covers for example the rights of children, the rights of individuals in relation to the state, and the rights of persons with disabilities [60]) and stakeholder consultation including consultation with relevant authorities; traceability, auditability and ex-ante oversight requirements; and an obligation to ensure appropriate by default and by design procedures to enable effective and immediate redress in case of mistakes, harms and/or other rights infringement.	E3, E2

159	41	Institutionalise a dialogue on AI policy with affected stakeholders to define red lines and discuss AI applications that may risk generating unacceptable harms, including applications that should be prohibited and/or tightly regulated or in specific situations where the risk for people's rights and freedoms would be too high and the impact of this technology would be detrimental to individuals or society as a whole. This could for instance be done through the European AI Alliance.	E3, E2
160	41	Develop auditing mechanisms for AI systems. This should allow public enforcement authorities as well as independent third party auditors to identify potentially illegal outcomes or harmful consequences generated by AI systems, such as unfair bias or discrimination.	E3
161	41	Ensure that the use of AI systems that entail interaction with end users is by default accompanied by procedures to support users in accessing effective redress in case of infringement of their rights under applicable laws. These procedures should be accompanied by simple explanations and a user-friendly procedure, and should entail interaction with a human interlocutor whenever possible and chosen by the user. Access to justice and effective redress are key elements of building consumer trust and thus are an important part of Trustworthy AI.	E3, E2
162	41	Foster the availability of redress-by-design mechanisms. This entails establishing – from the design phase – mechanisms to ensure alternative systems and procedures with an adequate level of human oversight (human in the loop, on the loop or in command approach) to be able to effectively detect, audit, and rectify incorrect decisions taken	E3

		by a "perfectly" functioning system, for those situations where the AI system's decisions significantly affects individuals.	
163	41	In addition, we urge policy-makers to refrain from establishing legal personality for AI systems or robots. We believe this to be fundamentally inconsistent with the principle of human agency, accountability and responsibility, and to pose a significant moral hazard.	E1
164	46	Encourage the Commission to work with European financial institutions, such as the European Investment Bank, to develop investment guidelines that take into account the Ethics Guidelines, leading to sustainable business developments. This could take the form of a criterion in the social proofing of future financial investments such as InvestEU. The appraisal of the Ethics Guidelines by all stakeholders, and notably industry and other international organisations, indicates how technologies with human-centric values are critical to ensuring societal acceptance.	E3
165	47	Europe has set its overarching ambition on a human-centric approach to Artificial Intelligence. In our first deliverable, this concept was captured in the notion of Trustworthy AI, which we characterised in terms of three components – being lawful, ethical and robust – and in line with the core tenets of the European Union: fundamental rights, democracy and the rule of law. Our Ethics Guidelines for Trustworthy AI hence constituted a crucial first step in delineating the type of AI that we want and do not want for Europe.	E1, E4
166	47	Taking the next step, this document therefore presents a set of policy and investment recommendations on how Trustworthy AI can actually be developed, deployed,	E3, E4

		fostered and scaled in Europe, all the while maximising its benefits whilst minimising and preventing its risks.	
167	47	We recall that Trustworthy AI is not an end itself, but can be a means to enhance individual and societal well-being. This requires sustainability, in order to safeguard our societal and natural environment for generations to come. It requires growth and competitiveness, so as to grow the pie, secure employment opportunities and generate beneficial progress. And it requires inclusion, to allow everyone to benefit therefrom.	E1
168	47	Using Trustworthy AI to enhance our well-being implies important prerequisites, in particular securing individual and societal empowerment and protection. First, individuals need to be aware of and understand the capabilities, limitations and impacts of AI. Second, they must have the necessary education and skills to use the technology, to ensure that they can truly benefit therefrom as well as being prepared for a transformed working environment where AI systems will become ever more prevalent. And third, they need adequate safeguards from any adverse impact that AI might bring.	E1
169	49	Ensuring Trustworthy AI requires an appropriate governance and regulatory framework. We advocate a risk-based approach that is focused on proportionate yet effective action to safeguard AI that is lawful, ethical and robust, and fully aligned with fundamental rights. A comprehensive mapping of relevant EU laws should be undertaken so as to assess the extent to which these laws are still fit for purpose in an AI-driven world. In addition, new legal measures and governance mechanisms may need to be put in place to	E1, E2

		ensure adequate protection from adverse impacts as well as enabling proper enforcement and oversight, without stifling beneficial innovation.	
170	24	With the Ethics Guidelines published on 8 April 2019, Europe has taken a strong initiative to lead the global debate on the applied ethics of AI. Consideration should be given to support the development of a Centre of Excellence in Trustworthy AI to maintain Europe's intellectual leadership	E5
171	37	Europe takes pride in its sound regulatory environment that enables and stimulates AI development and deployment through fostering legal certainty and providing a distinct global competitiveness element, while at the same time safeguarding fundamental rights and protecting individuals and society from risk or harm, guided principally by the proportionality principle.	E5
172	48	It is uniquely placed to deliver and promote human-centric and Trustworthy AI services, leading by example, while ensuring a strong protection of fundamental rights.	E5

Document 5: White Paper: On Artificial Intelligence – A European approach to excellence and trust

No.	Page	Citation	Category
173	1	At the same time, Artificial Intelligence (AI) entails a number of potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes.	E2
174	1	To address the opportunities and challenges of AI, the EU must act as one and define its own way, based on European values, to promote the development and deployment of AI.	E1, E3

175	1	Commission President Ursula von der Leyen announced in her political Guidelines2 a coordinated European approach on the human and ethical implications of AI as well as a reflection on the better use of big data for innovation.	E1
176	1	As digital technology becomes an ever more central part of every aspect of people's lives, people should be able to trust it. Trustworthiness is also a prerequisite for its uptake. This is a chance for Europe, given its strong attachment to values and the rule of law as well as its proven capacity to build safe, reliable and sophisticated products and services from aeronautics to energy, automotive and medical equipment.	E1, E5
177	1	Given the major impact that AI can have on our society and the need to build trust, it is vital that European AI is grounded in our values and fundamental rights such as human dignity and privacy protection.	E1
178	1	This White Paper presents policy options to enable a trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens.	E3, E1
179	1	The key elements of a future regulatory framework for AI in Europe that will create a unique 'ecosystem of trust'. To do so, it must ensure compliance with EU rules, including the rules protecting fundamental rights and consumers' rights, in particular for AI systems operated in the EU that pose a high risk [7]. Building an ecosystem of trust is a policy objective in itself, and should give citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI. The Commission strongly supports a human-centric approach based on the Communication on Building Trust in Human-Centric AI8 and will also take into account the input obtained	E3, E1

		during the piloting phase of the Ethics Guidelines prepared by the High-Level Expert Group on AI.	
180	9	As with any new technology, the use of AI brings both opportunities and risks. Citizens fear being left powerless in defending their rights and safety when facing the information asymmetries of algorithmic decision-making, and companies are concerned by legal uncertainty. While AI can help protect citizens' security and enable them to enjoy their fundamental rights, citizens also worry that AI can have unintended effects or even be used for malicious purposes. These concerns need to be addressed. Moreover, in addition to a lack of investment and skills, lack of trust is a main factor holding back a broader uptake of AI.	E2, E1
181	9	The Commission published a Communication31 welcoming the seven key requirements identified in the Guidelines of the High-Level Expert Group: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental wellbeing, and Accountability.	E1
182	9	A key result of the feedback process is that while a number of the requirements are already reflected in existing legal or regulatory regimes, those regarding transparency, traceability and human oversight are not specifically covered under current legislation in many economic sectors.	E1
183	9;10	On top of this set of non-binding Guidelines of the High-Level Expert Group, and in line with the President's political guidelines, a clear European regulatory framework would build trust among consumers and businesses in AI, and	E3, E1

		therefore speed up the uptake of the technology. Such a regulatory framework should be consistent with other actions to promote Europe's innovation capacity and competitiveness in this field. In addition, it must ensure socially, environmentally and economically optimal outcomes and compliance with EU legislation, principles and values. This is particularly relevant in areas where citizens' rights may be most directly affected, for example in the case of AI applications for law enforcement and the judiciary.	
184	10	Developers and deployers of AI are already subject to European legislation on fundamental rights (e.g. data protection, privacy, non-discrimination), consumer protection, and product safety and liability rules. Consumers expect the same level of safety and respect of their rights whether or not a product or a system relies on AI. However, some specific features of AI (e.g. opacity) can make the application and enforcement of this legislation more difficult. For this reason, there is a need to examine whether current legislation is able to address the risks of AI and can be effectively enforced, whether adaptations of the legislation are needed, or whether new legislation is needed.	E1
185	10	Given how fast AI is evolving, the regulatory framework must leave room to cater for further developments. Any changes should be limited to clearly identified problems for which feasible solutions exist.	E3
186	10	This harm might be both material (safety and health of individuals, including loss of life, damage to property) and immaterial (loss of privacy, limitations to the right of freedom of expression, human dignity, discrimination for	E2, E1

		instance in access to employment), and can relate to a wide variety of risks. A regulatory framework should concentrate on how to minimise the various risks of potential harm, in particular the most significant ones.	
187	10	The main risks related to the use of AI concern the application of rules designed to protect fundamental rights (including personal data and privacy protection and non-discrimination), as well as safety [32] and liability-related issues.	E2, E1
188	10	The use of AI can affect the values on which the EU is founded and lead to breaches of fundamental rights [33], including the rights to freedom of expression, freedom of assembly, human dignity, non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation, as applicable in certain domains, protection of personal data and private life, [34] or the right to an effective judicial remedy and a fair trial, as well as consumer protection. These risks might result from flaws in the overall design of AI systems (including as regards human oversight) or from the use of data without correcting possible bias (e.g. the system is trained using only or mainly data from men leading to suboptimal results in relation to women).	E2, E1
189	12	The specific characteristics of many AI technologies, including opacity ('black box-effect'), complexity, unpredictability and partially autonomous behaviour, may make it hard to verify compliance with, and may hamper the effective enforcement of, rules of existing EU law meant to protect fundamental rights.	E2, E1

190	13	Persons having suffered harm may not have effective access to the evidence that is necessary to build a case in court, for instance, and may have less effective redress possibilities compared to situations where the damage is caused by traditional technologies. These risks will increase as the use of AI becomes more widespread.	E2
191	13	An extensive body of existing EU product safety and liability legislation38, including sector-specific rules, further complemented by national legislation, is relevant and potentially applicable to a number of emerging AI applications.	E1
192	13	While the EU legislation remains in principle fully applicable irrespective of the involvement of AI, it is important to assess whether it can be enforced adequately to address the risks that AI systems create, or whether adjustments are needed to specific legal instruments.	E3
193	14	The Commission is of the opinion that the legislative framework could be improved to address the following risks and situations:	E2
194	14	Effective application and enforcement of existing EU and national legislation: the key characteristics of AI create challenges for ensuring the proper application and enforcement of EU and national legislation. The lack of transparency (opaqueness of AI) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights, attribute liability and meet the conditions to claim compensation. Therefore, in order to ensure an effective application and enforcement, it may be necessary to adjust or clarify existing legislation in	E2, E3

		certain areas, for example on liability as further detailed in the Report, which accompanies this White Paper.	
195	14	Limitations of scope of existing EU legislation: an essential focus of EU product safety legislation is on the placing of products on the market. While in EU product safety legislation software, when is part of the final product, must comply with the relevant product safety rules, it is an open question whether stand-alone software is covered by EU product safety legislation, outside some sectors with explicit rules45. General EU safety legislation currently in force applies to products and not to services, and therefore in principle not to services based on AI technology either (e.g. health services, financial services, transport services).	E2, E1
196	14	Changing functionality of AI systems: the integration of software, including AI, into products can modify the functioning of such products and systems during their lifecycle. This is particularly true for systems that require frequent software updates or which rely on machine learning. These features can give rise to new risks that were not present when the system was placed on the market. These risks are not adequately addressed in the existing legislation which predominantly focuses on safety risks present at the time of placing on the market.	E2
197	14	Uncertainty as regards the allocation of responsibilities between different economic operators in the supply chain: in general, EU legislation on product safety allocates the responsibility to the producer of the product placed on the market, including all components e.g. AI systems. But the rules can for example become unclear if AI is added after the product is placed on the market by a party that is not the	E2

		producer. In addition, EU product liability legislation provides for liability of producers and leaves national liability rules to govern liability of others in the supply chain.	
198	14, 15	Changes to the concept of safety: the use of AI in products and services can give rise to risks that EU legislation currently does not explicitly address. These risks may be linked to cyber threats, personal security risks (linked for example to new applications of AI such as to home appliances), risks that result from loss of connectivity, etc. These risks may be present at the time of placing products on the market or arise as a result of software updates or self-learning when the product is being used. The EU should make full use of the tools at its disposal to enhance its evidence base on potential risks linked to AI applications, including using the experience of the EU Cybersecurity Agency (ENISA) for assessing the AI threat landscape.	E2, E3
199	15	The autonomous behaviour of certain AI systems during its life cycle may entail important product changes having an impact on safety, which may require a new risk assessment. In addition, human oversight from the product design and throughout the lifecycle of the AI products and systems may be needed as a safeguard.	E2, E3
200	15	Explicit obligations for producers could be considered also in respect of mental safety risks of users when appropriate (ex. collaboration with humanoid robots).	E3
201	15	Union product safety legislation could provide for specific requirements addressing the risks to safety of faulty data at the design stage as well as mechanisms to ensure that quality	E3

		of data is maintained throughout the use of the AI products and systems.	
202	15	The opacity of systems based on algorithms could be addressed through transparency requirements.	E3
203	15	Existing rules may need to be adapted and clarified in the case of a stand-alone software placed as it is on the market or downloaded into a product after its placing on the market, when having an impact on safety.	E2, E3
204	15	Given the increasing complexity of supply chains as regards new technologies, provisions specifically requesting cooperation between the economic operators in the supply chain and the users could provide legal certainty.	E3
205	17	A risk-based approach is important to help ensure that the regulatory intervention is proportionate. However, it requires clear criteria to differentiate between the different AI applications, in particular in relation to the question whether or not they are 'high-risk' [49]. The determination of what is a high-risk AI application should be clear and easily understandable and applicable for all parties concerned. Nevertheless even if an AI application is not qualified as high-risk, it remains entirely subject to already existing EU-rules.	E1

206	17	More specifically, an AI application should be considered high-risk where it meets the following two cumulative criteria: First, the AI application is employed in a sector where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur. This first criterion ensures that the regulatory intervention is targeted on the areas where, generally speaking, risks are deemed most likely to occur. The sectors covered should be specifically and exhaustively listed in the new regulatory framework. For instance, healthcare; transport; energy and parts of the public sector. The list should be periodically reviewed and amended where necessary in function of relevant developments in practice; Second, the AI application in the sector in question is, in addition, used in such a manner that significant risks are likely to arise. This second criterion reflects the acknowledgment that not every use of AI in the selected sectors necessarily involves significant risks. For example, whilst healthcare generally may well be a relevant sector, a flaw in the appointment scheduling system in a hospital will normally not pose risks of such significance as to justify legislative intervention. The assessment of the level of risk of a given use could be based on the impact on the affected parties. For instance, uses of AI applications that produce legal or similarly significant effects for the rights of an individual or a company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities.	E4

207	18	Taking into account the guidelines of the High Level Expert Group and what has been set out in the foregoing, the requirements for high-risk AI applications could consist of the following key features, which are discussed in further detail in the subsections below: training data; data and record-keeping; information to be provided; robustness and accuracy; human oversight; specific requirements for certain particular AI applications, such as those used for purposes of remote biometric identification.	E3
208	18	It is more important than ever to promote, strengthen and defend the EU's values and rules, and in particular the rights that citizens derive from EU law. These efforts undoubtedly also extend to the high-risk AI applications marketed and used in the EU under consideration here.	E1, E5
209	19	Requirements aimed at providing reasonable assurances that the subsequent use of the products or services that the AI system enables is safe, in that it meets the standards set in the applicable EU safety rules (existing as well as possible complementary ones). For instance, requirements ensuring that AI systems are trained on data sets that are sufficiently broad and cover all relevant scenarios needed to avoid dangerous situations.	E3
210	19	Requirements to take reasonable measures aimed at ensuring that such subsequent use of AI systems does not lead to outcomes entailing prohibited discrimination. These requirements could entail in particular obligations to use data sets that are sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets;	E3, E1

211	19	Requirements aimed at ensuring that privacy and personal data are adequately protected during the use of AI-enabled products and services. For issues falling within their respective scope, the General Data Protection Regulation and the Law Enforcement Directive regulate these matters.	E1
212	20	Transparency is required also beyond the record-keeping requirements discussed in point c) above. In order to achieve the objectives pursued – in particular promoting the responsible use of AI, building trust and facilitating redress where needed – it is important that adequate information is provided in a proactive manner about the use of high-risk AI systems.	E1, E3
213	20	Ensuring clear information to be provided as to the AI system's capabilities and limitations, in particular the purpose for which the systems are intended, the conditions under which they can be expected to function as intended and the expected level of accuracy in achieving the specified purpose. This information is important especially for deployers of the systems, but it may also be relevant to competent authorities and affected parties.	E3
214	20	Separately, citizens should be clearly informed when they are interacting with an AI system and not a human being. Whilst EU data protection legislation already contain certain rules of this kind [54], additional requirements may be called for to achieve the abovementioned objectives. If so, unnecessary burdens should be avoided. Therefore, no such information needs to be provided, for instance, in situations where it is immediately obvious to citizens that they are interacting with AI systems. It is furthermore important that the information provided is objective, concise and easily	E1, E3

		understandable. The manner in which the information is to be provided should be tailored to the particular context.	
215	20	AI systems – and certainly high-risk AI applications – must be technically robust and accurate in order to be trustworthy. That means that such systems need to be developed in a responsible manner and with an ex-ante due and proper consideration of the risks that they may generate. Their development and functioning must be such to ensure that AI systems behave reliably as intended. All reasonable measures should be taken to minimise the risk of harm being caused.	E1
216	21	Human oversight helps ensuring that an AI system does not undermine human autonomy or cause other adverse effects. The objective of trustworthy, ethical and human-centric AI can only be achieved by ensuring an appropriate involvement by human beings in relation to high-risk AI applications.	E1
217	21	The gathering and use of biometric data [55] for remote identification [56] purposes, for instance through deployment of facial recognition in public places, carries specific risks for fundamental rights [57].	E2, E1
218	22	It follows that, in accordance with the current EU data protection rules and the Charter of Fundamental Rights, AI can only be used for remote biometric identification purposes where such use is duly justified, proportionate and subject to adequate safeguards.	E1
219	22	It is the Commission's view that, in a future regulatory framework, each obligation should be addressed to the actor(s) who is (are) best placed to address any potential risks.	E1, E3

220	23	In order to ensure that AI is trustworthy, secure and in respect of European values and rules, the applicable legal requirements need to be complied with in practice and be effectively enforced both by competent national and European authorities and by affected parties.	E1, E3
221	25	The European approach for AI aims to promote Europe's innovation capacity in the area of AI while supporting the development and uptake of ethical and trustworthy AI across the EU economy. AI should work for people and be a force for good in society.	E1, E5
222	2	Europe can combine its technological and industrial strengths with a high-quality digital infrastructure and a regulatory framework based on its fundamental values to become a global leader in innovation in the data economy and its applications as set out in the European data strategy [3].	E5
223	8	Europe is well positioned to exercise global leadership in building alliances around shared values and promoting the ethical use of AI. The EU's work on AI has already influenced international discussions. When developing its ethical guidelines, the High-Level Expert Group involved a number of non-EU organisations and several governmental observers. In parallel, the EU was closely involved in developing the OECD's ethical principles for AI [25]. The G20 subsequently endorsed these principles in its June 2019 Ministerial Statement on Trade and Digital Economy.	E5
224	8, 9	The Commission is convinced that international cooperation on AI matters must be based on an approach that promotes the respect of fundamental rights, including human dignity, pluralism, inclusion, non-discrimination and protection of	E5, E1

		privacy and personal data [26] and it will strive to export its values across the world [27].	

Document 6: Coordinated Plan on Artificial Intelligence 2021 Review

No.	Page	Citation	Category
225	2	The global leadership of Europe in adopting the latest technologies, seizing the benefits and promoting the development of human-centric, sustainable, secure, inclusive and trustworthy artificial intelligence (AI) depends on the ability of the European Union (EU) to accelerate, act and align AI policy priorities and investments [2]. This is the key message and a vision of this 2021 review of the Coordinated Plan.	E5
226	2	The 2021 review of the Coordinated Plan is the next step – it puts forward a concrete set of joint actions for the European Commission and Member States on how to create EU global leadership on trustworthy AI.	E5
227	3	In addition, the RRF provides an unprecedented opportunity to modernise and invest in AI to lead globally in the development and uptake of human-centric, trustworthy, secure and sustainable AI technologies [6].	E5
228	4	In order to accelerate, act and align to seize opportunities of AI technologies and to facilitate the European approach to AI, that is human-centric, trustworthy, secure, sustainable and inclusive AI, in full respect of our core European values, this review of the Coordinated Plan puts forward four key sets of proposals for the European Union and the Member States:	E1

229	20	have the aim that AI-related projects that receive R&I funding under the Horizon Europe adhere, as appropriate, to the 'ethics by design' principle, including for trustworthy AI.	E1
230	26	However, some uses of AI can also challenge rights protected by EU law and trigger new safety and security concerns [120], and affect labour markets. In the 2020 White Paper on AI121, the Commission put forward the European approach on AI that builds on an ecosystem of excellence and an ecosystem of trust for AI122.	E1, E2
231	29	support traineeships in digital areas, extending the possibility of participating to vocational education students and teaching staff, in addition to university students, with an increased focus on AI skills and with particular attention to the principle of non-discrimination and gender equality; and	E1
232	29	develop ethical guidelines on AI and data usage in teaching and learning for educators as well as the support of related research and innovation activities through Horizon Europe. This Action will build on the work of the High-Level Expert Group on AI on ethical guidelines [133]. The guidelines will be accompanied by a training programme for researchers and students on the ethical aspects of AI and include a target of 45 % of female participation in the training activities;	E1
233	31	Develop a policy framework to ensure trust in AI systems	E1
234	31	Trust is essential to facilitate the uptake of AI technologies. The European approach on AI, as proposed in the 2020 White Paper on AI, 'aims to promote Europe's innovation capacity in the area of AI while supporting the development and uptake of ethical and trustworthy AI across the EU	E1

		economy. AI should work for people and be a force for good in society' [140]. Given the major social and environmental impacts of AI technologies, a human-centric approach to their development and use, the protection of EU values and fundamental rights such as non-discrimination, privacy and data protection, and the sustainable and efficient use of resources are among the key principles that guide the European approach.	
235	31	Specifically, actions to facilitate trust have focused on issues relating to ethics, safety, fundamental rights, including the right not to be discriminated against, liability, the regulatory framework, innovation, competition [143], and intellectual property (IP).	E1
236	32	The main lessons learned are that the EU's approach should be human-centric, risk-based, proportionate and dynamic. One element of designing regulatory environments that are conducive to innovation, suggested by various stakeholders, is regulatory sandboxes. Regulatory sandboxes, in essence provide an experimentation facility for public regulation, and allow a more rapid evaluation of the impact of public intervention.	E1
237	32;33	The Commission will: Propose in 2021 legislative action on a horizontal framework for AI, focusing on issues of safety and the respect for fundamental rights specific to AI technologies. The proposed framework provides a definition of AI, it is risk-based (i.e. defines what a 'high risk' AI is) and lays down mandatory requirements for high-risk AI systems. It also proposes a governance mechanism that covers both ex ante conformity assessments and an ex post compliance and enforcement system. Outside of the high-	E1, E4

		risk category, all providers of AI systems are subject to the existing legislation and transparency requirements, and additionally could choose to subscribe to voluntary, non-binding, self-regulatory schemes, such as codes of conduct;	
238	33	The Commission will: propose in 2022 EU measures adapting the liability framework to the challenges of new technologies, including AI to ensure that victims who suffer damage to their life, health or property as a result of new technologies have access to the same compensation as victims of other technologies. This may include a revision of the Product Liability Directive [153], and a legislative proposal with regard to the liability for certain AI systems. Any new or amended provisions of existing legislation will take into account other existing EU legislation, as well as the proposed horizontal framework for AI;	E1
239	33	The Commission will: propose in 2021 and onwards as necessary revisions of existing sectoral safety legislation, including: targeted adaptations of the Machinery Directive154, the General Product Safety Directive, the Radio-Equipment Directive and the harmonised product legislation that follows the horizontal rules of the New Legislative Framework [155]. Any new or amended provisions of the existing legislation will take into account the existing EU health and safety at work legislation;	E1
240	34	Asserting Europe's global leadership and promoting the development of human-centric, sustainable, secure, inclusive and trustworthy AI will build further on the actions undertaken since the 2018 Coordinated Plan. In line with the Joint Communication on strengthening the EU's	E5

		contribution to rules-based multilateralism and as set out in the Commission 'Communication on 2030 Digital Compass: the European way for the Digital Decade', the international dimension is more essential than ever. The implications of new digital technologies such as AI transcend borders and need to be addressed globally	
241	34	The EU will promote ambitious global rules and standards, including strengthening cooperation with like-minded countries and the broader multi-stakeholder community and in a Team Europe spirit to support a human-centric and rules-based approach to AI. In order to be effective, the EU's approach will continue to be based on a proactive approach in various international bodies to build the strongest possible coalition of countries that share the desire for regulatory guardrails and democratic governance that benefit our societies. At the same time, the EU will reach out to other partners and seek common ground on an issue-by-issue basis to address the vast array of opportunities and challenges related to AI.	E5
242	35	The EU is a founding member of the new Global Partnership on AI (GPAI) launched in July 2020, with strong representation in the four working groups on: data governance, responsible AI (including a subgroup on pandemic response), the future of work; and commercialisation and innovation [161].	E5
243	35, 36	Dialogue with the United States on the development and roll out of trustworthy AI is ongoing. The Commission and the High Representative have jointly set out their ambitions for a new, forward-looking transatlantic agenda, including digital and other technology issues. The Commission is	E5

		notably proposing the setting up of an EU-US Trade and Technology council. Concretely the Commission will work towards an AI Agreement with the US [165]. There are several channels for discussion with US representatives (e.g. the EU-US Information Society Dialogue) [166] and various institutions/think tanks [167].	
244	36	The EU will step up its bilateral and multilateral efforts to support the establishment of a global level playing field for trustworthy and ethical use of AI, building notably on a strong transatlantic cooperation but also through a wider coalition of like-minded partners.	E5
245	36	continue to participate in, facilitate and support international, multilateral and bilateral discussions on trustworthy AI founded on an open value-based approach and promote the EU's approach to AI on the global stage, i.e. through regulatory cooperation, strategic communication and public diplomacy;	E5
246	36	foster the setting of global AI standards in close collaboration with international partners and continue to participate in the WIPO work on AI and IP rights; and	E5
247	36	continue their international outreach efforts on AI and ensure that Europe sends consistent messages on trustworthy AI to the world. Additionally, the Union will continue to contribute its expertise and dedicated financial means to anchor AI more firmly in diplomacy and in development policy with a particular focus on southern Mediterranean countries and Africa; and	E5

248	44	This development comes with a number of challenges. The changing labour landscape stresses the need to devise new working methods and to develop appropriate training in skills and competences for work alongside robots, and to understand their capabilities and limitations. Left unaddressed, these factors undermine trust in and acceptance of robotic technology. The Commission will continue to closely monitor the impacts on society, employment and labour conditions in the light of the development and uptake of AI technologies.	E1, E2
249	44	On the other hand, the specificity of robotics is linked to physical interaction with people and the environment. Robots will be increasingly autonomous and interacting with humans, be it co-working robots emerging from cages or robots providing services. This raises questions of safety: proximity to humans and interaction with them requires very high safety standards to prevent accidents and injuries. It also raises issues regarding ensuring accessibility and inclusiveness of persons with disabilities. Robots are also becoming more and more connected to each other and other types of devices and process more data, posing potential privacy and cybersecurity risks. All these considerations highlight the need to address testing, as planned in the future Testing and Experimentation Facilities, and to deal with issues such as certification and compliance with the regulatory framework, e.g. through regulatory sandboxes.	E1, E2
250	47	Through early adoption of AI, the public sector can be the first mover in adopting AI that is secure, trustworthy and sustainable [208].	E1

251	51	This also serves the objective that AI-enabled technologies fully comply with democratic values, the rule of law and fundamental rights and principles, including non-discrimination and data protection. These efforts will also contribute to the establishment of an ecosystem of trust.	E1
252	52	In order to ensure truly inclusive transport and mobility services, datasets used to train AI algorithms must be representative and balanced to avoid unintended results and potential discrimination of certain transport users.	E1
253	56	The objectives of the 2018 Coordinated Plan remain relevant and the overall direction set in the Coordinated Plan has proven to be the right one to contribute to Europe's ambition 'to become the world-leading region for developing and deploying cutting-edge, ethical and secure AI, (and) promoting a human-centric approach in the global context' [270].	E5

Document 7: Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts

No.	Page	Citation	Category
254	1	However, the same elements and techniques that power the socio-economic benefits of AI can also bring about new risks or negative consequences for individuals or the society.	E2
255	1	This proposal aims to implement the second objective for the development of an ecosystem of trust by proposing a legal framework for trustworthy AI. The proposal is based on EU values and fundamental rights and aims to give people and other users the confidence to embrace AI-based solutions, while encouraging businesses to develop them. AI should be	E1, E3

		a tool for people and be a force for good in society with the ultimate aim of increasing human well-being. Rules for AI available in the Union market or otherwise affecting people in the Union should therefore be human centric, so that people can trust that the technology is used in a way that is safe and compliant with the law, including the respect of fundamental rights.	
256	2	In 2017, the European Council called for a 'sense of urgency to address emerging trends' including 'issues such as artificial intelligence …, while at the same time ensuring a high level of data protection, digital rights and ethical standards' [5].	E1
257	2	The European Parliament has also undertaken a considerable amount of work in the area of AI. In October 2020, it adopted a number of resolutions related to AI, including on ethics [9], liability [10] and copyright [11].	E1
258	2	The EP Resolution on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies specifically recommends to the Commission to propose legislative action to harness the opportunities and benefits of AI, but also to ensure protection of ethical principles. The resolution includes a text of the legislative proposal for a regulation on ethical principles for the development, deployment and use of AI, robotics and related technologies	E1, E3
259	3	Against this political context, the Commission puts forward the proposed regulatory framework on Artificial Intelligence with the following specific objectives: ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values; ensure legal certainty to facilitate investment and innovation in AI;	E1, E3

		enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems; facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.	
260	3	To achieve those objectives, this proposal presents a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market. The proposal sets a robust and flexible legal framework.	E3
261	3	The proposal sets harmonised rules for the development, placement on the market and use of AI systems in the Union following a proportionate risk-based approach.	E3
262	3	Certain particularly harmful AI practices are prohibited as contravening Union values, while specific restrictions and safeguards are proposed in relation to certain uses of remote biometric identification systems for the purpose of law enforcement. The proposal lays down a solid risk methodology to define "high-risk" AI systems that pose significant risks to the health and safety or fundamental rights of persons. Those AI systems will have to comply with a set of horizontal mandatory requirements for trustworthy AI and follow conformity assessment procedures before those systems can be placed on the Union market. Predictable, proportionate and clear obligations are also placed on providers and users of those systems to ensure safety and respect of existing legislation protecting	E1, E3, E2

		fundamental rights throughout the whole AI systems' lifecycle. For some specific AI systems, only minimum transparency obligations are proposed, in particular when chatbots or 'deep fakes' are used.	
263	4	The horizontal nature of the proposal requires full consistency with existing Union legislation applicable to sectors where high-risk AI systems are already used or likely to be used in the near future.	E1
264	4	Consistency is also ensured with the EU Charter of Fundamental Rights and the existing secondary Union legislation on data protection, consumer protection, non-discrimination and gender equality. The proposal is without prejudice and complements the General Data Protection Regulation (Regulation (EU) 2016/679) and the Law Enforcement Directive (Directive (EU) 2016/680) with a set of harmonised rules applicable to the design, development and use of certain high-risk AI systems and restrictions on certain uses of remote biometric identification systems. Furthermore, the proposal complements existing Union law on non-discrimination with specific requirements that aim to minimise the risk of algorithmic discrimination, in particular in relation to the design and the quality of data sets used for the development of AI systems complemented with obligations for testing, risk management, documentation and human oversight throughout the AI systems' lifecycle.	E1, E4

265	7	The proposal builds on existing legal frameworks and is proportionate and necessary to achieve its objectives, since it follows a risk-based approach and imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety. For other, non-high-risk AI systems, only very limited transparency obligations are imposed, for example in terms of the provision of information to flag the use of an AI system when interacting with humans. For high-risk AI systems, the requirements of high quality data, documentation and traceability, transparency, human oversight, accuracy and robustness, are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks. Harmonised standards and supporting guidance and compliance tools will assist providers and users in complying with the requirements laid down by the proposal and minimise their costs.	E3, E2
266	7	The choice of a regulation as a legal instrument is justified by the need for a uniform application of the new rules, such as definition of AI, the prohibition of certain harmful AI-enabled practices and the classification of certain AI systems.	E4, E3
267	8	Stakeholders mostly requested a narrow, clear and precise definition for AI. Stakeholders also highlighted that besides the clarification of the term of AI, it is important to define 'risk', 'high-risk', 'low-risk', 'remote biometric identification' and 'harm'.	E4
268	8	In April 2019, the Commission supported23 the key requirements set out in the HLEG ethics guidelines for Trustworthy AI [24], which had been revised to take into account more than 500 submissions from stakeholders. The	E1

		key requirements reflect a widespread and common approach, as evidenced by a plethora of ethical codes and principles developed by many private and public organisations in Europe and beyond, that AI development and use should be guided by certain essential value-oriented principles.	
269	9	According to the Commission's established methodology, each policy option was evaluated against economic and societal impacts, with a particular focus on impacts on fundamental rights. The preferred option is option 3+, a regulatory framework for high-risk AI systems only, with the possibility for all providers of non-high-risk AI systems to follow a code of conduct. The requirements will concern data, documentation and traceability, provision of information and transparency, human oversight and robustness and accuracy and would be mandatory for high-risk AI systems.	E3
270	10	By requiring a restricted yet effective set of actions from AI developers and users, the preferred option limits the risks of violation of fundamental rights and safety of people and foster effective supervision and enforcement, by targeting the requirements only to systems where there is a high risk that such violations could occur.	E1, E3, E2
271	10	The preferred option will increase people's trust in AI, companies will gain in legal certainty, and Member States will see no reason to take unilateral action that could fragment the single market.	E1
272	10	This proposal lays down obligation that will apply to providers and users of high-risk AI systems.	E3

273	10	For companies using AI, it will promote trust among their customers. For national public administrations, it will promote public trust in the use of AI and strengthen enforcement mechanisms (by introducing a European coordination mechanism, providing for appropriate capacities, and facilitating audits of the AI systems with new requirements for documentation, traceability and transparency).	E1
274	11	With a set of requirements for trustworthy AI and proportionate obligations on all value chain participants, the proposal will enhance and promote the protection of the rights protected by the Charter: the right to human dignity (Article 1), respect for private life and protection of personal data (Articles 7 and 8), non-discrimination (Article 21) and equality between women and men (Article 23). It aims to prevent a chilling effect on the rights to freedom of expression (Article 11) and freedom of assembly (Article 12), to ensure protection of the right to an effective remedy and to a fair trial, the rights of defence and the presumption of innocence (Articles 47 and 48), as well as the general principle of good administration. Furthermore, as applicable in certain domains, the proposal will positively affect the rights of a number of special groups, such as the workers' rights to fair and just working conditions (Article 31), a high level of consumer protection (Article 28), the rights of the child (Article 24) and the integration of persons with disabilities (Article 26). The right to a high level of environmental protection and the improvement of the quality of the environment (Article 37) is also relevant, including in relation to the health and safety of people.	E1, E2

275	11	The obligations for ex ante testing, risk management and human oversight will also facilitate the respect of other fundamental rights by minimising the risk of erroneous or biased AI-assisted decisions in critical areas such as education and training, employment, important services, law enforcement and the judiciary. In case infringements of fundamental rights still happen, effective redress for affected persons will be made possible by ensuring transparency and traceability of the AI systems coupled with strong ex post controls.	E1, E2
276	11	This proposal imposes some restrictions on the freedom to conduct business (Article 16) and the freedom of art and science (Article 13) to ensure compliance with overriding reasons of public interest such as health, safety, consumer protection and the protection of other fundamental rights ('responsible innovation') when high-risk AI technology is developed and used. Those restrictions are proportionate and limited to the minimum necessary to prevent and mitigate serious safety risks and likely infringements of fundamental rights.	E1, E2
277	12;13	Title II establishes a list of prohibited AI. The regulation follows a risk-based approach, differentiating between uses of AI that create (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk. The list of prohibited practices in Title II comprises all those AI systems whose use is considered unacceptable as contravening Union values, for instance by violating fundamental rights. The prohibitions covers practices that have a significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities of specific	E1, E2, E4

		vulnerable groups such as children or persons with disabilities in order to materially distort their behaviour in a manner that is likely to cause them or another person psychological or physical harm. Other manipulative or exploitative practices affecting adults that might be facilitated by AI systems could be covered by the existing data protection, consumer protection and digital service legislation that guarantee that natural persons are properly informed and have free choice not to be subject to profiling or other practices that might affect their behaviour. The proposal also prohibits AI-based social scoring for general purposes done by public authorities. Finally, the use of 'real time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement is also prohibited unless certain limited exceptions apply.	
278	13	Title III contains specific rules for AI systems that create a high risk to the health and safety or fundamental rights of natural persons. In line with a risk-based approach, those high-risk AI systems are permitted on the European market subject to compliance with certain mandatory requirements and an ex-ante conformity assessment. The classification of an AI system as high-risk is based on the intended purpose of the AI system, in line with existing product safety legislation. Therefore, the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used.	E1, E2, E4
279	13	Chapter 2 sets out the legal requirements for high-risk AI systems in relation to data and data governance, documentation and recording keeping, transparency and	E3

		provision of information to users, human oversight, robustness, accuracy and security.	
280	14	As regards stand-alone high-risk AI systems that are referred to in Annex III, a new compliance and enforcement system will be established.	E3
281	14	A comprehensive ex-ante conformity assessment through internal checks, combined with a strong ex-post enforcement, could be an effective and reasonable solution for those systems, given the early phase of the regulatory intervention and the fact the AI sector is very innovative and expertise for auditing is only now being accumulated.	E3
282	14;15	Title IV concerns certain AI systems to take account of the specific risks of manipulation they pose. Transparency obligations will apply for systems that (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content ('deep fakes'). When persons interact with an AI system or their emotions or characteristics are recognised through automated means, people must be informed of that circumstance. If an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means, subject to exceptions for legitimate purposes (law enforcement, freedom of expression). This allows persons to make informed choices or step back from a given situation.	E2, E3
283	17	This Regulation pursues a number of overriding reasons of public interest, such as a high level of protection of health, safety and fundamental rights, and it ensures the free	E1

		movement of AI-based goods and services cross-border, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation.	
284	18	At the same time, depending on the circumstances regarding its specific application and use, artificial intelligence may generate risks and cause harm to public interests and rights that are protected by Union law. Such harm might be material or immaterial.	E2
285	18	A Union legal framework laying down harmonised rules on artificial intelligence is therefore needed to foster the development, use and uptake of artificial intelligence in the internal market that at the same time meets a high level of protection of public interests, such as health and safety and the protection of fundamental rights, as recognised and protected by Union law.	E1
286	20	In order to ensure a level playing field and an effective protection of rights and freedoms of individuals across the Union, the rules established by this Regulation should apply to providers of AI systems in a non-discriminatory manner, irrespective of whether they are established within the Union or in a third country, and to users of AI systems established within the Union.	E1
287	20	In order to ensure a consistent and high level of protection of public interests as regards health, safety and fundamental rights, common normative standards for all high-risk AI systems should be established. Those standards should be consistent with the Charter of fundamental rights of the European Union (the Charter) and should be non-	E1

		discriminatory and in line with the Union's international trade commitments.	
288	21	In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate. It is therefore necessary to prohibit certain artificial intelligence practices, to lay down requirements for high-risk AI systems and obligations for the relevant operators, and to lay down transparency obligations for certain AI systems.	E3
289	21	Aside from the many beneficial uses of artificial intelligence, that technology can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices. Such practices are particularly harmful and should be prohibited because they contradict Union values of respect for human dignity, freedom, equality, democracy and the rule of law and Union fundamental rights, including the right to non-discrimination, data protection and privacy and the rights of the child.	E1
290	21	The placing on the market, putting into service or use of certain AI systems intended to distort human behaviour, whereby physical or psychological harms are likely to occur, should be forbidden. Such AI systems deploy subliminal components individuals cannot perceive or exploit vulnerabilities of children and people due to their age, physical or mental incapacities. They do so with the intention to materially distort the behaviour of a person and in a manner that causes or is likely to cause harm to that or another person. The intention may not be presumed if the	E2

		distortion of human behaviour results from factors external to the AI system which are outside of the control of the provider or the user.	
291	21	AI systems providing social scoring of natural persons for general purpose by public authorities or on their behalf may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non-discrimination and the values of equality and justice. Such AI systems evaluate or classify the trustworthiness of natural persons based on their social behaviour in multiple contexts or known or predicted personal or personality characteristics. The social score obtained from such AI systems may lead to the detrimental or unfavourable treatment of natural persons or whole groups thereof in social contexts, which are unrelated to the context in which the data was originally generated or collected or to a detrimental treatment that is disproportionate or unjustified to the gravity of their social behaviour. Such AI systems should be therefore prohibited.	E2
292	21, 22	The use of AI systems for 'real-time' remote biometric identification of natural persons in publicly accessible spaces for the purpose of law enforcement is considered particularly intrusive in the rights and freedoms of the concerned persons, to the extent that it may affect the private life of a large part of the population, evoke a feeling of constant surveillance and indirectly dissuade the exercise of the freedom of assembly and other fundamental rights. In addition, the immediacy of the impact and the limited opportunities for further checks or corrections in relation to the use of such systems operating in 'real-time' carry	E2, E3

		heightened risks for the rights and freedoms of the persons that are concerned by law enforcement activities. The use of those systems for the purpose of law enforcement should therefore be prohibited, except in three exhaustively listed and narrowly defined situations, where the use is strictly necessary to achieve a substantial public interest, the importance of which outweighs the risks.	
293	22	In order to ensure that those systems are used in a responsible and proportionate manner, it is also important to establish that, in each of those three exhaustively listed and narrowly defined situations, certain elements should be taken into account, in particular as regards the nature of the situation giving rise to the request and the consequences of the use for the rights and freedoms of all persons concerned and the safeguards and conditions provided for with the use. In addition, the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement should be subject to appropriate limits in time and space, having regard in particular to the evidence or indications regarding the threats, the victims or perpetrator.	E3
294	22	Each use of a 'real-time' remote biometric identification system in publicly accessible spaces for the purpose of law enforcement should be subject to an express and specific authorisation by a judicial authority or by an independent administrative authority of a Member State.	E3
295	24	High-risk AI systems should only be placed on the Union market or put into service if they comply with certain mandatory requirements. Those requirements should ensure that high-risk AI systems available in the Union or whose	E1

		output is otherwise used in the Union do not pose unacceptable risks to important Union public interests as recognised and protected by Union law. AI systems identified as high-risk should be limited to those that have a significant harmful impact on the health, safety and fundamental rights of persons in the Union and such limitation minimises any potential restriction to international trade, if any.	
296	24	AI systems could produce adverse outcomes to health and safety of persons, in particular when such systems operate as components of products.	E2
297	24	The extent of the adverse impact caused by the AI system on the fundamental rights protected by the Charter is of particular relevance when classifying an AI system as high-risk. Those rights include the right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information, freedom of assembly and of association, and non-discrimination, consumer protection, workers' rights, rights of persons with disabilities, right to an effective remedy and to a fair trial, right of defence and the presumption of innocence, right to good administration. In addition to those rights, it is important to highlight that children have specific rights as enshrined in Article 24 of the EU Charter and in the United Nations Convention on the Rights of the Child (further elaborated in the UNCRC General Comment No. 25 as regards the digital environment), both of which require consideration of the children's vulnerabilities and provision of such protection and care as necessary for their well-being. The fundamental right to a high level of environmental	E1

		protection enshrined in the Charter and implemented in Union policies should also be considered when assessing the severity of the harm that an AI system can cause, including in relation to the health and safety of persons.	
298	26	As regards stand-alone AI systems, meaning high-risk AI systems other than those that are safety components of products, or which are themselves products, it is appropriate to classify them as high-risk if, in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity of the possible harm and its probability of occurrence and they are used in a number of specifically pre-defined areas specified in the Regulation.	E2
299	26	Technical inaccuracies of AI systems intended for the remote biometric identification of natural persons can lead to biased results and entail discriminatory effects. This is particularly relevant when it comes to age, ethnicity, sex or disabilities. Therefore, 'real-time' and 'post' remote biometric identification systems should be classified as high-risk.	E2
300	26	As regards the management and operation of critical infrastructure, it is appropriate to classify as high-risk the AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity, since their failure or malfunctioning may put at risk the life and health of persons at large scale and lead to appreciable disruptions in the ordinary conduct of social and economic activities.	E2

301	26	AI systems used in education or vocational training, notably for determining access or assigning persons to educational and vocational training institutions or to evaluate persons on tests as part of or as a precondition for their education should be considered high-risk, since they may determine the educational and professional course of a person's life and therefore affect their ability to secure their livelihood. When improperly designed and used, such systems may violate the right to education and training as well as the right not to be discriminated against and perpetuate historical patterns of discrimination.	E2
302	26	AI systems used in employment, workers management and access to self-employment, notably for the recruitment and selection of persons, for making decisions on promotion and termination and for task allocation, monitoring or evaluation of persons in work-related contractual relationships, should also be classified as high-risk, since those systems may appreciably impact future career prospects and livelihoods of these persons.	E2
303	26, 27	Throughout the recruitment process and in the evaluation, promotion, or retention of persons in work-related contractual relationships, such systems may perpetuate historical patterns of discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation. AI systems used to monitor the performance and behaviour of these persons may also impact their rights to data protection and privacy.	E2

304	27	In particular, AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems, since they determine those persons' access to financial resources or essential services such as housing, electricity, and telecommunication services. AI systems used for this purpose may lead to discrimination of persons or groups and perpetuate historical patterns of discrimination, for example based on racial or ethnic origins, disabilities, age, sexual orientation, or create new forms of discriminatory impacts.	E2
305	27	If AI systems are used for determining whether such benefits and services should be denied, reduced, revoked or reclaimed by authorities, they may have a significant impact on persons' livelihood and may infringe their fundamental rights, such as the right to social protection, non-discrimination, human dignity or an effective remedy. Those systems should therefore be classified as high-risk.	E2
306	27	Finally, AI systems used to dispatch or establish priority in the dispatching of emergency first response services should also be classified as high-risk since they make decisions in very critical situations for the life and health of persons and their property.	E2
307	27	In particular, if the AI system is not trained with high quality data, does not meet adequate requirements in terms of its accuracy or robustness, or is not properly designed and tested before being put on the market or otherwise put into service, it may single out people in a discriminatory or otherwise incorrect or unjust manner. Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as	E2, E1, E3

		the right of defence and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable and documented. It is therefore appropriate to classify as high-risk a number of AI systems intended to be used in the law enforcement context where accuracy, reliability and transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress.	
308	28	AI systems used in migration, asylum and border control management affect people who are often in particularly vulnerable position and who are dependent on the outcome of the actions of the competent public authorities. The accuracy, non-discriminatory nature and transparency of the AI systems used in those contexts are therefore particularly important to guarantee the respect of the fundamental rights of the affected persons, notably their rights to free movement, non-discrimination, protection of private life and personal data, international protection and good administration. It is therefore appropriate to classify as high-risk AI systems intended to be used by the competent public authorities charged with tasks in the fields of migration, asylum and border control management as polygraphs and similar tools or to detect the emotional state of a natural person; for assessing certain risks posed by natural persons entering the territory of a Member State or applying for visa or asylum; for verifying the authenticity of the relevant documents of natural persons; for assisting competent public authorities for the examination of applications for asylum, visa and residence permits and associated complaints with	E2, E1

		regard to the objective to establish the eligibility of the natural persons applying for a status.	
309	28	Certain AI systems intended for the administration of justice and democratic processes should be classified as high-risk, considering their potentially significant impact on democracy, rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial. In particular, to address the risks of potential biases, errors and opacity, it is appropriate to qualify as high-risk AI systems intended to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts.	E2, E1
310	29	To mitigate the risks from high-risk AI systems placed or otherwise put into service on the Union market for users and affected persons, certain mandatory requirements should apply, taking into account the intended purpose of the use of the system and according to the risk management system to be established by the provider.	E3
311	29	Requirements should apply to high-risk AI systems as regards the quality of data sets used, technical documentation and record-keeping, transparency and the provision of information to users, human oversight, and robustness, accuracy and cybersecurity. Those requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights, as applicable in the light of the intended purpose of the system, and no other less trade restrictive measures are reasonably available, thus avoiding unjustified restrictions to trade.	E3
312	29	For the development of high-risk AI systems, certain actors, such as providers, notified bodies and other relevant entities,	E3

		such as digital innovation hubs, testing experimentation facilities and researchers, should be able to access and use high quality datasets within their respective fields of activities which are related to this Regulation.	
313	30	Having information on how high-risk AI systems have been developed and how they perform throughout their lifecycle is essential to verify compliance with the requirements under this Regulation.	E3
314	30	To address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately.	E3
315	30	High-risk AI systems should be designed and developed in such a way that natural persons can oversee their functioning. For this purpose, appropriate human oversight measures should be identified by the provider of the system before its placing on the market or putting into service.	E3
316	30	High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity in accordance with the generally acknowledged state of the art. The level of accuracy and accuracy metrics should be communicated to the users.	E3
317	30	The technical robustness is a key requirement for high-risk AI systems. They should be resilient against risks connected to the limitations of the system (e.g. errors, faults, inconsistencies, unexpected situations) as well as against malicious actions that may compromise the security of the	E3

		AI system and result in harmful or otherwise undesirable behaviour.	
318	30	To ensure a level of cybersecurity appropriate to the risks, suitable measures should therefore be taken by the providers of high-risk AI systems, also taking into account as appropriate the underlying ICT infrastructure.	E3
319	31	It is appropriate that a specific natural or legal person, defined as the provider, takes the responsibility for the placing on the market or putting into service of a high-risk AI system, regardless of whether that natural or legal person is the person who designed or developed the system.	E3
320	32	In order to ensure a high level of trustworthiness of high-risk AI systems, those systems should be subject to a conformity assessment prior to their placing on the market or putting into service.	E3
321	33	Certain AI systems intended to interact with natural persons or to generate content may pose specific risks of impersonation or deception irrespective of whether they qualify as high-risk or not. In certain circumstances, the use of these systems should therefore be subject to specific transparency obligations without prejudice to the requirements and obligations for high-risk AI systems. In particular, natural persons should be notified that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use. Moreover, natural persons should be notified when they are exposed to an emotion recognition system or a biometric categorisation system. Such information and notifications should be provided in accessible formats for persons with disabilities. Further, users, who use an AI system to generate or	E1, E2, E3

		manipulate image, audio or video content that appreciably resembles existing persons, places or events and would falsely appear to a person to be authentic, should disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin.	
322	37	It is important that AI systems related to products that are not high-risk in accordance with this Regulation and thus are not required to comply with the requirements set out herein are nevertheless safe when placed on the market or put into service.	E3
323	43	The following artificial intelligence practices shall be prohibited: the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;	E1, E2
324	43	The following artificial intelligence practices shall be prohibited: the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;	E1, E2

325	43	The following artificial intelligence practices shall be prohibited: the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following: detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected; detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;	E1, E2
326	43	The following artificial intelligence practices shall be prohibited: the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement, unless and in as far as such use is strictly necessary for one of the following objectives: [...]	E1
327	45	Irrespective of whether an AI system is placed on the market or put into service independently from the products referred to in points (a) and (b), that AI system shall be considered high-risk where both of the following conditions are fulfilled: (a) the AI system is intended to be used as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II; (b) the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on	E2

		the market or putting into service of that product pursuant to the Union harmonisation legislation listed in Annex II.	
328	45	The Commission is empowered to adopt delegated acts in accordance with Article 73 to update the list in Annex III by adding high-risk AI systems where both of the following conditions are fulfilled: (a) the AI systems are intended to be used in any of the areas listed in points 1 to 8 of Annex III;	E1
329	50	High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of this Title.	E1
330	50	High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users.	E1
331	51	High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.	E1
332	51	Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.	E1, E2

333	51, 52	High-risk AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle.	E1
334	69	Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use.	E1
335	69	Users of an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto.	E1
336	69	Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated.	E1
337	1, 2	It supports the objective of the Union being a global leader in the development of secure, trustworthy and ethical artificial intelligence as stated by the European Council [3] and ensures the protection of ethical principles as specifically requested by the European Parliament [4].	E1, E5
338	18	By laying down those rules, this Regulation supports the objective of the Union of being a global leader in the development of secure, trustworthy and ethical artificial intelligence, as stated by the European Council [33], and it ensures the protection of ethical principles, as specifically requested by the European Parliament [34].	E1; E5

339	5	The proposal also strengthens significantly the Union's role to help shape global norms and standards and promote trustworthy AI that is consistent with Union values and interests. It provides the Union with a powerful basis to engage further with its external partners, including third countries, and at international fora on issues relating to AI.	E5
340	6	Only common action at Union level can also protect the Union's digital sovereignty and leverage its tools and regulatory powers to shape global rules and standards.	E5
341	4	High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:	E2
342	4	Biometric identification and categorisation of natural persons: (a) AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;	E2
343	4	Management and operation of critical infrastructure: (a) AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.	E2
344	4	Education and vocational training: (a) AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions; (b) AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.	E2

345	4	Employment, workers management and access to self-employment: (a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests; (b) AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.	E2
346	4	Access to and enjoyment of essential private services and public services and benefits: (a) AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services; (b) AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems put into service by small scale providers for their own use; (c) AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.	E2

347	4, 5	Law enforcement: (a) AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending or the risk for potential victims of criminal offences; (b) AI systems intended to be used by law enforcement authorities as polygraphs and similar tools or to detect the emotional state of a natural person; (c) AI systems intended to be used by law enforcement authorities to detect deep fakes as referred to in article 52(3); (d) AI systems intended to be used by law enforcement authorities for evaluation of the reliability of evidence in the course of investigation or prosecution of criminal offences; (e) AI systems intended to be used by law enforcement authorities for predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 or assessing personality traits and characteristics or past criminal behaviour of natural persons or groups; (f) AI systems intended to be used by law enforcement authorities for profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of detection, investigation or prosecution of criminal offences; (g) AI systems intended to be used for crime analytics regarding natural persons, allowing law enforcement authorities to search complex related and unrelated large data sets available in different data sources or in different data formats in order to identify unknown patterns or discover hidden relationships in the data.	E2, E1

348	5	Migration, asylum and border control management: (a) AI systems intended to be used by competent public authorities as polygraphs and similar tools or to detect the emotional state of a natural person; (b) AI systems intended to be used by competent public authorities to assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a natural person who intends to enter or has entered into the territory of a Member State; (c) AI systems intended to be used by competent public authorities for the verification of the authenticity of travel documents and supporting documentation of natural persons and detect non-authentic documents by checking their security features; (d) AI systems intended to assist competent public authorities for the examination of applications for asylum, visa and residence permits and associated complaints with regard to the eligibility of the natural persons applying for a status.	E2
349	5	Administration of justice and democratic processes: (a) AI systems intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts.	E2

Document 8: Executive Order 13859 Maintaining American Leadership in Artificial Intelligence

No.	Page	Citation	Category
341	1	The United States is the world leader in AI research and development (R&D) and deployment. Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities.The Federal Government plays an important role in facilitating AI R&D, promoting the trust of the American people in the development and deployment of AI-related technologies, training a workforce capable of using AI in their occupations, and protecting the American AI technology base from attempted acquisition by strategic competitors and adversarial nations. Maintaining American leadership in AI requires a concerted effort to promote advancements in technology and innovation, while protecting American technology, economic and national security, civil liberties, privacy, and American values and enhancing international and industry collaboration with foreign partners and allies. It is the policy of the United States Government to sustain and enhance the scientific, technological, and economic leadership position of the United States in AI R&D and deployment through a coordinated Federal Government strategy, the American AI Initiative (Initiative), guided by five principles.	U6, U7
342	1	(d) The United States must foster public trust and confidence in AI technologies and protect civil liberties, privacy, and	U1, U7

		American values in their application in order to fully realize the potential of AI technologies for the American people.	
343	2	(b) Enhance access to high-quality and fully traceable Federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections consistent with applicable laws and policies.	U2, U3
344	2	(c) Reduce barriers to the use of AI technologies to promote their innovative application while protecting American technology, economic and national security, civil liberties, privacy, and values.	U2, U3
345	2	(d) Ensure that technical standards minimize vulnerability to attacks from malicious actors and reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies; and develop international standards to promote and protect those priorities.	U2, U3, U6
346	4	Within 180 days of the date of this order, the Secretary of Commerce, through the Director of the National Institute of Standards and Technology (NIST), shall issue a plan for Federal engagement in the development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies.	U3

Document 9: A Plan for Federal Engagement in Developing Technical Standards and Related Tools

No.	Page	Citation	Category
347	3	Other aspects, such as trustworthiness, are only now being considered.	U2
348	3	This plan identifies the following nine areas of focus for AI standards: Concepts and terminology, Data and knowledge, Human interactions, Metrics, Networking, Performance testing and reporting methodology, Safety, Risk management, Trustworthiness	U4
349	3	Trustworthiness standards include guidance and requirements for accuracy, explainability, resiliency, safety, reliability, objectivity, and security.	U4
350	4	It is important for those participating in AI standards development to be aware of, and to act consistently with, U.S. government policies and principles, including those that address societal and ethical issues, governance, and privacy.	U1,
351	4	Standards should be complemented by related tools to advance the development and adoption of effective, reliable, robust, and trustworthy AI technologies.	U3
352	4	U.S. government agencies should prioritize involvement in AI standards efforts that are: inclusive and accessible, open and transparent, consensus-based, globally relevant, and non-discriminatory.	U3
353	4	This plan recommends that the Federal government commit to deeper, consistent, long-term engagement in AI standards development activities to help the United States to speed the pace of reliable, robust, and trustworthy AI technology development	U3

354	5	Promote focused research to advance and accelerate broader exploration and understanding of how aspects of trustworthiness can be practically incorporated within standards and standards-related tools.	U3
355	5	Support and expand public-private partnerships to develop and use AI standards and related tools to advance reliable, robust, and trustworthy AI.	U3
356	5	4. Strategically engage with international parties to advance AI standards for U.S. economic and national security needs. Champion U.S. AI standards priorities in AI standards development activities around the world. Accelerate the exchange of information between Federal officials and counterparts in like-minded countries through partnering on development of AI standards and related tools. Track and understand AI standards development strategies and initiatives of foreign governments and entities.	U3, U6, U7
357	8	Increasing trust in AI technologies is a key element in accelerating their adoption for economic growth and future innovations that can benefit society. Today, the ability to understand and analyze the decisions of AI systems and measure their trustworthiness is limited. Among the characteristics that relate to trustworthy AI technologies are accuracy, reliability, resiliency, objectivity, security, explainability, safety, and accountability. Ideally, these aspects of AI should be considered early in the design process and tested during the development and use of AI technologies. AI standards and related tools, along with AI risk management strategies, can help to address this limitation and spur innovation.	U2, U3, U7

358	8	AI standards that articulate requirements, specifications, guidelines, or characteristics can help to ensure that AI technologies and systems meet critical objectives for functionality, interoperability, and trustworthiness—and that they perform accurately, reliably, and safely.	U4, U3
359	8	In contrast, standards that are not fit-for-purpose, are not available when needed, or that are designed around less than ideal technological solutions may hamper innovation and constrain the effective or timely development and deployment of reliable, robust, and trustworthy AI technologies.	U2
360	8, 9	Global cooperation and coordination on AI standards will be critical for having a consistent set of "rules of the road" to enable market competition, preclude barriers to trade, and allow innovation to flourish. The U.S. government should ensure cooperation and coordination across Federal agencies and partner with private sector stakeholders to continue to shape international dialogues in regards to AI standards development.	U3, U6, U4, U7
361	10	There are several existing technology standards applicable to AI that were originally developed for other technologies. Standards related to data formats, testing methodology, transfer protocols, cybersecurity, and privacy are examples.	U4
362	11	Lastly, even where standards are noted as available or being developed, each area could likely benefit from additional standards to advance or keep pace with AI technologies, and their widespread use, in a reliable, robust, and trustworthy manner.	U3
363	12	By defining common vocabularies, establishing the essential characteristics of reliable, robust, and trustworthy AI	U3

		technologies, and identifying best practices within the life cycle of an AI system, these standards can accelerate the pace of innovation.	
364	12	Trustworthiness standards include guidance and requirements for: accuracy, explainability, resiliency, safety, reliability, objectivity, and security.	U4
365	12, 13	In terms of developing standards for societal and ethical considerations, it is important to distinguish between technical and non-technical standards. Not all societal and ethical issues of AI can be addressed by developing technical standards [23]. Non-technical standards can inform policy and human decision-making [24].	U1, U4
366	13	Standards should be complemented by an array of related tools to advance the development and adoption of effective, reliable, robust, and trustworthy AI technologies.	U3
367	15	Like several other pioneering areas of science and technology, the development of AI raises a host of legal, ethical, and societal issues that create real and perceived challenges for developers, policy makers, and users—including the general public.	U1
368	15, 16	In this arena, standards flow from principles, and a first step toward standardization will be reaching broad consensus on a core set of AI principles. These kinds of principles are being forged by multiple organizations, including the Organisation for Economic Cooperation and Development (OECD), whose member countries (including the United States) recently adopted such principles [32].	U1, U6
369	16	While stakeholders in the development of this plan expressed broad agreement that societal and ethical	U1

		considerations must factor into AI standards, it is not clear how that should be done and whether there is yet sufficient scientific and technical basis to develop those standards provisions. Moreover, legal, societal, and ethical considerations should be considered by specialists trained in law and ethics.	
370	16	The degree to which ethical considerations might be incorporated into standards should be tied tightly to the type, likelihood, degree, and consequence of risk to humans,	U1
371	16	Privacy risks are different depending on the use case, the type of data involved, the societal and cultural context, and many other factors. Privacy considerations should be included in any standards governing the collection, processing, sharing, storage, and disposal of personal information, and	U2
372	16	Standards should facilitate AI systems that function in a robust, secure and safe way throughout their life cycles.	U3
373	16	Legal, ethical, and societal considerations also can come into play as developers and policy makers consider whether and how to factor in the management of risk to individuals, communities, and society at large. Some standards and standards-related tools aim to provide guidance for evaluating risks, which can be used by developers and policy makers in considering how to manage those risks. Ultimately, it is up to system owners and users to determine what risks they are willing to accept, mitigate, or avoid within existing regulations and policies.	U1, U2
374	16	The degree of potential risk presented by particular AI technologies and systems will help to drive decision making	U2

		about the need for specific AI standards and standards-related tools.	
375	19	Human-centered to ensure that human interactions and values—including abilities, disabilities, diversity—are considered during AI data collection, model development, testing, and deployment.	U1
376	19	Sensitive to ethical considerations, identifying and minimizing bias, and incorporating provisions that protect privacy and reflect the broader community's notions of acceptability.	U1
377	19	U.S. engagement in establishing AI standards is critical; AI standards developed without the appropriate level and type of involvement may exclude or disadvantage U.S.-based companies in the marketplace as well as U.S. government agencies. Furthermore, due to the foundational nature of standards, the lack of U.S. stakeholder engagement in the development of AI standards can degrade the innovativeness and competitiveness of the U.S. in the long term.	U6, U7
378	22	In addition to the guidance provided regarding priorities and levels of engagement called for in the previous section of this plan, the Federal government should commit to deeper, consistent, long-term engagement in AI standards development activities to help the United States to speed the pace of reliable, robust, and trustworthy AI technology development.	U3, U7

Document 10: Guidance for Regulation of Artificial Intelligence Applications

No.	Page	Citation	Category
379	1	When considering regulations or policies related to AI applications, agencies should continue to promote advancements in technology and innovation, while protecting American technology, economic and national security, privacy, civil liberties, and other American values, including the principles of freedom, human rights, the rule of law, and respect for intellectual property.	U1, U7
380	2	The importance of developing and deploying AI requires a regulatory approach that fosters innovation, growth, and engenders trust, while protecting core American values, through both regulatory and non-regulatory actions and reducing unnecessary barriers to the development and deployment of AI.	U1, U3
381	3	Given that many AI applications do not necessarily raise novel issues, these considerations also reflect longstanding Federal regulatory principles and practices that are relevant to promoting the innovative use of AI. Promoting innovation and growth of AI is a high priority of the United States government. Fostering innovation and growth through forbearing from new regulations may be appropriate. Agencies should consider new regulation only after they have reached the decision, in light of the foregoing section and other considerations, that Federal regulation is necessary.	U3
382	2	To that end, Federal agencies must avoid regulatory or non-regulatory actions that needlessly hamper AI innovation and growth. Where permitted by law, when deciding whether	U2, U7

		and how to regulate in an area that may affect AI applications, agencies should assess the effect of the potential regulation on AI innovation and growth. Agencies must avoid a precautionary approach that holds AI systems to such an impossibly high standard that society cannot enjoy their benefits. Where AI entails risk, agencies should consider the potential benefits and costs of employing AI, when compared to the systems AI has been designed to complement or replace.	
383	3	Public Trust in AI: AI is expected to have a positive impact across sectors of social and economic life, including employment, transportation, education, finance, healthcare, personal security, and manufacturing. At the same time, AI applications could pose risks to privacy, individual rights, autonomy, and civil liberties that must be carefully assessed and appropriately addressed. Its continued adoption and acceptance will depend significantly on public trust and validation. It is therefore important that the government's regulatory and non-regulatory approaches to AI promote reliable, robust, and trustworthy AI applications, which will contribute to public trust in AI. The appropriate regulatory or non-regulatory response to privacy and other risks must necessarily depend on the nature of the risk presented and the appropriate mitigations.	U2, U3
384	3	Public Participation Public participation, especially in those instances where AI uses information about individuals, will improve agency accountability and regulatory outcomes, as well as increase public trust and confidence. Agencies should provide ample opportunities for the public to provide information and participate in all stages of the	U1, U3

		rulemaking process, to the extent feasible and consistent with legal requirements (including legal constraints on participation in certain situations, for example, national security preventing imminent threat to or responding to emergencies). Agencies are also encouraged to the extent practicable, to inform the public and promote awareness and widespread availabilitiy of standards and the creation of other infomative documents.	
385	4	Scientific Integrity and Information Quality: The government's regulatory and non-regulatory approaches to AI applications should leverage scientific and technical information and processes. Agencies should hold information, whether produced by the government or acquired by the government from third parties, that is likely to have a clear and substantial influence on important public policy or private sector decisions (including those made by consumers) to a high standard of quality, transparency, and compliance. Consistent with the principles of scientific integrity in the rulemaking and guidance processes, agencies should develop regulatory approaches to AI in a manner that both informs policy decisions and fosters public trust in AI. Best practices include transparently articulating the strengths, weaknesses, intended optimizations or outcomes, bias mitigation, and appropriate uses of the AI application's results. Agencies should also be mindful that, for AI applications to produce predictable, reliable, and optimized outcomes, data used to train the AI system must be of sufficient quality for the intended use.	U1, U3

386	4	Risk Assessment and Management: Regulatory and non-regulatory approaches to AI should be based on a consistent application of risk assessment and risk management across various agencies and various technologies. It is not necessary to mitigate every foreseeable risk; in fact, a foundational principle of regulatory policy is that all activities involve tradeoffs. Instead, a risk-based approach should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits. Agencies should be transparent about their evaluations of risk and re-evaluate their assumptions and conclusions at appropriate intervals so as to foster accountability. Correspondingly, the magnitude and nature of the consequences should an AI tool fail, or for that matter succeed, can help inform the level and type of regulatory effort that is appropriate to identify and mitigate risks. Specifically, agencies should follow the direction in Executive Order 12866, "Regulatory Planning and Review,"4 to consider the degree and nature of the risks posed by various activities within their jurisdiction. Such an approach will, where appropriate, avoid hazard-based and unnecessarily precautionary approaches to regulation that could unjustifiably inhibit innovation.	U2, U3
387	4, 5	Benefits and Costs: When developing regulatory and non-regulatory approaches, agencies will often consider the application and deployment of AI into already-regulated industries. Presumably, such significant investments would not occur unless they offered significant economic potential. As in all technological transitions of this nature, the introduction of AI may also create unique challenges. For	U3

		example, while the broader legal environment already applies to AI applications, the application of existing law to questions of responsibility and liability for decisions made by AI could be unclear in some instances, leading to the need for agencies, consistent with their authorities, to evaluate the benefits, costs, and distributional effects associated with any identified or expected method for accountability. Executive Order 12866 calls on agencies to "select those approaches that maximize net benefits (including potential economic, environmental, public health and safety, and other advantages; distributive impacts; and equity)" [6]. Agencies should, when consistent with law, carefully consider the full societal costs, benefits, and distributional effects before considering regulations related to the development and deployment of AI applications. Such consideration will include the potential benefits and costs of employing AI, when compared to the systems AI has been designed to complement or replace, whether implementing AI will change the type of errors created by the system, as well as comparison to the degree of risk tolerated in other existing ones. Agencies should also consider critical dependencies when evaluating AI costs and benefits, as technological factors (such as data quality) and changes in human processes associated with AI implementation may alter the nature and magnitude of the risks and benefits. In cases where a comparison to a current system or process is not available, evaluation of risks and costs of not implementing the system should be evaluated as well.	

388	5	Flexibility: When developing regulatory and non-regulatory approaches, agencies should pursue performance-based and flexible approaches that can adapt to rapid changes and updates to AI applications. Rigid, design-based regulations that attempt to prescribe the technical specifications of AI applications will in most cases be impractical and ineffective, given the anticipated pace with which AI will evolve and the resulting need for agencies to react to new information and evidence. Targeted agency conformity assessment schemes, to protect health and safety, privacy, and other values, will be essential to a successful, and flexible, performance-based approach. To advance American innovation, agencies should keep in mind international uses of AI, ensuring that American companies are not disadvantaged by the United States' regulatory regime.	U3
389	5	Fairness and Non-Discrimination Agencies should consider in a transparent manner the impacts that AI applications may have on discrimination. AI applications have the potential of reducing present-day discrimination caused by human subjectivity. At the same time, applications can, in some instances, introduce real-world bias that produces discriminatory outcomes or decisions that undermine public trust and confidence in AI. When considering regulations or non-regulatory approaches related to AI applications, agencies should consider, in accordance with law, issues of fairness and non-discrimination with respect to outcomes and decisions produced by the AI application at issue, as well as whether the AI application at issue may reduce levels of unlawful, unfair, or otherwise unintended discrimination as compared to existing processes.	U1, U3

390	6	Disclosure and Transparency: In addition to improving the rulemaking process, transparency and disclosure can increase public trust and confidence in AI applications. At times, such disclosures may include identifying when AI is in use, for instance, if appropriate for addressing questions about how the application impacts human end users. Agencies should be aware that some applications of AI could increase human autonomy. Agencies should carefully consider the sufficiency of existing or evolving legal, policy, and regulatory environments before contemplating additional measures for disclosure and transparency. What constitutes appropriate disclosure and transparency is context-specific, depending on assessments of potential harms, the magnitude of those harms, the technical state of the art, and the potential benefits of the AI application.	U1, U3
391	6	Safety and Security: Agencies should promote the development of AI systems that are safe, secure, and operate as intended, and encourage the consideration of safety and security issues throughout the AI design, development, deployment, and operation process. Agencies should pay particular attention to the controls in place to ensure the confidentiality, integrity, and availability of the information processed, stored, and transmitted by AI systems. Agencies should give additional consideration to methods for guaranteeing systemic resilience, and for preventing bad actors from exploiting AI system weaknesses, including cybersecurity risks posed by AI operation, and adversarial use of AI against a regulated entity's AI technology. When evaluating or introducing AI policies, agencies should be mindful of any potential safety and security risks, as well as	U1, U3

		the risk of possible malicious deployment and use of AI applications.	
392	6	Interagency Coordination: A coherent and whole-of-government approach to AI oversight requires interagency coordination. Agencies should coordinate with each other to share experiences and to ensure consistency and predictability of AI-related policies that advance American innovation and growth in AI, while appropriately protecting privacy, civil liberties, and American values and allowing for sector-and application-specific approaches when appropriate. When OMB's Office of Information and Regulatory Affairs (OIRA) designates AI-related draft regulatory action as "significant" for purposes of interagency review under Executive Order 12866, OIRA will ensure that all agencies potentially affected by or interested in a particular action will have an opportunity to provide input.	U3
393	6	Agencies should promote the development of AI systems that are safe, secure, and operate as intended, and encourage the consideration of safety and security issues throughout the AI design, development, deployment, and operation process.	U3
394	6	When evaluating or introducing AI policies, agencies should be mindful of any potential safety and security risks, as well as the risk of possible malicious deployment and use of AI applications.	U2
395	8	Consistent with the principles described in this Memorandum, agencies should communicate with the	U1

		public about the benefits and risks of AI in a manner that gives the public appropriate trust and understanding of AI.	
396	9	Executive Order 13859 calls for Federal engagement in the development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies. To promote innovation, use, and adoption of AI applications, standards could address many technical aspects, such as AI performance, measurement, safety, security, privacy, interoperability, robustness, trustworthiness, and governance.	U3
397	9	Accordingly, agencies should engage in dialogues to promote consistent regulatory approaches to AI that promote American AI innovation while protecting privacy, civil rights, civil liberties, and American values. Such discussions, including those with the general public, can provide valuable opportunities to share best practices, data, and lessons learned, and ensure that America remains at the forefront of AI development.	U3,

Document 11: Executive Order 13960 Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government

No.	Page	Citation	Category
398	1	The ongoing adoption and acceptance of AI will depend significantly on public trust. Agencies must therefore design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable law and the goals of Executive Order 13859.	U1, U7

399	1	Purpose. Artificial intelligence (AI) promises to drive the growth of the United States economy and improve the quality of life of all Americans.	U7
400	1;2	It is the policy of the United States to promote the innovation and use of AI, where appropriate, to improve Government operations and services in a manner that fosters public trust, builds confidence in AI, protects our Nation's values, and remains consistent with all applicable laws, including those related to privacy, civil rights, and civil liberties.	U1
401	2	It is the policy of the United States that responsible agencies, as defined in section 8 of this order, shall, when considering the design, development, acquisition, and use of AI in Government, be guided by the common set of Principles set forth in section 3 of this order, which are designed to foster public trust and confidence in the use of AI, protect our Nation's values, and ensure that the use of AI remains consistent with all applicable laws, including those related to privacy, civil rights, and civil liberties	U1, U7
402	2	(a) Lawful and respectful of our Nation's values. Agencies shall design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation's values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.	U1
403	2	(b) Purposeful and performance-driven. Agencies shall seek opportunities for designing, developing, acquiring, and using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed.	U1

404	2	(c) Accurate, reliable, and effective. Agencies shall ensure that their application of AI is consistent with the use cases for which that AI was trained, and such use is accurate, reliable, and effective.	U1
405	2	(d) Safe, secure, and resilient. Agencies shall ensure the safety, security, and resiliency of their AI applications, including resilience when confronted with systematic vulnerabilities, adversarial manipulation, and other malicious exploitation.	U2
406	2	(e) Understandable. Agencies shall ensure that the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate.	U1
407	2	(f) Responsible and traceable. Agencies shall ensure that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI. Agencies shall ensure that AI is used in a manner consistent with these Principles and the purposes for which each use of AI is intended. The design, development, acquisition, and use of AI, as well as relevant inputs and outputs of particular AI applications, should be well documented and traceable, as appropriate and to the extent practicable.	U1, U3
408	2	(g) Regularly monitored. Agencies shall ensure that their AI applications are regularly tested against these Principles. Mechanisms should be maintained to supersede, disengage, or deactivate existing applications of AI that demonstrate performance or outcomes that are inconsistent with their intended use or this order.	U3

409	2	(h) Transparent. Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders, including the Congress and the public, to the extent practicable and in accordance with applicable laws and policies, including with respect to the protection of privacy and of sensitive law enforcement, national security, and other protected information.	U1
410	2;3	(i) Accountable. Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI, and shall monitor, audit, and document compliance with those safeguards. Agencies shall provide appropriate training to all agency personnel responsible for the design, development, acquisition, and use of AI.	U1, U3

Document 12: National Artificial Intelligence Initiative Act

No.	Page	Citation	Category
411	3	ESTABLISHMENT; PURPOSES.—The President shall establish and implement an initiative to be known as the "National Artificial Intelligence Initiative". The purposes of the Initiative shall be to (2) lead the world in the development and use of trustworthy artificial intelligence systems in the public and private sectors;	U1, U7
412	6	(d) RESPONSIBILITIES.—The Interagency Committee shall—	U3
413	6	(2) not later than 2 years after the date of the enactment of this Act, develop a strategic plan for artificial intelligence (to be updated not less than every 3 years) that establishes goals,	U3

		priorities, and metrics for guiding and evaluating how the agencies carrying out the Initiative will—	
414	6	(C) support research and other activities on ethical, legal, environmental, safety, security, bias, and other appropriate societal issues related to artificial intelligence;	U1
415	6	(D) provide or facilitate the availability of curated, standardized, secure, representative, aggregate, and privacy-protected data sets for artificial intelligence research and development;	U3
416	7	(d) DUTIES.—The Advisory Committee shall advise the President and the Initiative Office on matters related to the Initiative, including recommendations related to— (10) whether ethical, legal, safety, security, and other appropriate societal issues are adequately addressed by the Initiative;	U1
417	8	(11) opportunities for international cooperation with strategic allies on artificial intelligence research activities, standards development, and the compatibility of international regulations;	U6
418	8	(12) accountability and legal rights, including matters relating to oversight of artificial intelligence systems using regulatory and nonregulatory approaches, the responsibility for any violations of existing laws by an artificial intelligence system, and ways to balance advancing innovation while protecting individual rights; and	U1
419	9	(2) ADVICE.—The subcommittee shall provide advice to the President on matters relating to the development of artificial intelligence relating to law enforcement, including advice on the following:	U3

420	9	(A) Bias, including whether the use of facial recognition by government authorities, including law enforcement agencies, is taking into account ethical considerations and addressing whether such use should be subject to additional oversight, controls, and limitations.	U1
421	9	(B) Security of data, including law enforcement's access to data and the security parameters for that data.	U2
422	9	(C) Adoptability, including methods to allow the United States Government and industry to take advantage of artificial intelligence systems for security or law enforcement purposes while at the same time ensuring the potential abuse of such technologies is sufficiently mitigated.	U3
423	9	(D) Legal standards, including those designed to ensure the use of artificial intelligence systems are consistent with the privacy rights, civil rights and civil liberties, and disability rights issues raised by the use of these technologies.	U1
424	17	The National Institute of Standards and Technology Act (15 U.S.C. 271 et seq.) is amended by inserting after section 22 the following:	U4
425	18	"(a) MISSION.—The Institute shall—	U3
426	18	"(1) advance collaborative frameworks, standards, guidelines, and associated methods and techniques for artificial intelligence;	U3
427	18	"(2) support the development of a risk-mitigation framework for deploying artificial intelligence systems;	U3
428	18	"(3) support the development of technical standards and guidelines that promote trustworthy artificial intelligence systems; and	U3

429	18	"(4) support the development of technical standards and guidelines by which to test for bias in artificial intelligence training data and applications.	U3
430	18	"(b) SUPPORTING ACTIVITIES.—The Director of the National Institute of Standards and Technology may—	U3
431	18	"(1) support measurement research and development of best practices and voluntary standards for trustworthy artificial intelligence systems, which may include—	U3
432	18	"(A) privacy and security, including for datasets used to train or test artificial intelligence systems and software and hardware used in artificial intelligence systems;	U2
433	18	"(D) safety and robustness of artificial intelligence systems, including assurance, verification, validation, security, control, and the ability for artificial intelligence systems to withstand unexpected inputs and adversarial attacks;	U2
434	18	"(E) auditing mechanisms and benchmarks for accuracy, transparency, verifiability, and safety assurance for artificial intelligence systems;	U3
435	19	"(c) RISK MANAGEMENT FRAMEWORK.—Not later than 2 years after the date of the enactment of this Act, the Director shall work to develop, and periodically update, in collaboration with other public and private sector organizations, including the National Science Foundation and the Department of Energy, a voluntary risk management framework for trustworthy artificial intelligence systems. The framework shall—	U3
436	19	"(1) identify and provide standards, guidelines, best practices, methodologies, procedures and processes for—	U3
437	19	"(A) developing trustworthy artificial intelligence systems;	U3

438	19	"(B) assessing the trustworthiness of artificial intelligence systems; and	U3
439	19	"(C) mitigating risks from artificial intelligence systems;	U3
440	19	"(2) establish common definitions and characterizations for aspects of trustworthiness, including explainability, transparency, safety, privacy, security, robustness, fairness, bias, ethics, validation, verification, interpretability, and other properties related to artificial intelligence systems that are common across all sectors;	U4
441	20	"(3) provide case studies of framework implementation;	U3
442	20	"(4) align with international standards, as appropriate;	U6
443	20	"(5) incorporate voluntary consensus standards and industry best practices; and	U6
444	20	"(6) not prescribe or otherwise require the use of specific information or communications technology products or services.	U3
445	20	"(d) PARTICIPATION IN STANDARD SETTING ORGANIZATIONS.—	U3
446	20	"(1) REQUIREMENT.—The Institute shall participate in the development of standards and specifications for artificial intelligence.	U3
447	20	"(2) PURPOSE.—The purpose of this participation shall be to ensure—	U3
448	20	"(A) that standards promote artificial intelligence systems that are trustworthy; and	U1
449	20	"(B) that standards relating to artificial intelligence reflect the state of technology and are fit-for-purpose and developed in transparent and consensus-based processes that are open to all stakeholders.	U1

Document 13: Draft Taxonomy of AI Risk

No.	Page	Citation	Category
450	1	Among other things, in that RFI, NIST proposed eight characteristics of trustworthy AI. This paper aims to provide context to the eight characteristics of trustworthy AI mentioned in the RFI, clarify the distinction between characteristics and principles, and advance discussions about AI risks and forge agreements across organizations and internationally to the benefit AI design, development, use, and evaluation.	U4, U2, U6
451	2	The National Institute of Standards and Technology (NIST) aims to cultivate trust in the design, development, use, and governance of Artificial Intelligence (AI) technologies and systems in ways that enhance economic security and improve quality of life. NIST focuses on improving measurement science, technology, standards, and related tools – including evaluation and data.	U4
452	2	The paper starts by identifying several relevant policy directives that identify sources or types of risk across the AI lifecycle. For example, the Organisation for Economic Co-operation and Development (OECD) AI principles1 specify that AI needs to have: Traceability to human values such as rule of law, human rights, democratic values, and diversity, and ensuring fairness and justice, Transparency and responsible disclosure so people can understand and challenge AI-based outcome, Robustness, security, and safety, through the AI lifecycle to manage risks, Accountability in line with these principles	U2, U6

453	2	Similarly, the European Union Digital Strategy's Ethics Guidelines for Trustworthy AI [2] identifies seven key principles of trustworthy AI: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination, and fairness, Environmental and societal well-being, Accountability	U6
454	2;3	Finally, US Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government3 specifies that AI should be: Lawful and respectful of our Nation's values. Purposeful and performance-driven… using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed, Accurate, reliable, and effective Safe, secure, and resilient, Understandable…by subject matter experts, users, and others, as appropriate, Responsible and traceable, Regularly monitored, Transparent, Accountable.	U6
455	3	Those three documents indicate that AI system stakeholders must account for several different sources of risk in the AI lifecycle. This proposed taxonomy seeks to simplify the categorization of these risks so that stakeholders may better recognize and manage them. The approach is hierarchical. First, it is recognized that there are three broad categories of risk sources related to AI systems:	U2
456	3	1) Technical design attributes. This refers to the factors that are under the direct control of system designers and developers, and which may be measured using standard evaluation criteria that have traditionally been applied to machine learning systems, or that may be applied in an automated way in the future. Examples include accuracy and	U2

		related measures (e.g., false positive and false negative rates, precision, recall, F-score) but also sources of statistical error that might be measured by applying AI tools to new data (e.g., discrepancies between performance on test and holdout sets). Finally, data generated from experiments that are designed to evaluate system performance also fall into this category, and might include tests of causal hypotheses, assessments of robustness to adversarial attack, etc.	
457	3	2) How AI systems are perceived. This refers to mental representations of models, including whether the output provided is sufficient to evaluate compliance (transparency), whether model operations can be easily understood (explainability), and whether they provide output that can be used to make a meaningful decision (interpretability). In general, any judgment or assessment of an AI system, or its output, that is made by a human or needs human interpretation rather than by an automated process falls into this category.	U2
458	3	3) Guiding policies and principles. This refers to broader societal determinations of value, such as privacy, accountability, fairness, justice, equity, etc., which cannot be measured consistently across domains because of their dependence on context.	U1
459	4	Accuracy: This trustworthiness attribute captures the broad notion of whether the machine learning model is correctly capturing a relationship that exists within training data.	U4
460	4	Reliability: A model is reliable if its output is insensitive to small changes in its input, and if it is free from measurement bias.	U4

461	4	Robustness: A model is robust if it applies to multiple settings beyond which it was trained.	U4
462	4	Resilience or Security: A model that is insensitive to adversarial attacks, or more generally, to unexpected changes in its environment or use, may be said to be resilient and secure.	U4
463	5	Human judgment must be employed when deciding on the specific metrics, and the precise values of these metrics. Additionally, human users will also make judgments regarding what these metrics, and the associated models, mean when applied to daily life. Thus, a second broad category of risk pertains to how these human judgments are made. These include:	U2
464	5	Explainability: Attempts to increase explainability seek to provide a programmatic description of how model predictions are generated [9]. The underlying assumption is that perceptions of risk stem from a lack of technical background knowledge on the part of the user. Even given all the information required to make a model fully transparent, a human must apply what technical expertise they have to understand how the model works. Explainability refers to the user's perception of how the model works – such as what output may be expected for a given input. Risks due to explainability may arise if humans incorrectly infer a model's operation and it does not operate as expected.	U2
465	5	Interpretability: Attempts to increase interpretability seek to fill a meaning deficit [10]. The underlying assumption is that perceptions of risk stem from a lack of ability to make sense of, or contextualize, model output appropriately.	U2

466	5	Interpretability is the glue that links transparency – information provided along with a model's output – to determinations that have to do with values (e.g., privacy, safety).	U2
467	6	Privacy. Like safety and security, specific technical features of a system may promote privacy and assessors can identify how the processing of data could create privacy-related problems. However, determinations of likelihood and severity of impact of these problems are contextual and vary between cultures and individuals. Furthermore, ensuring fairness may require violating privacy and vice versa (since fairness determinations often require obtaining data that some consider private).	U2
468	6	Safety. In the context of medical devices and drugs, safety is a categorical determination made by domain experts: a drug is either deemed "safe and efficacious" or it is not. These determinations are made relative to the state of the art in the field, and relative to society's expectations.	U2
469	6	Managing bias. Schwartz et al. [15] point out that bias is neither new nor unique to AI, nor can bias be eliminated entirely. Rather, biases which are harmful must be identified and, to the extent possible, understood, measured, managed, and reduced. Furthermore, perceptions of bias are also human judgments. Thus, perceptions of bias are intimately related to interpretations of model output.	U2
470	6	Human judgments are premised on guiding policies and principles – broad social constructs that indicate societal priorities. AI has the potential to benefit nearly all aspects of our society, but the development and use of new AI-based technologies, products, and services bring technical and	U1,

		societal challenges and risks, including risks to ethical values. While there is no objective standard for ethical values, as they are grounded in the norms and legal expectations of specific societies or cultures, it is widely agreed that AI must be developed in a trustworthy manner. This trustworthiness can support the development and deployment of AI in ways that meet a given set of ethical values.	
471	6;7	Several of the policy documents cited above outline broad statements of values to which AI should adhere.	U1
472	7	Principles relevant to AI include: Fairness. Like safety, standards of fairness are culturally determined, and perceptions of fairness differ between cultures, with societal determinations of fairness litigated in courts. Engineers often assume that machine learning algorithms are inherently fair because the same procedure applies regardless of user; however, this perception has eroded recently as awareness of biased algorithms and biased datasets has increased. Arguably, absence of harmful bias is a necessary condition for fairness	U1
473	7	Accountability: Determinations of accountability are closely related to notions of risk and "blame" – that is, the responsible party in the event that a risky outcome is realized. Anthropologists, including Mary Douglas [16], have written extensively on how perceptions of risk and blame associated with technology differ systematically between cultures, and legal scholars [17] have developed psychometric measures of cultural cognition that are theorized to vary with these risk perceptions.	U1

474	7	Transparency: Attempts to increase transparency seek to fill a perceived information deficit. The underlying assumption is that perceptions of risk stem from an absence of information. Transparency reflects the extent to which information is available to a decision-maker when making a judgment about an AI system, and may span the scope from what data were included in model training, the structure of the model, its intended use case, to how decisions were made, by whom, when, etc. Absent transparency, users are left to guess about these factors and may make unwarranted assumptions regarding model provenance. Although it is impossible to remove a subject's background knowledge from their evaluations of a model, making adequate knowledge available is a precursor to building trust. This risk may be mitigated by a transparent process – one in which users can get answers regarding what decisions were made and what resources (e.g., data, energy, etc.) were used throughout the lifecycle, and why these decisions were made. This highlights the importance of documenting information in a standardized manner throughout the development lifecycle of an AI algorithm (i.e., the need for a "transparency toolkit.") Beyond such a toolkit, users' perceptions of systems as transparent are crucial. This emphasizes the need to develop approaches (e.g., a convenient user interface and cataloguing system, and possibly human contact) to surface this information when needed or requested, potentially in a context-sensitive manner. Finally, transparency is often framed as an instrumental value – a means to the end of achieving a broader value, such as accountability.	U2, U3

Document 14: Blueprint for an AI Bill of Rights

No.	Page	Citation	Category
475	8	This framework applies to (1) automated systems that (2) have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services. These rights, opportunities, and access to critical resources of services should be enjoyed equally and be fully protected, regardless of the changing role that automated systems may play in our lives.	U1, U4
476	8	This framework describes protections that should be applied with respect to all automated systems that have the potential to meaningfully impact individuals' or communities' exercise of: RIGHTS, OPPORTUNITIES, OR ACCESS	U4
477	8	Civil rights, civil liberties, and privacy, including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy and other freedoms in both public and private sector contexts;	U1
478	8	Equal opportunities, including equitable access to education, housing, credit, employment, and other programs; or,	U1
479	8	Access to critical resources or services, such as healthcare, financial services, safety, social services, on-deceptive information about goods and services, and government benefits.	U1
480	8	Considered together, the five principles and associated practices of the Blueprint for an AI Bill of Rights form an overlapping set of backstops against potential harms. This purposefully overlapping framework, when taken as a	U1, U2

		whole, forms a blueprint to help protect the public from harm. The measures taken to realize the vision set forward in this framework should be proportionate with the extent and nature of the harm, or risk of harm, to people's rights, opportunities, and access.	
481	9	The Blueprint for an AI Bill of Rights is meant to assist governments and the private sector in moving principles into practice	U3
482	9	This framework instead shares a broad, forward-leaning vision of recommended principles for automated system development and use to inform private and public involvement with these systems where they have the potential to meaningfully impact rights, opportunities, or access.	U4, U3
483	14	The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence.	U4
484	15	SAFE AND EFFECTIVE SYSTEMS: You should be protected from unsafe or ineffective systems. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards. Outcomes of these protective measures should include the possibility of not deploying the system or	U1, U3

		removing a system from use. Automated systems should not be designed with an intent or reasonably foreseeable possibility of endangering your safety or the safety of your community. They should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts of automated systems. You should be protected from inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of its reuse. Independent evaluation and reporting that confirms that the system is safe and effective, including reporting of steps taken to mitigate potential harms, should be performed and the results made public whenever possible.	
485	17	In order to ensure that an automated system is safe and effective, it should include safeguards to protect the public from harm in a proactive and ongoing manner; avoid use of data inappropriate for or irrelevant to the task at hand, including reuse that could cause compounded harm; and demonstrate the safety and effectiveness of the system.	U3
486	23	ALGORITHMIC DISCRIMINATION Protections: You should not face discrimination by algorithms and systems should be used and designed in an equitable way. Algorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran	U1, U3

		status, genetic infor-mation, or any other classification protected by law. Depending on the specific circumstances, such algorithmic discrimination may violate legal protections. Designers, developers, and deployers of automated systems should take proactive and continuous measures to protect individuals and communities from algorithmic discrimination and to use and design systems in an equitable way. This protection should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities in design and development, pre-deployment and ongoing disparity testing and mitigation, and clear organizational oversight. Independent evaluation and plain language reporting in the form of an algorithmic impact assessment, including disparity testing results and mitigation information, should be performed and made public whenever possible to confirm these protections.	
487	24	There is extensive evidence showing that automated systems can produce inequitable outcomes and amplify existing inequity. Data that fails to account for existing systemic biases in American society can result in a range of consequences.	U2
488	24	Instances of discriminatory practices built into and resulting from AI and other automated systems exist across many industries, areas, and contexts. While automated systems have the capacity to drive extraordinary advances and innovations, algorithmic discrimination protections should be built into their design, deployment, and ongoing use.	U1, U2

489	24	The guardrails protecting the public from discrimination in their daily lives should include their digital lives and impacts—basic safeguards against abuse, bias, and discrimination to ensure that all people are treated fairly when automated systems are used.	U1
490	26	Any automated system should be tested to help ensure it is free from algorithmic discrimination before it can be sold or used. Protection against algorithmic discrimination should include designing to ensure equity, broadly construed. Some algorithmic discrimination is already prohibited under existing anti-discrimination law.	U1, U3
491	30	Data Protection: You should be protected from abusive data practices via built-inprotections and you should have agency over how data about you is used. You should be protected from violations of privacy through design choices that ensure such protections are included by default, including ensuring that data collection conforms to reasonable expectations and that only data strictly necessary for the specific context is collected. Designers, developers, and deployers of automated systems should seek your permission and respect your decisions regarding collection, use, access, transfer, and deletion of your data in appropriate ways and to the greatest extent possible; where not possible, alternative privacy by design safeguards should be used. Systems should not employ user experience and design decisions that obfuscate user choice or burden users with defaults that are privacy invasive. Consent should only be used to justify collection of data in cases where it can be appropriately and meaningfully given. Any consent requests should be brief, be understandable in plain	U1, U3

		language, and give you agency over data collection and the specific context of use; current hard-to-understand notice-and-choice practices for broad uses of data should be changed. Enhanced protections and restrictions for data and inferences related to sensitive domains, including health, work, education, criminal justice, and finance, and for data pertaining to youth should put you first. In sensitive domains, your data and related inferences should only be used for necessary functions, and you should be protected by ethical review and use prohibitions. You and your communities should be free from unchecked surveillance; surveillance technologies should be subject to heightened oversight that includes at least pre-deployment assessment of their potential harms and scope limits to protect privacy and civil liberties. Continuous surveillance and monitoring should not be used in education, work, housing, or in other contexts where the use of such surveillance technologies is likely to limit rights, opportunities, or access. Whenever possible, you should have access to reporting that confirms your data decisions have been respected and provides an assessment of the potential impact of surveillance technologies on your rights, opportunities, or access.	
492	31	Data privacy is a foundational and cross-cutting principle required for achieving all others in this framework. Surveillance and data collection, sharing, use, and reuse now sit at the foundation of business models across many industries, with more and more companies tracking the behavior of the American public, building individual profiles based on this data, and using this granular-level	U2, U1

		information as input into automated systems that further track, profile, and impact the American public.	
493	31	Additional protections would assure the American public that the automated systems they use are not monitoring their activities, collecting information on their lives, or otherwise surveilling them without context-specific consent or legal authority.	U1
494	33	The American public should be protected via built-in privacy protections, data minimization, use and collection limitations, and transparency, in addition to being entitled to clear mechanisms to control access to and use of their data—including their metadata—in a proactive, informed, and ongoing way. Any automated system collecting, using, sharing, or storing personal data should meet these expectations.	U1, U3
495	40	Notice and Explanation: You should know that an automated system is being used, and understand how and why it contributes to outcomes that impact you. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible. Such notice should be kept up-to-date and people impacted by the system should be notified of significant use case or key functionality changes. You should know how and why an outcome	U1, U3

		impacting you was determined by an automated system, including when the automated system is not the sole input determining the outcome. Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context. Reporting that includes summary information about these automated systems in plain language and assessments of the clarity and quality of the notice and explanations should be made public whenever possible.	
496	41	In order to guard against potential harms, the American public needs to know if an automated system is being used. Clear, brief, and understandable notice is a prerequisite for achieving the other protections in this framework. Likewise, the public is often unable to ascertain how or why an automated system has made a decision or contributed to a particular outcome. The decision-making processes of automated systems tend to be opaque, complex, and, therefore, unaccountable, whether by design or by omission. These factors can make explanations both more challenging and more important, and should not be used as a pretext to avoid explaining important decisions to the people impacted by those choices. In the context of automated systems, clear and valid explanations should be recognized as a baseline requirement.	U1, U3
497	41	While notice and explanation requirements are already in place in some sectors or situations, the American public deserve to know consistently and across sectors if an automated system is being used in a way that impacts their	U1

		rights, opportunities, or access. This knowledge should provide confidence in how the public is being treated, and trust in the validity and reasonable use of automated systems.	
498	43	An automated system should provide demonstrably clear, timely, understandable, and accessible notice of use, and explanations as to how and why a decision was made or an action was taken by the system.	U3
499	46	HUMAN ALTERNATIVES, CONSIDERATION, AND FALLBACK: You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter. You should be able to opt out from automated systems in favor of a human alternative, where appropriate. Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts. In some cases, a human or other alternative may be required by law. You should have access to timely human consideration and remedy by a fallback and escalation process if an automated system fails, it produces an error, or you would like to appeal or contest its impacts on you. Human consideration and fallback should be accessible, equitable, effective, maintained, accompanied by appropriate operator training, and should not impose an unreasonable burden on the public. Automated systems with an intended use within sensitive domains, including, but not limited to, criminal justice, employment, education, and health, should additionally be tailored to the purpose, provide meaningful access for oversight, include training for	U1, U3

		any people interacting with the system, and incorporate human consideration for adverse or high-risk decisions. Reporting that includes a description of these human governance processes and assessment of their timeliness, accessibility, outcomes, and effectiveness should be made public whenever possible.	
500	47	There are many reasons people may prefer not to use an automated system: the system can be flawed and can lead to unintended outcomes; it may reinforce bias or be inaccessible; it may simply be inconvenient or unavailable; or it may replace a paper or manual process to which people had grown accustomed.	U2
501	47	The American public deserves the assurance that, when rights, opportunities, or access are meaningfully at stake and there is a reasonable expectation of an alternative to an automated system, they can conveniently opt out of an automated system and will not be disadvantaged for that choice.	U1
502	47	In addition to being able to opt out and use a human alternative, the American public deserves a human fallback system in the event that an automated system fails or causes harm. No matter how rigorously an automated system is tested, there will always be situations for which the system fails. The American public deserves protection via human review against these outlying or unexpected scenarios. In the case of time-critical systems, the public should not have to wait—immediate human consideration and fallback should be available.	U1, U3

503	47	The American people deserve the reassurance that such procedures are in place to protect their rights, opportunities, and access. People make mistakes, and a human alternative or fallback mechanism will not always have the right answer, but they serve as an important check on the power and validity of automated systems.	U1
504	48	An automated system should provide demonstrably effective mechanisms to opt out in favor of a human alternative, where appropriate, as well as timely human consideration and remedy by a fallback system, with additional human oversight and safeguards for systems used in sensitive domains, and with training and assessment for any human- based portions of the system to ensure effectiveness.	U3
505	49	Automated systems used within sensitive domains, including criminal justice, employment, education, and health, should meet the expectations laid out throughout this framework, especially avoiding capricious, inappropriate, and discriminatory impacts of these technologies.	U1, U3

Document 15: Artificial Intelligence Risk Management Framework

No.	Page	Citation	Category
506	1	AI technologies, however, also pose risks that can negatively impact individuals, groups, organizations, communities, society, the environment, and the planet. Like risks for other types of technology, AI risks can emerge in a variety of ways and can be characterized as long- or short-term, high or low-probability, systemic or localized, and high- or low-impact.	U2

507	1	AI systems, for example, may be trained on data that can change over time, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to understand. AI systems and the contexts in which they are deployed are frequently complex, making it difficult to detect and respond to failures when they occur. AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed.	U2
508	1	AI risk management is a key component of responsible development and use of AI systems. Responsible AI practices can help align the decisions about AI system design, development, and uses with intended aim and values. Core concepts in responsible AI emphasize human centricity, social responsibility, and sustainability.	U1, U3
509	1	Understanding and managing the risks of AI systems will help to enhance trustworthiness, and in turn, cultivate public trust.	U1
510	2	As directed by the National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283), the goal of the AI RMF is to offer a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. The Framework is intended to be voluntary, rights-preserving, non-sector-	U3

		specific, and use-case agnostic, providing flexibility to organizations of all sizes and in all sectors and throughout society to implement the approaches in the Framework.	
511	2, 3	Next, AI risks and trustworthiness are analyzed, outlining the characteristics of trustworthy AI systems, which include valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed.	U1
512	3	It describes four specific functions to help organizations address the risks of AI systems in practice. These functions – GOVERN, MAP, MEASURE, and MANAGE – are broken down further into categories and subcategories. While GOVERN applies to all stages of organizations' AI risk management processes and procedures, the MAP, MEASURE, and MANAGE functions can be applied in AI system-specific contexts and at specific stages of the AI lifecycle.	U3
513	4	AI risk management offers a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, while also providing opportunities to maximize positive impacts. Addressing, documenting, and managing AI risks and potential negative impacts effectively can lead to more trustworthy AI systems.	U1, U2
514	4	In the context of the AI RMF, risk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018).	U2

		When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016). Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet.	
515	4	While risk management processes generally address negative impacts, this Framework offers approaches to minimize anticipated negative impacts of AI systems and identify opportunities to maximize positive impacts. Effectively managing the risk of potential harms could lead to more trustworthy AI systems and unleash potential benefits to people (individuals, communities, and society), organizations, and systems/ecosystems.	U1, U2
516	5	Risks related to third-party software, hardware, and data: Third-party data or systems can accelerate research and development and facilitate technology transition. They also may complicate risk measurement. Risk can emerge both from third-party data, software or hardware itself and how it is used. Risk metrics or methodologies used by the organization developing the AI system may not align with the risk metrics or methodologies uses by the organization deploying or operating the system. Also, the organization developing the AI system may not be transparent about the risk metrics or methodologies it used. Risk measurement and management can be complicated by how customers use or integrate thirdparty data or systems into AI products or services, particularly without sufficient internal governance structures and technical safeguards. Regardless, all parties	U2

		and AI actors should manage risk in the AI systems they develop, deploy, or use as standalone or integrated components.	
517	6	Availability of reliable metrics: The current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases, is an AI risk measurement challenge. Potential pitfalls when seeking to measure negative risk or harms include the reality that development of metrics is often an institutional endeavor and may inadvertently reflect factors unrelated to the underlying impact. In addition, measurement approaches can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts.	U2
518	6	Inscrutability: Inscrutable AI systems can complicate risk measurement. Inscrutability can be a result of the opaque nature of AI systems (limited explainability or interpretability), lack of transparency or documentation in AI system development or deployment, or inherent uncertainties in AI systems.	U2
519	7	While the AI RMF can be used to prioritize risk, it does not prescribe risk tolerance. Risk tolerance refers to the organization's or AI actor's (see Appendix A) readiness to bear the risk in order to achieve its objectives. Risk tolerance can be influenced by legal or regulatory requirements (Adapted from: ISO GUIDE 73).	U4
520	7	Policies and resources should be prioritized based on the assessed risk level and potential impact of an AI system.	U3

521	8	When applying the AI RMF, risks which the organization determines to be highest for the AI systems within a given context of use call for the most urgent prioritization and most thorough risk management process. In cases where an AI system presents unacceptable negative risk levels – such as where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently managed. If an AI system's development, deployment, and use cases are found to be low-risk in a specific context, that may suggest potentially lower prioritization.	U3
522	8	The AI RMF may be utilized along with related guidance and frameworks for managing AI system risks or broader enterprise risks. Some risks related to AI systems are common across other types of software development and deployment. Examples of overlapping risks include: privacy concerns related to the use of underlying data to train AI systems; the energy and environmental implications associated with resource-heavy computing demands; security concerns related to the confidentiality, integrity, and availability of the system and its training and output data; and general security of the underlying software and hardware for AI systems.	U2
523	9	The OECD has developed a framework for classifying AI lifecycle activities according to five key socio-technical dimensions, each with properties relevant for AI policy and governance, including risk management [OECD (2022) OECD Framework for the Classification of AI systems — OECD Digital Economy Papers].	U6

524	12	For AI systems to be trustworthy, they often need to be responsive to a multiplicity of criteria that are of value to interested parties. Approaches which enhance AI trustworthiness can reduce negative AI risks. This Framework articulates the following characteristics of trustworthy AI and offers guidance for addressing them. Characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. Creating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use. While all characteristics are socio-technical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external setting. Neglecting these characteristics can increase the probability and magnitude of negative consequences.	U1
525	12	Addressing AI trustworthiness characteristics individually will not ensure AI system trustworthiness; tradeoffs are usually involved, rarely do all characteristics apply in every setting, and some will be more or less important in any given situation. Ultimately, trustworthiness is a social concept that ranges across a spectrum and is only as strong as its weakest characteristics.	U1
526	13	Trustworthiness characteristics explained in this document influence each other. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate but secure, privacy-enhanced, and transparent systems are all undesirable. A comprehensive approach to	U1, U3

		risk management calls for balancing tradeoffs among the trustworthiness characteristics. It is the joint responsibility of all AI actors to determine whether AI technology is an appropriate or necessary tool for a given context or purpose, and how to use it responsibly. The decision to commission or deploy an AI system should be based on a contextual assessment of trustworthiness characteristics and the relative risks, impacts, costs, and benefits, and informed by a broad set of interested parties.	
527	13	Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000, 2015). Deployment of AI systems which are inaccurate, unreliable, or poorly generalized to data and settings beyond their training creates and increases negative AI risks and reduces trustworthiness.	U4, U2
528	13	Reliability is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723, 2022). Reliability is a goal for overall correctness of AI system operation under the conditions of expected use and over a given period of time, including the entire lifetime of the system.	U4
529	14	Accuracy and robustness contribute to the validity and trustworthiness of AI systems, and can be in tension with one another in AI systems.	U1
530	14	Accuracy is defined by ISO/IEC TS 5723, 2022 as "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." Measures of accuracy should consider computational-centric	U4

		measures (e.g., false positive and false negative rates), human-AI teaming, and demonstrate external validity (generalizable beyond the training conditions). Accuracy measurements should always be paired with clearly defined and realistic test sets – that are representative of conditions of expected use – and details about test methodology; these should be included in associated documentation. Accuracy measurements may include disaggregation of results for different data segments.	
531	14	Robustness or generalizability is defined as the "ability of a system to maintain its level of performance under a variety of circumstances" (Source: ISO/IEC TS 5723, 2022). Robustness is a goal for appropriate system functionality in a broad set of conditions and circumstances, including uses of AI systems not initially anticipated. Robustness requires not only that the system perform exactly as it does under expected uses, but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected setting.	U4
532	14	Measurement of validity, accuracy, robustness, and reliability contribute to trustworthiness and should take into consideration that certain types of failures can cause greater harm. AI risk management efforts should prioritize the minimization of potential negative impacts, and may need to include human intervention in cases where the AI system cannot detect or correct errors.	U1, U3
533	14	AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723, 2022). Safe operation of AI systems is improved through:	U1

		responsible design, development, and deployment practices; clear information to deployers on responsible use of the system; responsible decision-making by deployers and end users; and explanations and documentation of risks based on empirical evidence of incidents.	
534	14	Safety risks that pose a potential risk of serious injury or death call for the most urgent prioritization and most thorough risk management process.	U1
535	15	AI systems, as well as the ecosystems in which they are deployed, may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary (Adapted from: ISO/IEC TS 5723, 2022).	U4
536	15	Security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and goes beyond the provenance of the data to encompass unexpected or adversarial use (or abuse or misuse) of the model or data.	U1
537	15	Trustworthy AI depends upon accountability. Accountability presupposes transparency. Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so. Meaningful transparency provides access to appropriate levels of information based on the	U1

		stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system. By promoting higher levels of understanding, transparency increases confidence in the AI system.	
538	16	A transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system.	U1
539	16	The role of AI actors should be considered when seeking accountability for the outcomes of AI systems. The relationship between risk and accountability associated with AI and technological systems more broadly differs across cultural, legal, sectoral, and societal contexts. When consequences are severe, such as when life and liberty are at stake, AI developers and deployers should consider proportionally and proactively adjusting their transparency and accountability practices. Maintaining organizational practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems.	U1, U3
540	16	Measures to enhance transparency and accountability should also consider the impact of these efforts on the implementing entity, including the level of necessary resources and the need to safeguard proprietary information.	U1
541	16	Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the context of their designed functional purposes. Together, explainability and interpretability assist those operating or overseeing an AI system, as well as users of an AI system,	U4

		to gain deeper insights into the functionality and trustworthiness of the system, including its outputs.	
542	17	Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation).	U4
543	17	Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals.	U1
544	17	Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination.	U1
545	17	Systems in which harmful biases are mitigated are not necessarily fair.	U1
546	18	Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. Systemic bias can be present in AI datasets, the organizational norms,	U4, U2

		practices, and processes across the AI lifecycle, and the broader society that uses AI systems. Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.	
547	18	Bias exists in many forms and can become ingrained in the automated systems that help make decisions about our lives. While bias is not always a negative phenomenon, AI systems can potentially increase the speed and scale of biases and perpetuate and amplify harms to individuals, groups, communities, organizations, and society. Bias is tightly associated with the concepts of transparency as well as fairness in society.	U2
548	42	The AI RMF strives to: Be risk-based, resource-efficient, pro-innovation, and voluntary.	U3

Document 16: OECD Recommendation of the Council on Artificial Intelligence

No.	Page	Citation	Category
549	3	The Recommendation aims to foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI while ensuring respect for human rights and democratic values.	B1, B3
550	3	The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI and calls on AI actors to promote and implement them: inclusive growth, sustainable development and well-being; human-centred values and fairness; transparency and explainability; robustness, security and safety; accountability.	B1, B3, B4
551	3	Alongside benefits, AI also raises challenges for our societies and economies, notably regarding economic shifts and inequalities, competition, transitions in the labour market, and implications for democracy and human rights.	B2
552	3	This work has demonstrated the need to shape a stable policy environment at the international level to foster trust in and adoption of AI in society. Against this background, the OECD Committee on Digital Economy Policy (CDEP) agreed to develop a draft Council Recommendation to promote a humancentric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and applies to all stakeholders.	B3, B5
553	4	Principles for responsible stewardship of trustworthy AI: the first section sets out five complementary principles relevant to all stakeholders: i) inclusive growth, sustainable	B1, B3, B4

		development and well-being; ii) human-centred values and fairness; iii) transparency and explainability; iv) robustness, security and safety; and v) accountability. This section further calls on AI actors to promote and implement these principles according to their roles.	
554	5	However, in order to make the most of these innovative solutions, AI systems need to be designed, developed and deployed in a trustworthy manner, consistent with the Recommendation: they should respect human rights and privacy; be transparent, explainable, robust, secure and safe; and actors involved in their development and use should remain accountable.	B1, B3
555	6	RECOGNISING that trust is a key enabler of digital transformation; that, although the nature of future AI applications and their implications may be hard to foresee, the trustworthiness of AI systems is a key factor for the diffusion and adoption of AI; and that a well-informed whole-of-society public debate is necessary for capturing the beneficial potential of the technology, while limiting the risks associated with it;	B1, B2, B3
556	6	UNDERLINING that certain existing national and international legal, regulatory and policy frameworks already have relevance to AI, including those related to human rights, consumer and personal data protection, intellectual property rights, responsible business conduct, and competition, while noting that the appropriateness of some frameworks may need to be assessed and new approaches developed;	B1, B5, B6
557	6	RECOGNISING that given the rapid development and implementation of AI, there is a need for a stable policy	B1, B3

		environment that promotes a human-centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and that applies to all stakeholders according to their role and the context;	
558	7	Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.	B1, B3
559	7	Human-centred values and fairness: AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, nondiscrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.	B1
560	8	Transparency and explainability: AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art: i. to foster a general understanding of AI systems, ii. to make stakeholders aware of their interactions with AI systems, including in the workplace, iii. to enable those affected by an AI system to understand the outcome, and, iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.	B1, B3, B4

561	8	Robustness, security and safety: a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.	B1, B3, B4
562	8	Accountability: AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.	B1, B3
563	11	Recommendations are adopted by Council and are not legally binding. They represent a political commitment to the principles they contain and entail an expectation that Adherents will do their best to implement them.	B3

Document 17: EU-US Inaugural Joint Statement of the TTT

No.	Page	Citation	Category
564	1	The European Union and the United States reaffirm the TTC's objectives to: coordinate approaches to key global technology, economic, and trade issues; and to deepen transatlantic trade and economic relations, basing policies on shared democratic values.	B5
565	1	We intend to cooperate on the development and deployment of new technologies in ways that reinforce our shared democratic values, including respect for universal human rights, advance our respective efforts to address the climate change crisis, and encourage compatible standards and regulations. We intend to cooperate to effectively address the misuse of technology, to protect our societies from information manipulation and interference, promote secure and sustainable international digital connectivity, and support human rights defenders.	B1, B3, B5
566	2, 3	The European Union and the United States acknowledge that AI technologies yield powerful advances but also can threaten our shared values and fundamental freedoms if they are not developed and deployed responsibly or if they are misused. The European Union and the United States affirm their willingness and intention to develop and implement AI systems that are innovative and trustworthy and that respect universal human rights and shared democratic values.	B1, B2, B3
567	3, 4	The European Union and the United States support the development of technical standards in line with our core values, and recognise the importance of international	B1, B3, B5

		standardisation activities underpinned by core WTO principles.	
568	11	The European Union and the United States acknowledge that AI-enabled technologies have risks associated with them if they are not developed and deployed responsibly or if they are misused.	B2
569	11	The European Union and the United States affirm their willingness and intention to develop and implement trustworthy AI and their commitment to a human-centred approach that reinforces shared democratic values and respects universal human rights, which they have already demonstrated by endorsing the OECD Recommendation on AI. Moreover, the European Union and the United States are founding members of the Global Partnership on Artificial Intelligence, which brings together a coalition of like-minded partners seeking to support and guide the responsible development of AI that is grounded in human rights, inclusion, diversity, innovation, economic growth, and societal benefit.	B1, B3, B5
570	11	The European Union and the United States are committed to working together to ensure that AI serves our societies and economies and that it is used in ways consistent with our common democratic values and human rights. Accordingly, the European Union and the United States are opposed to uses of AI that do not respect this requirement, such as rights-violating systems of social scoring.	B1, B3
571	11	The European Union and the United States have significant concerns that authoritarian governments are piloting social scoring systems with an aim to implement social control at scale. These systems pose threats to fundamental freedoms	B2

		and the rule of law, including through silencing speech, punishing peaceful assembly and other expressive activities, and reinforcing arbitrary or unlawful surveillance systems.	
572	11	The European Union and the United States underline that policy and regulatory measures should be based on, and proportionate to the risks posed by the different uses of AI.	B1, B3
573	11	The United States notes the European Commission's proposal for a risk-based regulatory framework for AI. The framework defines high-risk uses of AI, which are to be subject to a number of requirements. The EU also supports a number of research, innovation and testing projects on trustworthy AI as part of its AI strategy.	B1, B2, B3
574	11	The European Union notes the US government's development of an AI Risk Management Framework, as well as ongoing projects on trustworthy AI as part of the US National AI Initiative.	B1, B2, B3
575	12	We are committed to working together to foster responsible stewardship of trustworthy AI that reflects our shared values and commitment to protecting the rights and dignity of all our citizens. We seek to provide scalable, research-based methods to advance trustworthy approaches to AI that serve all people in responsible, equitable, and beneficial ways.	B1, B3
576	12	The European Union and the United States are committed to the responsible stewardship of trustworthy AI and intend to continue to uphold and implement the OECD Recommendation on Artificial Intelligence. The European Union and the United States seek to develop a mutual understanding on the principles underlining trustworthy and responsible AI.	B1, B3, B5

577	12	The European Union and the United States intend to discuss measurement and evaluation tools and activities to assess the technical requirements for trustworthy AI, concerning, for example, accuracy and bias mitigation.	B3, B4
578	12	The European Union and the United States intend to collaborate on projects furthering the development of trustworthy and responsible AI to explore better use of machine learning and other AI techniques towards desirable impacts. We intend to explore cooperation on AI technologies designed to enhance privacy protections, in full compliance with our respective rules, as well as additional areas of cooperation to be defined through dedicated exchanges.	B3, B5

Document 18: TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management

No.	Page	Citation	Category
587	1	Effective risk management and assessment can help earn and increase trust in the development, deployment, and use of AI systems. Recognizing the power of AI to address the world's challenges, we also acknowledge AI systems entail risk. By minimizing the negative impacts of AI systems on individuals, culture, the economy, societies, and the planet, we can maximize the positive impacts and benefits of AI systems that support the shared values underpinning like-minded democracies. Towards that goal, the U.S.-EU Joint Statement of the Trade and Technology Council (May 2022) expressed an intention to develop a joint roadmap ("Joint	B1, B2, B3, B5

		Roadmap") on evaluation and measurement tools for trustworthy AI and risk management.	
588	1	This Joint Roadmap aims to guide the development of tools, methodologies, and approaches to AI risk management and trustworthy AI by the EU and the United States and to advance our shared interest in supporting international standardization efforts and promoting trustworthy AI on the basis of a shared dedication to democratic values and human rights. The roadmap takes practical steps to advance trustworthy AI and uphold our shared commitment to the Organisation for Economic Co-operation and Development (OECD) Recommendation on AI.	B1, B3, B5
589	1	The United States and EU acknowledge that a risk-based approach and a focus on trustworthy AI systems can provide people with confidence in AI-based solutions, while inspiring enterprises to develop trustworthy AI technologies. This approach supports common values, protects the rights and dignity of people, sustains the planet, and encourages market innovation. Both parties are pursuing risk-based approaches that operationalize these values.	B1, B2, B3
590	1	Both sides apply risk-based approaches that consider the combination of societal and technical factors (socio-technical perspective) to advance trustworthy AI. EU examples are represented in the proposed EU AI Act and the work of the High-Level Expert Group (HLEG) on AI. United States examples can be seen in the National Institute of Standards and Technology (NIST) draft AI Risk Management	B1, B2, B3

		Framework as well as the White House Office of Science and Technology Policy (OSTP) Blueprint for an AI Bill of Rights.	
591	1, 2	While the EU and United States may have different views on regulatory approaches – including allocation of responsibility for risk assessment, possible legal responsibility for the establishment of a risk management system, and the appropriate balance between regulatory and voluntary measures – the EU and United States risk-based approaches recognize that our shared values can guide the advancement of emerging technologies.	B1, B3, B6
592	2	Shared terminologies and taxonomies are essential for operationalizing trustworthy AI and risk management in an interoperable fashion. The activities in this section support the EU's and United States' work on interoperable definitions of key terms such as trustworthy, risk, harm, risk threshold, and socio-technical characteristics such as bias, robustness, safety, interpretability, and security. Developing a shared understanding of basic terms will offer an interoperable taxonomy when developing standards and identifying responsibilities, practices, and policies.	B3, B4, B5
593	2	This work will leverage the global work already done and ongoing (such as within the International Organization for Standardization [ISO], OECD, and Institute of Electrical and Electronics Engineers [IEEE]). It will consider related work by the United States (such as the NIST AI Risk Management Framework and the Blueprint for an AI Bill of Rights) and the EU (such as the EU AI Act, HLEG, and European Standardisation Organisations). The EU and United States affirm the importance of a shared understanding and consistent application of concepts and terminology that	B1, B3, B4, B5

		include, but are not limited to - risk, risk management, risk tolerances, risk perception, and the socio-technical characteristics of trustworthy AI.	
594	3	The EU and United States affirm that AI technologies should be shaped by our shared democratic values and commitment to protecting and respecting human rights. Leadership in standards for AI and emerging technologies should promote safety, security, fairness, non-discrimination, interoperability, innovation, transparency, diverse markets, compatibility, and inclusiveness. Both sides are committed to supporting multi-stakeholder approaches to standards development, and recognize the importance of procedures that advance transparency, openness, fair processes, impartiality, and inclusiveness.	B1, B3
595	3	AI standards that articulate requirements, specifications, test methodologies, or guidelines relating to trustworthy characteristics can help ensure that AI technologies and systems meet critical objectives (e.g., functionality, interoperability) and performance characteristics (e.g., accuracy, reliability, and safety). In contrast, standards that are not fit for purpose, not yet available, not broadly accessible (notably to start-ups and small and medium-sized enterprises), or not designed around valid technological solutions may hamper innovation and the timely development and deployment of trustworthy AI technologies.	B1, B3, B6
596	3	Global leadership, participation, and cooperation on international AI standards will be critical for consistent "rules of the road" that enable market competition, preclude barriers to trade, and allow innovation to flourish. This may	B1, B3, B5

		enable governments to align with an international approach when developing internal policies for safeguarding and advancing respect for human rights and democratic values.	
597	3	As like-minded partners, the EU and United States seek to support and provide leadership in international standardization efforts. This can be achieved by contributing and cooperating on technical AI standards development, currently underway in international standards organizations. These standards impact the design, operation, and evaluation and measurement of trustworthy AI and risk management.	B1, B3, B5
598	5	A tracker of existing and emergent risks and risk categories based on context, use cases, and empirical data on AI incidents, impacts, and harms. A values-based understanding of existing risks serves as a baseline for detecting and analyzing both existing and emergent risks.	B1, B2

Document 19: EU-US 2nd Joint Statement of the TTC

No.	Page	Citation	Category
579	1	The EU-U.S. partnership is a cornerstone of our shared strength, prosperity, andcommitment to freedom, democracy, and respect for human rights.	B1, B5
580	2;3	We intend to accelerate our actions to promote the responsible use of technologies,including by working together on policies, standards and technology governance, to foster the use of critical and emerging technologies in line with democratic values and protection of human rights. We are committed to promoting the responsibility to refrain from the arbitrary or unlawful use of surveillance products or services. We are also committed to promoting respect for	B1, B3, B5, B6

		human rights by businesses, including by highlighting best practices in due diligence, and engaging with civil society and the private sector. The European Union and United States also plan to step up actions against the misuse of technologies as tools of repression and as tools of arbitrary or unlawful surveillance, coercion, and cyber threats. These actions will include building further digital and cyber capacities. We resolve to strengthen our cooperation on protecting human rights defenders online, promoting the open, free, global, interoperable, reliable, and secure Internet, and combatting government-imposed Internet shutdowns.	
581	3	Formation of an Artificial Intelligence ("AI") sub-group to realise our commitment to the responsible stewardship of trustworthy AI and our joint support for the Organisation for Economic Co-operation and Development ("OECD") Recommendation on AI. This sub-working group is working to develop a joint roadmap on evaluation and measurement tools for trustworthy AI and risk management, as well as a common project on privacy-enhancing technologies. We will continue to collaborate on the implementation of the OECD AI principles to further our mutual understanding of how to integrate trustworthy and responsible AI into society. This includes working together to identify and oppose rights-violating systems of social scoring.	B1, B2, B3, B5
582	8	In addition, a dedicated subgroup on Artificial Intelligence ("AI") was established to advance work on specific deliverables, and ensure a coordinated approach on AI given its transversal character across several of the TTC working groups.	B3, B5

583	8, 9	We reaffirm our commitment to collaboration in developing and implementing trustworthy AI through a human-centered approach that reinforces shared democratic values and respects human rights. We are jointly exploring how to implement existing AI principles and related efforts within our respective jurisdictions and policy and regulatory landscapes. Mutual understanding on this topic will help lay the foundation for future cooperation on AI initiatives.	B1, B3, B5
584	9	We maintain that a risk-based approach to AI can enable trustworthy AI systems that enhance innovation, lower barriers to trade, bolster market competition, operationalise common values and protect the human rights and dignity of our citizens. The U.S. National Institute of Standards and Technology ("NIST") has released the first draft of an AI Risk Management Framework based on feedback from industry, academia, and civil society, as well as a special publication on bias in AI. In the European Union, the European Commission proposal for a regulatory framework for AI contains dedicated requirements for AI trustworthiness and AI risk management. The requirements will be supported by harmonised standards developed by European Standardisation Organisations ("ESOs"). The ESOs have already started work related to a risk management and a unified approach to trustworthiness, taking into account relevant international standards. The European Commission, standardisation experts and NIST have initiated cooperation concerning foundational elements related to measurement and evaluation tools, risk management and technical and socio-technical requirements for trustworthy AI.	B1, B2, B3, B4, B5, B6

585	9	We are working towards the development of interoperable approaches for managing AI risks. In conjunction with more trustworthy AI systems; such approaches can enable globally beneficial products and services. We intend to work on interoperable terminology related to technical characteristics such as robustness and accuracy, and on socio-technical characteristics including safety.	B3, B4
586	10	Finally, in the Pittsburgh Statement on AI, we stated our opposition to rights-violating systems of social scoring. The European Commission has commissioned a survey to map the use and forms of social scoring worldwide, which will inform our development of a common understanding on social scoring systems, the risks they may pose, and possible mitigation steps.	B1, B2, B3

Document 20: EU-US 3rd Joint Statement of the TTC

No.	Page	Citation	Category
599	2	To fulfill our commitment on developing and implementing trustworthy AI, the United States and the European Union have issued a first Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management (AI Roadmap) and collected perspectives from relevant stakeholders. This roadmap will inform our approaches to AI risk management and trustworthy AI on both sides of the Atlantic, and advance collaborative approaches in international standards bodies related to AI. In conjunction with this effort, we aim to build a shared repository of metrics for measuring AI trustworthiness and risk management methods, which would support ongoing	B1, B2, B3, B4, B5, B6

		work in other settings such as the OECD and GPAI. Our cooperation will enable trustworthy AI systems that enhance innovation, lower barriers to trade, bolster market competition, operationalise common values, and protect the universal human rights and dignity of our citizens. Recognising the importance of privacy in advancing responsible AI development, the European Union and the United States will work on a pilot project to assess the use of privacy-enhancing technologies and synthetic data in health and medicine, in line with applicable data protection rules.	

Document 21: EU-US 4th Joint Statement of the TTC

No.	Page	Citation	Category
600	1	We are committed to make the most of the potential of emerging technologies, while at the same time limiting the challenges they pose to universal human rights and shared democratic values.	B1, B2, B3
601	2	AI is a transformative technology with great promise for our people, offering opportunities to increase prosperity and equity. But in order to seize the opportunities it presents, we must mitigate its risks. The European Union and the United States reaffirm their commitment to a risk-based approach to AI to advance trustworthy and responsible AI technologies. Cooperating on our approaches is key to promoting responsible AI innovation that respects rights and safety and ensures that AI provides benefits in line with our shared democratic values.	B1, B2, B3, B5

602	2	Recent developments in generative AI highlight the scale of the opportunities and the need to address the associated risks. These developments further highlight the urgency and importance of successful cooperation on AI already taking place under the TTC through the implementation of the Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management, as further outlined below. The European Union and the United States decided to add special emphasis on generative AI, including its opportunities and risks, to the work on the Roadmap. This work will complement the G7 Hiroshima AI process.	B2, B3, B5
603	2	The groups have (i) issued a list of 65 key AI terms essential to understanding risk-based approaches to AI, along with their EU and U.S. interpretations and shared EU-US definitions; and (ii) mapped the respective involvement of the European Union and the United States in standardisation activities with the goal of identifying relevant AI-related standards of mutual interest. Going forward, we will continue to consult and be informed by industry, civil society, and academia. We intend to expand shared AI terms, continue our progress towards advancing AI standards and tools for AI risk management, and develop a catalogue of existing and emergent risks, including an understanding of the challenges posed by generative AI.	B2, B3, B4, B5

Document 22: EU-US Terminology and Taxonomy for AI

No.	Page	Citation	Category
604	1	The European Union (EU) and the United States (U.S.) are committed to cooperating on technologies and a digital transformation based on shared democratic values.	B5
605	1	As policy frameworks on AI emerge both in the EU and in the U.S., as well as in many other like-minded countries worldwide, the importance of aligning terminology and conceptual frameworks is becoming increasingly evident. Converging, interoperable approaches to defining and framing AI risks and trustworthiness are essential to enhance legal certainty, promote effective risk management, speed up the identification of emerging risks and reduce compliance costs and administrative burdens. This, in turn, is expected to foster innovation, maximising the benefits of AI systems and at the same time managing its risks. Ultimately the alignment of terminologies will help foster the EU-U.S. joint leadership in the development of an international standard for Trustworthy AI based on a mutual respect for human rights and democratic values.	B2, B3, B5
606	1	The identified terms reflect a shared technical, socio-technical and values-based understanding of AI systems between the EU and U.S. and will serve as a foundation for future definitions, as well as future transatlantic cooperation on AI terminology and taxonomy. This list should be considered as preliminary, to be further expanded and validated also with input from experts and stakeholders in the coming months.	B5

607	1	The EU and U.S. understanding is based on the term "Trustworthy AI." According to the EU HLEG Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. According to the NIST AI Risk Management Framework (AI RMF), characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy enhanced, and fair with their harmful biases managed. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle.	B1, B2, B4
608	1;2	The EU and U.S. agree on the pursuit of a human-centric approach to AI: this requires that the terminology adopted to implement our shared approach to AI centres human, societal and environmental well-being, as well as the rule of law, human rights, democratic values and sustainable development.	B1, B3
609	7	Accuracy: Closeness of computations or estimates to the exact or true values that the statistics were intended to measure. The goal of an AI model is to learn patterns that generalise well for unseen data. It is important to check if a trained AI model is performing well on unseen examples that have not been used for training the model. To do this, the model is used to predict the answer on the test dataset and	B4

		then the predicted target is compared to the actual answer. The concept of accuracy is used to evaluate the predictive capability of the AI model. Informally, accuracy is the fraction of predictions the model got right. A number of metrics are used in machine learning (ML) to measure the predictive accuracy of a model. The choice of the accuracy metric to be used depends on the ML task.	
610	8;9	human values for AI: Values are idealised qualities or conditions in the world that people find good. AI systems are not value-neutral. The incorporation of human values into AI systems requires that we identify whether, how and what we want AI to mean in our societies. It implies deciding on ethical principles, governance policies, incentives, and regulations. And it also implies that we are aware of differences in interests and aims behind AI systems developed by others according to other cultures and principles. The EU and U.S. are committed to the development of Trustworthy AI systems based on shared democratic values including the respect for the rule of law and human rights.	B1, B4, B5
611	9	human-centric AI: An approach to AI that prioritises human ethical responsibility, dynamic qualities, understanding and meaning. It encourages the empowerment of humans in design, use and implementation of AI systems. Human-Centric AI systems are built on the recognition of a meaningful human-technology interaction. They are designed as components of socio-technical environments in which humans assume meaningful agency. Human-Centric AI is not designed as an end in itself, but as tools to serve people with the ultimate aim of increasing human and	B1, B4

		environmental well-being with respect for the rule of law, human rights, democratic values and sustainable development.	
612	10	Auditability: Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and Intellectual Property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enable the system's auditability.	B4, B3
613	11	Accessibility: Extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (which includes direct use or use supported by assistive technologies).	B4
614	11	Accountability: Accountability relates to an allocated responsibility. The responsibility can be based on regulation or agreement or through assignment as part of delegation. In a systems context, accountability refers to systems and/or actions that can be traced uniquely to a given entity. In a governance context, accountability refers to the obligation of an individual or organisation to account for its activities, to complete a deliverable or task, to accept the responsibility for those activities, deliverables or tasks, and to disclose the results in a transparent manner.	B4

615	11	AI Bias: Harmful AI bias describes systematic and repeatable errors in AI systems that create unfair outcomes, such as placing. privileged groups at systematic advantage and unprivileged groups at systematic disadvantage. Different types of bias can emerge and interact due to many factors, including but not limited to, human or system decisions and processes across the AI lifecycle. Bias can be present in AI systems resulting from pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; by being used in unanticipated contexts; or by non-representative design specifications.	B2, B4
616	11	Reliability: An AI system is said to be reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested earlier.	B4
617	11	robustness: Robustness of an AI system encompasses both its technical robustness (ability of a system to maintain its level of performance under a variety of circumstances) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur.	B1, B2, B4
618	11	Safety: AI systems should not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered.	B2, B4
619	12	Security: The protection mechanisms, design and maintenance of an AI system and infrastructure's AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms.	B4

620	12	Traceability: Ability to track the journey of a data input through all stages of sampling, labelling, processing and decision making.	B4, B3
621	12	Trustworthy AI: Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Characteristics of Trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle.	B1, B2, B4