Nanoethics (2024) 18:11
https://doi.org/10.1007/s11569-024-00454-9

ORIGINAL RESEARCH PAPER

Normative Challenges of Risk Regulation of Artificial Intelligence

Carsten Orwat · Jascha Bareis · Anja Folberth · Jutta Jahnel · Christian Wadephul

Received: 16 November 2022 / Accepted: 21 May 2024

Abstract Approaches aimed at regulating artificial intelligence (AI) include a particular form of risk regulation, i.e. a risk-based approach. The most prominent example is the European Union's Artificial Intelligence Act (AI Act). This article addresses the challenges for adequate risk regulation that arise primarily from the specific type of risks involved, i.e. risks to the protection of fundamental rights and fundamental societal values. This is mainly due to the normative ambiguity of such rights and societal values when attempts are made to select, interpret, specify or operationalise them for the purposes of risk assessments and risk mitigation. This is exemplified by (1) human dignity, (2) informational self-determination, data protection and privacy, (3) anti-discrimination, fairness and justice, and (4) the common good. Normative ambiguities require normative choices, which are assigned to different actors under the regime of the AI Act. Particularly critical normative choices include selecting normative concepts by which to operationalise and specify risks, aggregating and quantifying risks (including the use of metrics), balancing value conflicts, setting levels of acceptable risks, and standardisation. To ensure that these normative choices do not lack democratic legitimacy and to avoid legal uncertainty, further political processes and scientific debates are suggested.

Keywords Risk regulation · Risk governance · Artificial intelligence · Human rights · Standardisation · Quantification

C. Orwat (✉) · J. Bareis · J. Jahnel · C. Wadephul
Karlsruhe Institute of Technology, Institute for Technology Assessment and Systems Analysis, Karlstrasse 11, 76133 Karlsruhe, Germany
e-mail: orwat@kit.edu

A. Folberth
University of Heidelberg, Institute of Political Science, Heidelberg, Germany

Introduction

Proposals and approaches explicitly aimed at regulating artificial intelligence (AI) have tended to advocate risk governance or risk regulation using a risk-based approach [1–7]. In general, the umbrella term "risk regulation" encompasses various approaches involving "governmental interference with market or social processes to control potential adverse consequences" [3, 8]. The notion of "risk governance" refers to complex configurations of governmental, semi-private or private actors endeavouring to identify, assess and manage or regulate risks [9–11]. Both risk regulation and risk governance aim to identify and provide information about risks to society. They involve deciding how to weigh up risks against benefits, determining the levels of acceptable risks for those affected and defining the measures that should be taken to manage those risks, especially with the goal of minimising them.

Published online: 23 August 2024

The risk-based approach is one of the many types of risk regulation.[1] One of its aims is to adjust or prioritise regulatory activities according to the risk levels attributed to the concerns to be regulated. Above all, it seeks to not overly stifle innovations and to save the resources of the regulatory authority, e.g. by focusing efforts mainly on objects of regulation that are assigned a high risk ([10]:330; [13–17]). This form of risk-based approach is most often proposed for regulating AI applications, and the Artificial Intelligence Act[2] (hereinafter: AI Act) has adopted it as its main regulatory approach. In contrast, another form of risk regulation, namely the rights-based approach, attempts to prevent risks by establishing regulatory rules, such as prescriptions or rights of persons affected, that apply independently of presumed risk levels of the objects of regulation.[3]

Several studies[4] and official statements[5] emphasise the risks to fundamental rights and societal values caused by AI applications, especially when it comes to human rights and fundamental rights. The right to protection of human dignity, the right to life, the right to equal treatment and non-discrimination, the right to free development of personality and autonomy, the right to informational self-determination, data protection and privacy, the right to freedom of expression and information, the right to a fair trial and to an effective remedy, are among the many fundamental rights and values addressed there.

The purpose of this article is to highlight specific challenges entailed in regulating the risks of AI applications. Such challenges arise because of the specific types of risks, i.e. risks to fundamental rights and societal values. This implies that many normative choices and value judgements are involved; these are the main subjects of the following sections. In this paper, the term fundamental rights[6] includes constitutional rights and the fundamental rights of the European Union. Fundamental societal values are those values that are of public interest and considered essential for the functioning of a democratic society, but are not or not fully enshrined as rights in constitutions or the body of human rights. An example of a fundamental societal value is the common good. The following argumentation is also based on the premise that it is primarily the duty of the state to protect fundamental rights (e.g., [33]:38-41; [34]:103; [35]). Thus, the duty to safeguard and guarantee these rights is also one of the main justifications for state actors to engage in risk regulatory activities.

While numerous proposals for regulating AI are under debate (overview in [36]), there are too few studies that consider the specific characteristics of the risks to be regulated.[7] This paper aims to help fill this gap, as the authors believe that it is essential to understand the specific risk characteristics in order to establish and maintain legitimate, effective and efficient risk regulation. To address the manifold risk characteristics, the article adopts an interdisciplinary perspective that encompasses political science, philosophy, law, risk research and technology assessment.

After some light is shed on the characteristics of AI applications as objects of risk regulation (Section "Applications of artificial intelligence as objects of risk regulation"), the European approach to AI regulation (the AI Act) is outlined in Section "Risk regulation in the European Artificial Intelligence Act". Section "Operationalising risks: options and challenges" discusses the need for the risks to fundamental rights and societal values to be interpreted and operationalised given that the AI Act is seen as a regulatory framework that needs to be filled with normative choices. The normative ambiguities in regulating the risks of AI are presented using four

[1] For the difference and relationship between the 'risk-based approach' (or risk-based regulation) as a specific approach and risk regulation as a more general approach, see ([12]:187; [13]:508-510).

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L series, 12 July 2024. ELI: https://data.europa.eu/eli/reg/2024/1689/oj.

[3] For a discussion in the field of data protection, see [17].

[4] In particular the studies of the Council of Europe [3, 18–22], the studies of the High Level Expert Group established by the European Commission [23, 24], the study of the European Union Agency for Fundamental Rights [25], and in Germany, the Data Ethics Commission [6] and the Enquete Kommission Künstliche Intelligence [7]. See also [26–30].

[5] [2, 5, 31, 32].

[6] Human rights are also referred to in specified places.

[7] Exceptions are [37, 38].

examples of fundamental rights and values in Section "Normative ambiguities in selecting and interpreting fundamental rights and values". On the basis of these examples, we generalise and extend the discussion on normative ambiguities in Section "Necessary normative choices in the design of risk regulation" and point to the normative choices needing to be made for legitimate risk regulation. This leads to the question of who should make such normative choices; this is addressed in Section "Problematic distribution of normative risk decisions in the AI Act", which looks at the partly unclear distribution of normative decisions in the AI Act.

Applications of Artificial Intelligence as Objects of Risk Regulation

One of the areas of AI applications that raises considerable societal concern is their use for automated identification and differentiation of persons as part of semi- or fully automated decision-making (ADM). Differentiation is mainly achieved by using algorithms to create categories to classify persons as individuals or groups to these categories, by selecting individuals or groups or by assigning scores, ratings or rankings to them. This is done using criteria that have been identified as relevant based on the discovery of correlations or the recognition of patterns in data sets with the help of machine learning systems. It enables cost-effective, fine-grained targeting of groups or individuals with information about jobs, products or services or with political statements, etc. These practices are often referred to by terms such as "micro-targeting", "personalisation", "individualisation", "customisation" or "psychological targeting". Some AI applications have other features such as real-time inference and dynamic adaptation of decision rules through the continuous analysis of data streams [39]. AI applications for differentiating between persons can affect a significantly larger number of people. Unlike the limited reach of decision-making by a single human decision-maker, these systems can use the same set of decision rules to target all persons affected, with potential adverse effects. This also means that even small biases or errors in data sets or algorithms can have an impact on large sections of the population ([40]:39; [41]:22). Furthermore, with some types of AI applications (e.g., generative AI, predictive policing), feedback loops are created when algorithms are used to make decisions and information about these decisions is subsequently used to further train these algorithms.

The opacity and "black box" phenomenon, i.e. the incomprehensibility and unpredictability of AI applications, is seen as another critical characteristic. This lack of understandability provides rationales for regulation such as the AI Act (explanatory memorandum to the proposed AI Act, p. 30, also in [2, 42]). However, not all algorithms are equally opaque and unpredictable. First, it is important to clarify whether this incomprehensibility exists for developers and deployers or for external actors such as affected persons or oversight bodies. It is also argued that algorithms, even those generated by machine learning, can help to prove discrimination because programmed decision rules can be used as evidence [43]. However, a certain degree of incomprehensibility is inherent to many AI systems on account of the learning processes and large number of variables, the complexity of relations between them (e.g., artificial neural networks) or their constant adaptation to data streams [44, 45]. In principle, it is possible to use software testing and empirical investigations of outcomes to detect biases and discriminations. Thus, the incomprehensibility of certain types of AI is a question not only of technical complexity but also of the efforts and resources allocated to increase comprehensibility. Second, the lack of external comprehensibility may also be due to the legal protection of proprietary software and trade and business secrets [45, 46]. Third, some applications in business and administration that use algorithm-based personalisation or individualisation of information, products, services, etc. can also lead to reduced comparability of individual treatments or outcomes with those of other persons.

All these factors give rise to situations in which the comprehensibility of the algorithmically generated decisions and outcomes, and of their underlying criteria and weighting between them, is decreased, and in which the reasons for certain decisions ultimately become incomprehensible. Affected individuals are as a result less able to detect, prove and contest adverse outcomes such as manipulation or discrimination [47, 48]. Therefore, the regulatory approach taken in the existing legal frameworks for anti-discrimination and for data protection appears to be inadequate in some respects, when it comes to the

risks of AI and automated decision-making. This regulatory approach provides legal support for affected persons only once they have become aware of adverse effects or violations of their rights or freedoms after these have already occurred, and who seek redress and compensation. However, the difficulty of recognising and mitigating the adverse impacts of AI *ex post* is one of the main justifications for *ex ante* prevention in the area of risk regulation.

AI models and systems are increasingly made available and applied in the form of foundation models for implementation in other systems [49] or in the form of AI as a service [50]. Foundation models are described as those models that are trained using a massive amount of data and can be adapted to many downstream tasks, for the most part by fine-tuning them [49]. Today, foundation models tend to be generative AI that is capable of producing content such as text, images, videos or sound, and includes large language models (e.g., ChatGPT, Gemini, LLama, Luminous or models by Mistral). Generative AI entails risks to fundamental rights such as stereotypes and discrimination, deep fakes and disinformation, or privacy violations (e.g., [51–53]).

Generative AI in the form of foundation models is an example of general-purpose AI that can be deployed and used in many applications and contexts or integrated into numerous other systems. In contrast to so-called narrow AI, what distinguishes the risks of general-purpose AI is that they can spread along downstream systems and applications that utilise them, issues concerning the distribution of responsibilities and liabilities for damage arise, and the risks can become increasingly systemic in character. Additionally, providers of general-purpose AI can become to some extent "system-relevant actors" on which other providers of services or products depend.

Risk Regulation in the European Artificial Intelligence Act

The AI Act contains various instruments for regulating risks that are classified according to the risk categories defined by the European Commission.[8] (1) AI practices with "unacceptable risks" are prohibited, including certain manipulative, subliminal and deceptive techniques, certain social scoring practices, certain predictive policing systems, and under certain conditions some remote biometric systems, including facial recognition and emotion recognition techniques (Article 5 of the AI Act). (2) In the case of "high-risk" AI applications, the provider of an AI system has to fulfil several requirements, such as establishing and maintaining a quality management system that includes a documented risk management system, applying certain data governance and management practices, creating technical documentation and logging, and managing human oversight, etc. (Chapter III of the AI Act). (3) The AI Act lays down special requirements for general-purpose AI, with additional requirements for general-purpose AI with systemic risks (Chapter V of the AI Act). A distinction is made between general-purpose AI with and without systemic risks according to its capabilities, using criteria defined by the European Commission (Annex XIII of the AI Act), or on the basis of a threshold, measured in floating point operations (FLOPs), of greater than 10^{25} FLOPs for general-purpose AI with systemic risks. (4) For AI systems rated as entailing "limited risks" (e.g., AI system directly interacting with persons like chatbots or generating synthetic content), it is mainly transparency obligations that are required (Article 50 of the AI Act). (5) No binding provisions are stipulated for AI systems with "minimal risk or no risk"; there is only the option of voluntarily applying codes of conduct (Article 95 of the AI Act).

The AI Act follows a hybrid governance approach, in particular for high-risk AI systems, in the sense that it utilises various regulatory approaches and instruments. The conformity assessments carried out by most providers of high-risk systems themselves (internal control) or by private third parties (notified bodies) (Article 43 of the AI Act) are a form of self-assessment and self-certification. Furthermore, the post-market monitoring conducted by providers (Article 72 of the AI Act) and the performance of a fundamental rights impact assessment by certain deployers (Article 27 of the AI Act) are elements of "regulated self-regulation". In addition, the AI Act provides for governmental supervisory and sanctioning measures for national supervisory authorities (as market surveillance authorities) (Chapter IX Sections 3 and 4 of the AI Act) or other competent authorities that

[8] The AI Act focuses on risks to health, safety and fundamental rights including democracy, rule of law, and environmental protection (e.g., Recitals 1 or 5 of the AI Act). In the following, mainly risks to fundamental rights are considered.

supervise and enforce fundamental rights (Article 77 of the AI Act). Under certain circumstances, the AI Act assigns to national supervisory authorities the rights and duties to carry out evaluations, require corrective measures, prohibit AI systems and withdraw or recall them (Article 79 of the AI Act). For general-purpose AI systems, the AI Office plays a supervisory and enforcement role (Articles 75(1) and 88-94 of the AI Act).

Operationalising Risks: Options and Challenges

Usually, the effectiveness of risk regulation is assessed according to whether it actually achieves its protection goals such as the prevention of damage, for which regulators are required to provide accountability ([10]:332-336). Sufficiently unambiguous and concrete criteria or principles to define what constitutes relevant risks are necessary not only to avoid arbitrary assessments, but also to provide legal certainty in risk assessments or compliance tests and to derive the appropriate type of regulatory measures such as tests, approvals, requirements, bans or moratoria. Regulators themselves can use such criteria and principles to avoid objections and legal actions due to allegedly incorrect or inaccurate assessments. Last but not least, clear risk definitions also provide developers, providers and operators with the necessary orientation in the anticipatory and preventive development, design and application of systems and, thus, security in financial investments. However, potential problems associated with the inappropriate selection and prioritisation of protection goals and risks, and with their inappropriate interpretation and operationalisation can, in the worst case, result in a failure to achieve the actual protection goals of regulation. Legal uncertainty can materialise in later legal disputes.

This raises questions about how to arrive at the necessary interpretations, specifications and operationalisations. It is doubtful whether the now numerous ethical guidelines developed mainly by academic research, NGOs or private companies in the context of AI[9] can be used directly for risk operationalisation in risk regulation. Given their non-binding character, they can only serve as inspiration for identifying and interpreting those values affected by AI applications. Some fail to address European or national specificities such as the historical development of certain constitutional rights and their position within a society's value structure. Moreover, they differ greatly in terms of the values and principles they encompass. In some cases, the processes of selecting, justifying and interpreting values and principles lacks transparency and is difficult to reconstruct. The same applies to the underlying normative perspectives or interests. The question arises of whether the requirements of some guidelines actually fall short of existing legal requirements and thus constitute nothing more than "ethics-washing" [57].

Furthermore, the existing secondary legislation framework provides only some points of reference for interpretation and operationalisation. For instance, the General Data Protection Regulation[10] (GDPR) and the German General Equal Treatment Act[11], which due to their regulatory scope are particularly relevant to AI in the national context, are not only considered inadequate to protect against the risks of AI ([3, 18]:21-25, 35; [58]:128-142]). Indeed, they are themselves often compromise texts that serve to weigh up various fundamental rights against the interests of different stakeholders. They do not provide pure interpretations of fundamental rights. Additionally, they frequently leave too much scope in interpretations for an *ex ante* specification of risks because they are often designed for *ex post*, context-dependent judicial decisions that require interpretations by judges (in general, [59]:13). Such legal rules for contextual judicial decisions are difficult to transform into generalised standards for the purposes of *ex ante*-oriented risk regulation and risk management.

[9] Overviews are provided by [54–56].

[10] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1.

[11] General Act on Equal Treatment of 14 August 2006 (Federal Law Gazette I p. 1897), as last amended by Article 8 of the SEPA Accompanying Act of 3 April 2013 (Federal Law Gazette I p. 610), in German: Allgemeines Gleichbehandlungsgesetz vom 14. August 2006 (BGBl. I S. 1897), das zuletzt durch Artikel 8 des SEPA-Begleitgesetz vom 3. April 2013 (BGBl. I S. 610) geändert worden ist.

Rather, the practices and case law relating to the protection of fundamental rights, as well as their broad and well-founded scientific debate and further development, can inform the interpretation of their protection goals, essence and options for specification and operationalisation ([23, 60–62]: 82-83). Although fundamental rights are not always unambiguous in terms of content nor consistently established and accepted in different jurisdictions (e.g., [33]:35-52; [37]; [63]:20), they have undergone a considerable degree of concretisation in detail thanks to the long-term case law in a wide variety of situations and contexts and the extensive elaboration in the international body of fundamental rights. In addition, fundamental rights are binding and guaranteed in many jurisdictions. The way these jurisdictions interpret and operationalise fundamental rights needs to be continuously adapted to keep pace with the rapid socio-technical developments in AI (see "Value conflicts" Section). Likewise, valuable guidance when it comes to identifying and operationalising risks can be found in the comprehensive ethical research in general, ethics in specific sectors and ethical research on AI. That said, the challenge associated with ethical research is to deal with the plurality of its approaches and worldviews. These approaches, legal and ethical, have yet to be linked to the interpretational needs and operationalisations involved in the normative basis of risk regulating AI applications. In other words, normative decisions are required to determine which of the many possible operationalisations are acceptable.

Normative Ambiguities in Selecting and Interpreting Fundamental Rights and Values

The notion of "normative ambiguity" refers to differences in the way risk regulation actors understand and accept the meaningful and legitimate values, concepts, priorities, assumptions or boundaries to be applied in risk appraisals ([9]:77f., 153; [64, 65]). As a result, no clear evaluation criteria can exist. Value concepts for evaluating risks can even conflict with each other. In the following, four examples of fundamental rights and societal values affected by AI, considered from a European and German perspective, are used to show that the operationalisation or concretisation required for risk regulation is challenged by normative ambiguities. Human dignity is considered the most central fundamental right. The rights to informational self-determination, data protection and privacy, as well as to anti-discrimination, equality, justice and fairness are the most discussed rights affected by AI and algorithmic decision-making. The societal value of the common good is considered here, because it goes beyond the protection of individual rights and freedoms.

Human Dignity

AI applications can affect the inviolable right to human dignity ([1]: Article 5; [3]:27-28; [6]:43; [19]:33-34; [23, 66]:10; [25]:60; [67, 68]). Most machine learning applications use data about statistically generated groups of people and about the past to build models for algorithmic decision-making using classifications, scores or predictions. In many cases, they use correlations and constructions of persona types, such as in machine learning methods for predicting credit worthiness, likelihood of recidivism, potential suitability as an employee etc. Individuals are evaluated, judged and differentiated between on the basis of one, several or many numerical values. These data analyses do not take account of individuality of persons and their unique quality. Instead, they sort individuals into "prefabricated" categories of group characteristics based on single or multiple variables. A violation of human dignity can result from persons being treated as mere means, instruments or objects. The so-called "object formula" prohibits the reduction of individuals and their treatment as mere objects to achieve the goals of others, because individuals have an intrinsic moral worth (e.g., [69]).

Although the treatment of individuals as numbers is inherent to information and communication technologies with personal data processing, and also happens during decision-making by humans, a distinctive feature of fully automated decision-making is that no further information about the person affected or their situation tends to be considered. Individuals are assessed and judged on the basis of numerical values derived by processing data about large numbers of people categorised into specific groups. Any judgement of individuals based only on numerical values derived from the processing of data about groupings is a normative choice between efficiency gains on the one hand and recognition of the individual and their uniqueness and dignity on the other. Humans make

decisions based on hermeneutical processes, not the kind of decisions based on machine learning with the use of data [59]. Problematically, data does not speak or justify itself, but needs to be interpreted and normatively justified by humans. This is one of the reasons why algorithmic decisions about differentiating between people are more problematic than human-based decisions. Algorithms cannot give reasons and justifications for the decisions they take, which is important if such decisions are to be contested in instances in which they violate human dignity.

Another phenomenon is the potential opacity of the rules, criteria and their respective weightings that determine decisions in algorithmic decision-making, as discussed above. This touches on a crucial dimension of dignity, namely the consent of persons to the treatment, since those affected may not understand what they are agreeing to [70], especially given the unknown criteria that are used in decision-making. Furthermore, Karen Yeung points out the problem that economic operators of algorithmic differentiations primarily aim to derive economic value from customer relationships rather than considering the actual rationale behind their behaviour ([19]:30f.). Comprehensive personality profiles, "super scores" or social scoring that imply total surveillance of people constitute another potential factor that can lead to violations of human dignity ([6]:43). However, profiles and scores that are transferred and combined between organisations are already part of many business and government practices. Therefore, normative decisions at the tipping points between justified and unjustified comprehensive profiles need to be taken in specific contexts.

However, the controversies surrounding the interpretation of this fundamental right highlight the normative choices and clarifications that a society must make with regard to AI applications. At issue are questions of what we mean when we say that the right to dignity is inviolable and absolute, as enshrined in the international and national body of human and fundamental rights law.[12] One question in particular involves ascertaining which precise specifications follow from this norm in order to determine when it has been violated. Although inviolability is established by the prohibition of certain AI applications in Article 5 of the AI Act, which is justified by the need to protect the right to human dignity (e.g., Recitals 28 and 31 of the AI Act), this issue is still relevant to those AI applications that are exempted from such prohibition or classified as "high risk".

No conceptual features for further operationalising human dignity have yet been developed to the extent that they could provide appropriate criteria for risk assessments or specifications for the design of AI applications. Examples of crucial aspects include specifying the object formula, prohibiting the instrumentalisation and reduction of humans, the conditions under which an affected individual would consent to a certain treatment or under which contempt, humiliation or manipulation would be deemed present, the conditions needed to ensure the necessary respect for an individual's qualities [73] and for people to be able to act in a self-determined manner, for the free personal development, or the possibility of leading an autonomous life [e.g., [74]). The concept of human dignity can also help further specify the role of those people involved in AI-based decision-making who are required for the purposes of human oversight under Article 14 of the AI Act and Article 22 of the GDPR. The concept envisages that humans function not only as supervisors of technical functionality but also consider (additional) information about the actual personalities and situations of those affected, as well as specific concerns relating to the self-determination of their lives; in addition, they provide explanations and justifications for the decisions to the persons affected. The extent to which a fundamental right can be encroached without violating its essence – often defined with reference to the inviolability of human dignity – also needs to be clarified. This could be done, for instance, to distinguish conventional scoring from social scoring and illegitimate comprehensive profiles which fully determine a person's personality without the person affected having any chance to influence or contest this. Additionally, operationalising human dignity can be helpful when it comes to identifying further AI technologies to be prohibited (approaches can be found in [75, 76]).

Secondary law does not provide sufficient guidance for operationalising human dignity. In data protection

[12] In particular, Article 1 of the Charter of Fundamental Rights of the European Union (hereinafter Charter) and Article 1(1) of the German Basic Law. Internationally, however, and also in Europe, human dignity is institutionalised differently in constitutions, is specified in different ways and to varying degrees, and occupies different positions in the value structures of the respective societies [71, 72].

law, particularly in Article 22 of the GDPR, the original intention in prohibiting exclusively automated decision-making was to protect people from being treated as mere objects of an assessment of personal data based solely on algorithms (e.g., [77]: para. 3). Due to several exemptions in the GDPR, Article 22 cannot be seen as a general prohibition. On the contrary, it serves rather to regulate the conditions under which fully automated decision-making is allowed. Unclear provisions regarding the conditions under which a human is and must be involved in decision-making or regarding the data controller's obligation to explain the "logic involved" ([41]:77-82) raise doubts about whether, overall, human dignity is sufficiently protected by the provisions of the GDPR.

The specific operationalisation of the right to dignity should therefore be made through *ex ante* definitions of the conditions under which AI-based automated decision-making should be prohibited. This could be particularly relevant in cases where the various criteria used by automated decision-making and the weighting relationships between these criteria cannot be explained by the deployer or user to the affected persons, or where specific goods or positions that are essential for a decent life and for the free development of personality are the objects of automated decision-making.

Informational Self-Determination, Data Protection and Privacy

AI techniques can continue and accelerate trends of data mining, big data analytics, predictive analytics, profiling and targeted advertising that are considered problematic in terms of informational self-determination, data protection and privacy. AI techniques increase the potential for data aggregation and data repurposing, the ability to de-anonymise and re-identify persons even from non-personal or anonymised data, and the ability to assess, categorise, rank or score persons. AI is capable, in particular, of making automated inferences about a person's identity, personality characteristics, or other sensitive facts even from apparently innocuous or mundane forms of data (e.g., communication in social networks), and of providing algorithms and models in automated decision-making that affect a person's ability to self-develop and lead a decent life or the prerequisites for achieving this. AI-based systems and devices are increasingly intruding into people's most intimate sphere (e.g., AI-based personal assistants). Since AI-based biometric identification technologies, including facial or emotion recognition systems, exhibit many of these capabilities, they are particularly problematic (e.g., [19, 25, 26]).

Although the fundamental rights to data protection and privacy and to informational self-determination have been established for decades, a certain degree of normative ambiguity still remains. For one thing, these fundamental rights are persistently interpreted in different ways in order to make them adaptable to new socio-technological developments. At the European level, the rights to the protection of personal data and to privacy are enshrined above all in Articles 7 and 8 of the Charter, Article 16(1) of the Treaty on the Functioning of the European Union (TFEU), Article 8 of the European Convention on Human Rights (ECHR) and the Council of Europe's Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Conventions 108 and 108+). The rights are further developed by the Court of Justice of the EU (CJEU) and the European Court of Human Rights (ECtHR) through case law.[13] At the German national level, for instance, the right to informational self-determination has been developed to protect Article 2(1) on the free development of personality in conjunction with Article 1(1) on human dignity of the Basic Law.[14] The right to informational

[13] De Terwangne [78] states that Conventions 108 and 108+, and the case law of the European Court of Human Rights (ECtHR), have defined data protection as the right to informational self-determination with the right to control. As far as the rights to privacy and data protection are concerned, Brkan [79] shows that the case law of the Court of Justice of the European Union (CJEU) does not yet constitute a clear normative framework for the essence of both rights. Fuster and Gutwirth [80] point to contrasting interpretations of the fundamental right to data protection of the Charter with respect to the interpretation of informational self-determination as either a prohibitive or a permissive (or regulatory) notion.

[14] For AI, protection of the right to informational self-determination is demanded by Germany's Federal Government ([32]:10, 16, 29 etc.) and the Council of Europe - European Committee of Ministers ([5]:Appendix Point B.2.1). This right has been established and developed by a series of decisions by the German Federal Constitutional Court, such as BVerfGE 65, 1 ('Volkszählung', judgement of 15 Dec 1983), BVerfGE 113, 29 ('Anwaltsdaten', judgement of 12 Apr 2005), BVerfGE 115, 320 ('Rasterverhandung II', judgement of 4 Apr 2006), and BVerfGE 118, 168 ('Kontostammdaten', judgement of 16 Jun 2007).

self-determination has often been taken to mean that the individuals or data subjects affected by data collection and processing should have control over their personal data. This fundamental right is enshrined in data protection law, specifically in the GDPR, for the most part by ensuring that the individuals concerned have the right to be informed, to be asked for consent, to rectify or erase data, to restrict data processing, or to object to fully automated decisions (Articles 12-22 of the GDPR). However, the GDPR itself gives rise to contradictions, because it allows data controllers, under certain circumstances, to process personal data for the purposes of their legitimate interests without the consent of data subjects (Article 6(1)f of the GDPR).

Besides this interpretation of data control, another related interpretation of the right to informational self-determination more directly addresses the protection of human dignity and the free development of personality (see Decision by the German Federal Constitutional Court, BVerfGE 65,1 pp. 42ff.) [81, 82]. This interpretation calls for predominantly self-determined options for action to be preserved in informational contexts, and for the free formation of identity, which should still be perceived for the most part as one's own. Furthermore, it demands that the uninhibited use of digital services and products be ensured and that the chilling effects, which can be caused by uncertainties, be avoided. Such uncertainties arise from data processing and its consequences, which are no longer traceable or understandable for the persons concerned (Decision BVerfGE 65,1 pp. 42ff.). Among other things, the right to informational self-determination in this interpretation should prevent impediments to political participation and freedom of expression. In particular, this interpretation could justify tighter regulation of algorithm- and data-based decisions or automated decision-making. Risk assessments in regulatory approaches would then need to address not only the possible loss of control over personal data, but also the possible violations of human dignity and restrictions of the scope to free self-determination, of the right to self-representation, of the ability of those affected to influence decisions, and of the ability to form one's own identity. These aspects would need to be concretized with specific criteria and principles.

In general, there are many different (theoretical) concepts of (informational) privacy (e.g., [13, 83–85]:537-539), and the rights to informational self-determination, data protection and privacy have also been developed on the basis of some of these concepts. One consequence of this diversity of concepts is that there is no clearly delineated and generally accepted list of adverse impacts or risks to the rights of informational self-determination, data protection and privacy.[15] The adverse impacts comprise various forms of restrictions of freedoms, including the infringements of the above-mentioned freedoms of action and self-determination in personality development and identity formation, as well as adverse impacts caused by chilling effects. In addition, since data protection also serves to prevent discrimination, the potential violating impacts include unjustified unequal treatment and the unjust attribution of individual characteristics to persons. Associated with this are adverse impacts such as the violation of human dignity, stigmatisation, stereotyping, damage to reputation, abuse of information power and structural superiority. Other adverse impacts include conformity pressure through surveillance, identity theft, a failure to meet confidentiality expectations or persistent risks due to the permanence of data storage and data processing (e.g., [13, 86, 87]:537).

Another consequence of this conceptual diversity is that attempts to quantify these fundamental rights have different underlying worldviews about what data protection and privacy are for and how to ensure them. In particular, recent approaches that involve quantifying and measuring privacy by means of privacy metrics (e.g., [88]) such as k-anonymity or differential privacy are based on a narrow understanding of privacy as anonymity, secrecy or the confidentiality of systems. This is an inappropriate simplification given that the above-mentioned protective goals of the rights to informational self-determination and data protection go far beyond this narrow interpretation of privacy [89].

[15] The wide range of adverse impacts is also illustrated by enumerations in Recitals 75 and 85 of the GDPR, though they are not completely or exhaustively defined there. The open nature of the range of impacts means that new types of adverse impact brought about by socio-technological developments can potentially be included, though it does leave a certain degree of uncertainty when it comes to the *ex ante* determination of risk factors that need to be considered in risk assessment and risk management.

Anti-Discrimination, Fairness and Justice

A third example of normative ambiguity relates to the concepts of anti-discrimination, fairness and justice.[16] According to the existent anti-discrimination law, AI applications entail a variety of discrimination risks. The main causes of these risks include cognitive bias, technical or organisational errors and subjectivity in decisions about the development, adaptation or use of the systems, including the use of biased or non-representative data sets for machine learning, inappropriate data labelling and the inappropriate selection of AI algorithms or parameters. These can lead to unjustified unequal treatment on the basis of attributes protected under anti-discrimination law (e.g., gender, age, ethnic origin or religion, also known as grounds of prohibited discrimination) or of proxies related to these protected attributes [18, 41, 47, 90–92].

When it comes to risk assessments, anti-discrimination laws such as the German General Equal Treatment Act can only provide a rough basis and scope for concretisation, since the Act itself leaves room for interpretation and ambiguity ([41]:74-76). It is in general unclear whether such context-independent rules for less discriminatory algorithm-based differentiations can be defined at a general level at all (see also [48]). The principle of necessity and proportionality that underlies anti-discrimination law implies rather that context-specific balancing decisions need to be made about the legitimacy of a differentiation, taking into account factors such as the purposes and contexts of the various types of differentiation, the legitimate aims, the necessity and appropriateness of the differentiation or the availability of less adverse alternatives. This makes it difficult to develop AI systems *ex ante* with a lower discrimination risk, especially those AI models and systems that should be applied to different contexts such as foundation models or general-purpose AI.

Researchers and developers strive to define and quantify the discriminatory risks of machine learning algorithms using so-called fairness definitions or fairness metrics and to develop fairness measures in order to debias datasets and algorithms (overviews in [92–95]). However, notions of how fairness should be defined are predicated on highly different normative worldviews, assumptions and ideas of justice and fairness [48, 96–100]. Some fairness metrics depict relations between error rates, including false positives and false negatives, and the rates of true positives and true negatives.[17] Several normative decisions are necessary in the quest to apply fairness metrics or fairness definitions and to design less discriminatory algorithms. Widely discussed trade-offs include the fairness-accuracy trade-off [101, 102] and the decision between group fairness and individual fairness [103]. Additionally, new inequalities may result if fairness measures are applied such as those based on strategies of blinding (i.e. not using legally protected characteristics), equalising decision rates or equalising error rates [104]. Despite these fairness measures, there is still a risk of inflicting injustice, meaning that some of the definitions and measurements of fairness must be understood as expressions of residual risks (i.e. those risks that remain after preventive or risk mitigating measures have been taken).

Furthermore, the use of such fairness metrics and fairness measures already takes for granted the application of AI, algorithmic decision-making and the use of decision criteria derived from generalisations and algorithmic inferences. However, the legitimacy of those applications themselves may be questionable. Anti-discrimination law is not only based on the constitutional rights to justice and equal treatment, but also upholds human dignity and the free development of personality by avoiding stigmatisation and the unjustified attribution of characteristics to persons and their negative consequences for those affected. Many fairness definitions are aimed at the fair treatment of groups and do not consider individual justice, individualised justice or justice on the basis of case-by-case judgements ("Einzelfallgerechtigkeit"). Individual justice takes account of individuals and their personal qualities and life situations (e.g., based on personal interviews); this is usually abandoned when fully automated decision-making is used. Automated

[16] At the European fundamental rights level, the rights to equal treatment and non-discrimination are enshrined in Articles 20, 21 and 23 of the Charter, Article 2 of the Treaty on European Union (TEU), Articles 8 and 10 of the TFEU and Article 14 of the ECHR and Protocol 12.

[17] For instance, the fairness metrics 'equalised odds', 'demographic parity' (or statistical parity), 'equal opportunity' and 'treatment equality' are different ratios built upon rates of false positives, false negatives, true positives or true negatives [92, 93].

decision-making and AI-based decision rules are usually based on generalisations made from data about groups or entire populations and the use of detected patterns ([87, 103]:519).[18] The problem is also demonstrated by one of the first judicial discrimination cases in Europe, in which the court not only prohibited the exclusive use of protected attributes in automated credit decisions. It also emphasised the inadequacy of using statistical figures based on data collected about other individuals to assess the creditworthiness of a particular individual (Decision of the National Non-Discrimination and Equality Tribunal of Finland [105]).

Linking this perspective to the wider debate about justice, allowing the use of AI and automated decision-making systems in specific contexts is already a normative decision per se. Normative choices also mean determining for specific contexts the rules, criteria or parameters that may or may not be legitimately used in automated decisions and for which purposes, how persons affected can communicate their views or which options they have when it comes to challenging and redressing automated decisions.

If one wishes to operationalise the risks of AI applications to social justice, equity and equality, a number of normative issues would need to be decided beforehand by society itself given that the use of algorithm-based differentiation inevitably has an impact on justice, equity and equality. Several different concepts of justice and equality exist, however.[19] This reflects not only the plurality of and controversies in contemporary societies in this respect but also the fact that different concepts are prevalent in different areas of societies. Therefore, it is necessary to identify the contexts in which these fundamental rights are affected and how, and which respective concept of justice should be applied in a particular context.

Specifically, concepts of justice and equality differ in terms of their objects of consideration, e.g., respect for equal moral worth and dignity and valuing individuals as equals, equality of rights and duties, equality in the distribution of welfare, resources and opportunities, equality of human capabilities or of political and democratic status. They also differ in terms of the societal conditions that allow inequalities to be recognised and overcome with a view to achieving social solidarity and respect among the members of that society. One of the most controversial political issues concerns the conditions under which possible deviations from equality are justified and whom this might affect. Furthermore, specific concepts are based on different operationalisation approaches, such as emphasising procedural justice (e.g., due process, right to remedy, presumption of innocence) or distributive justice with a focus on decision-making outcomes. The concepts also differ with respect to what they imply for state interventions based on equality rationales, such as compensating for disadvantage (and maintaining social structures), ensuring opportunities for a decent life or eradicating social structures that are seen as unjust or oppressive ([58]:219-226; [96, 97, 106]).

The Common Good

A fourth and final example of a normative perspective that lacks clarity is the common good. Recent political strategy documents emphasise that AI applications should not only be viewed from the perspective of the individual, but should also contribute to the common good ([23]:4f.; [32]:7, 9, 10, 45, [47]; 2:2). However, the concept of the common good is one of the most contested and vague. In the liberal tradition, the common good embodies a shared standpoint for practical reasoning among the members of society that urges them to engage in "a way of thinking and acting that constitutes the appropriate form of mutual concern" ([107]:Section 4.1). This concern raises awareness of the need for certain inclusive facilities, institutions, collective or public goods that the community has an obligation to maintain, such as public education, local transport, health care or energy infrastructure. In this context, the strength of the concept of the common good lies in the recognition that individual rights, such as the free development of personality, are also in a dependent relationship with this communal realm (see also [108]:487). Collective goods have an enabling function in the sense that they

[18] This discussion of the risk of injustice from AI-based applications of algorithmic decisions extends and intensifies the discussion on statistical discrimination (e.g., [96]).

[19] The complexity of the justice discourse is reflected in the wide variety of justice theories and concepts, ranging from liberal and communitarian traditions to different (cap)ability approaches and more specific discussions about equality of opportunity, gender equality or intergenerational equity, to list but a few of the social science approaches.

encourage individuals to become self-determined and autonomous selves.

In the context of AI, it is still unclear how this broad concept of the common good could be translated into concrete principles, criteria for risk assessments or concepts of systemic risks. There are a number of possible though not exhaustive concretisations: (1) These include adverse outcomes for entire groups, or at the collective level, that are not covered by the statutory data protection and anti-discrimination framework, which is mainly focused on the individual (e.g., [109]). Adverse outcomes at the macro level may also result from a loss of trust or confidence, intimidation or chilling effects. These can be caused by abstract uncertainty about the processing of personal data or its transfer to third parties, about belonging to artificially fabricated groupings and about the actual criteria used in algorithm-based decisions and determining their consequences. These uncertainties can prompt certain population groups to withdraw from using digital services and products. This can lead to social segregation or damage processes of democratic decision-making, as digital services and products are increasingly used in community-building and political exchange. (2) Furthermore, as AI applications contribute to the fine-grained differentiation, personalisation and individualisation of services and products, the applications may displace practices based on solidarity or community services, changing the way a broad range of public or collective goods are provided (e.g., health services). (3) New models for using data and AI methods, especially in the form of open data concepts in public administration to support social innovations ([7]:198f., 201, 206f., 215) or as AI-enhanced platforms to optimise public and multimodal transport (ibid. 385f.), are also discussed in terms of their ability to serve the common good. (4) The mitigation of substitutional effects, job losses or surveillance issues caused by AI-based systems in the world of work are likewise debated with regard to the common good (ibid. 140f.). (5) Last but not least, the normative questions of the societal distribution of risks, the distributive effects of feedback loops and the distribution of efficiency gains and other benefits from AI and algorithmic decision-making can be approached from the perspective of the common good. All the above issues require normative choices to be made when establishing risk regulation if the common good is to be addressed as a relevant value for risk assessments and risk mitigations.

Interim Conclusions

The examples described above demonstrate sources of normative ambiguities. First, they may result from a certain degree of interpretative openness about and indeterminacy of abstract fundamental rights. A general openness in interpretation and specification is necessary so that fundamental rights can be applied to different contexts, situations and times, as well as to new socio-technical developments. Second, the secondary laws that exist in terms of data protection and anti-discrimination are often considered inadequate to protect fundamental rights from the risks of AI. The AI Act establishes more of a regulatory framework that necessitates the interpretation and operationalisation of fundamental rights and cannot therefore be used directly as orientation when it comes to specifying fundamental rights. Third, certain fundamental rights and societal values may be endangered by multiple activities that could potentially affect them. Normative ambiguities can result from uncertainties and subjectivity in the selection and weighting of these relevant risk factors. Fourth, for some fundamental societal values such as the common good, not enough research, consolidated knowledge and above all public discourse and consensus with regard to AI applications are available to select appropriate interpretations and specifications for regulating the risks of AI applications.

Necessary Normative Choices in the Design of Risk Regulation

Designing risk regulation, including risk assessment and risk management, usually involves multiple normative choices [110–117]. These range from the definition and interpretation of regulatory protection goals and the determination of what constitutes a risk to the acceptable levels of risk that can be imposed on society. In the following, the numerous normative choices needing to be made when designing risk regulation regimes for AI are identified and discussed, including how the regulator intends the AI Act to resolve normative ambiguities or leave them open.

Choosing Between Risk-Regulatory Approaches

Applying a risk-based approach rather than a principles-based approach is already a normative decision in its own right. Prioritising certain risks while excluding others and selecting specific areas of intervention for risk regulation entails the danger that this could violate the state's guarantee of human rights protection for everyone. This has the consequence that a certain degree of risk of violating fundamental rights is, by this, allowed for the sake of other societal values such as innovation or efficiency gains. This implies that the chosen risk-based approach will not protect everyone equally, especially when cases involving "non-high" but still relevant risk are neglected, thereby leading to the unequal treatment of holders of fundamental rights (see similar [34]:102). Another risk regulation option would be to establish rights, duties, principles or other rules for risk prevention independently of the specific risk levels assessed for AI applications.[20] In the case of the AI Act, this regulatory option has been deliberately rejected by the European Commission, which cites the administrative costs and burdens on regulatees despite itself considering this option to be a more effective way of enforcing existing laws on fundamental rights ([118]:64-85). Science and technology studies have pointed out that narratives, framings and cultural contexts influence the way risk regulation schemes are designed (e.g., [113]). As far as the AI Act is concerned, the narrative of choosing the specific form of risk-based approach reveals the underlying assumption of the legislator that some degree of risk in the form of potential derogations of fundamental rights is worth saving administrative resources, enabling innovation, or enabling profits by gains in efficiency made by the providers or deployers.

Choosing between a horizontal or sector-specific regulatory approach is also a normative decision. The AI Act follows a horizontal approach in attempting to establish the same framework of rules across different sectors[21] and opens up the possibility to fill this framework with context-specific standardisations, yet it is uncertain if standardisation organisations would do this. A cross-sectoral scheme based on single interpretations of fundamental rights (e.g., a unique interpretation of justice or fairness) can negate the possibility of differing conceptions in different areas of society. For instance, a single definition of fairness would mean that one conception of justice would be relevant to all AI applications in different contexts. However, this would negate the idea that different conceptions of justice and fairness can be relevant to specific contexts by taking account of their peculiarities. Over decades of societal development, for example, the concept of equality of opportunity has become more relevant in certain contexts (e.g., work, education), while the concept of equality of legal and political status has gained importance in other areas (e.g., democratic processes) ([96]:6f.). Even in one and the same sector such as medicine, different fairness concepts can be relevant to different purposes, e.g., group-based fairness for hospital management and health policy-making, and individual-based concepts for patient-level decision-making ([119]:172).

Selection, Prioritisation and Operationalisation of Risks

When a risk-based approach is chosen, further normative decisions are taken when certain fundamental rights and societal values or protection goals are selected and prioritised (for risk assessments in general, see [115, 120]). This includes decisions about which risks to include or omit in risk regulation assessments, how to weigh different risk types against activities with adverse effects, and which regulatory measures (bans, required organisational measures, etc.) to link to different risk levels or categories [110]. There is a risk that certain types of risk factors or adverse activities will be ignored, especially when there is a lack of detailed guidance and many diverse risks to certain fundamental rights and societal values. For example, the omission of the media sector as an area involving the high-risk applications

[20] General principles independent of the level of risks were included in the European Parliament's AI Act proposal of 14 June 2023. However, there was uncertainty as to whether the "make their best effort" provision contained therein would have limited the effectiveness of the principles. In the EU AI Act trilogues, the principles were later dropped. For a discussion of the rights-based and risk-based approach in data protection law see [17].

[21] One exception are financial services for which special provisions are made in the AI Act.

defined in Annex III of the AI Act can be criticised in this respect. Additionally, it is questionable whether the value common good (or parts of it) is addressed as a protection goal by the AI Act. The Act does not include for instance any consideration of chilling effects or the societal distribution of risks and benefits.

The above-mentioned examples of normative ambiguities demonstrate that one of the central normative decisions a society has to make is to interpret and operationalise fundamental rights and societal values. These involves choosing between the various different ways of interpreting fundamental rights and values. The decisions also include adapting the concepts to the contexts of AI applications and result in regulatory requirements being operationalised. The examples of the possible implementation of conceptions of justice and fairness or the common good show that these seemingly practical matters of risk regulation actually involve fundamental decisions about how to live together in a society. We argue in the conclusions that such fundamental societal decisions should be made in legitimate democratic processes.

Aggregation, Comparison and Quantification of Risks

Generally speaking, several concepts of risks exist, which involve attempts to aggregate, quantify and measure risks as well as to perform qualitative risk assessments. They are the subject of long-lasting debates in science and risk governance about the appropriateness of using quantitative or qualitative concepts for specific regulatory purposes ([9, 110, 111]:12-45; [112, 117]).

The AI Act also provides for the use of quantitative measurements and metrics (in particular Recital 74, Article 9(8) or Article 13(3)b(ii) of the AI Act). Such metrics may include fairness metrics, privacy metrics, explainability metrics [121], or performance metrics such as accuracy metrics (e.g., [122]:101ff.). In general, aggregating multiple dimensions of risk assessments into a few more easily manageable numbers – or indeed a single number – may motivate regulators, developers, providers or deployers to strive for quantification. Numerical values ideally allow threshold values or quantitative targets to be set for acceptable risk levels or justify the sorting of systems into certain risk categories according to specific threshold values, scores or rankings. Quantitative measures would also facilitate conformity assessments, certification procedures, inspections and the continuous monitoring of the use of AI applications over their lifetime. Even automated checks to determine whether certain prescribed thresholds have been reached are conceivable. Numbers would make it easier to compare the risks resulting from alternative system designs and optimise them in development processes.

Aggregating and quantifying the risks to fundamental rights entails several problems and limitations, however. First, science and technology studies have shown that aggregation and quantification can disguise the underlying normativity of the decision situation and lead to inappropriate simplifications. Aggregation and quantification can create an "aura of objective truth" ([123]:1) and scientific neutrality, but in fact include several subjective assumptions, value judgements and political decisions [110, 113, 123–125]. Furthermore, the regulatory regimes and the problems, which are intended to be governed, are "co-produced" ([126]:422; [127]) meaning that not only the problem influences the design of the regulatory regime, but also the regulatory regime shapes the problem. Developing a measurement or standardised scale shapes the way the world is experienced and co-produces the phenomenon it claims to measure ([123, 124]:12, 28). If a certain metric is established for AI applications, this can give rise at the same time to a specific understanding of dignity, privacy, justice and other fundamental rights and societal values. Focusing on one fairness metric could influence the way justice is interpreted, for instance.

Second, individual AI and ADM applications usually entail risks to several fundamental rights and societal values at once. The example of data protection and privacy shows that this fundamental right can be violated by a multitude of activities. When certain rights, values, impacts or protection goals or parts thereof are selected for possible aggregation, for use as a measurement instrument or for commensuration (i.e. making comparable on the basis of a common measurement), this constitutes a normative decision that requires them to be weighed up against one another. This includes weighting them according to their importance or forming risk categories (e.g., high risk or low risk, or using "traffic light" approaches to

qualitative risk assessment). As exemplified above, selecting one privacy indicator or metric (e.g., differential privacy) could lead to an unacceptable simplification of the multiple protection goals of this fundamental right. Furthermore, the example of fairness metrics demonstrates that individual metrics are each based on their own very different understandings of fairness or justice. Selecting one fairness metric to be applied in risk regulation is a value judgement in favour of one justice concept with far-reaching effects on society as a whole; it also undermines the importance of other value concepts und worldviews. Accepting one or just a few metrics suggests that the understandings of their normativity are less ambiguous than they are.

Third, a further limitation results from the context-dependency of such value judgements [111]. This hinders the application of common quantitative measures to all areas of society. For example, deciding whether to reduce the false positive rate or the false negative rate is a normative decision that usually differs from one area of society to another. While it tends to be more important in medicine for example to avoid false negatives (when a person who does have a specific disease or condition is wrongly diagnosed as not having it), in criminal law it is usually more important to avoid false positives (when a person who is in fact innocent is wrongly convicted) ([116]:124f.).

Fourth, quantitative measures can raise false expectations about the protection of fundamental rights. Metrics often provide a scale or range of numerical values. This may imply that fundamental rights can be partially derogated (i.e. to a certain degree). In fact, fundamental rights have the status of universal moral boundaries from which it is permissible to deviate only in narrowly defined circumstances (see also "Value conflicts" Section). Ranking fundamental rights violations "on a sliding scale from trivial to serious" is therefore problematic ([128]:10). Accepting that a certain percentage of a population will be exposed to a fundamental right risk would mean accepting that this part of the population does not have equal moral worth and does not enjoy legal equality. As mentioned above, many fairness metrics use error rates that have to be viewed as residual risks that can still lead to violations of fundamental rights. In concrete terms, this implies for instance that a certain percentage of people may be discriminated as a result of the application of AI even when the conformity of these systems with the provisions of the AI Act has been (self-)certified.

Fifth, it is notoriously difficult to mathematically formalise, quantitatively estimate, measure or aggregate many risks to fundamental rights, which is why this is often deliberately neglected. This is due to the characteristics of the fundamental rights. Many fundamental rights and societal values are hardly comparable or commensurable. It is virtually impossible for instance to assess violations of human dignity or restrictions of freedoms in quantitative terms, above all because the severity of the restriction of one fundamental right cannot be quantitatively expressed and compared to another. Furthermore, it is inherent to the ethical nature of human dignity that humans cannot be evaluated in quantitative terms and should only be evaluated according to their unique individual qualities. Some human rights, such as the inviolable right to human dignity, should in fact not be subject to derogations or limitations nor balanced against other fundamental rights.

Sixth, the above examples of normative values also indicate that fundamental rights are complex value structures in the sense that they can have supportive or instrumental interrelations among themselves. The importance of human dignity as an underlying value for understanding other fundamental rights, such as the right to informational self-determination, is just one example. To operationalise specific risks, it is necessary to understand the complex value structures with their supportive and derivative relations and acknowledge that the meaning of each fundamental right depends on other fundamental rights. This makes it difficult to operationalise just one individual risk, such as operationalising the right to privacy as constituting mere technical anonymity or the right to equal treatment as being solely about the "fair" distribution of outcomes.

Seventh, creating quantitative measures entails the additional problem that risks that are not quantified may be ignored ([129]:57). Fairness metrics may for example shift the focus of risk assessments and risk management to the composition of data sets and the selection and training of algorithms, possibly neglecting the fact that bias may also occur when the systems are used in interaction with humans. Focusing on inappropriate metrics may lead to neglect of the consideration to not use an AI system at all or to

use it only with specific measures such as a human operator to ensure that specific information about an affected person is taken into account.

Eighth, consolidated knowledge and consensus in scientific and public discourse are needed to establish quantitative measures that will be commonly accepted. The relevant knowledge about the actual causes and extent of the risks of AI applications is asymmetrically distributed in society, i.e. it is to a large extent in the hands of private providers. This is particularly problematic because it makes it difficult to replicate, contest, verify or falsify methods and procedures to measure and estimate the severity of risks since the actual extent of the harm is not known. Furthermore, this knowledge is characterised by a lack of clarity and certainty, especially regarding the evaluative criteria for risk assessments, as outlined in the case of the normative ambiguities above.

Overall, there is still a lack of commonly agreed-upon, inter-subjectively established measurements that have been replicated, contested, proven and validated in scientific and public discourse. It should be clarified which of the various fundamental rights should not be quantified, which can be quantified and what the most appropriate measurements are, considering the many adverse side effects associated with quantification.

Value Conflicts

Generally speaking, protecting one fundamental right or societal value may lead to conflicts with other fundamental rights or societal values. With regard to AI, value conflicts seem to be fairly ubiquitous rather than the exception. One of the most relevant conflicts is the conflict between, on the one hand, the rights to dignity, free development of personality, informational self-determination, privacy and individual justice, and on the other hand the rights to freedom to conduct a business and freedom of contract, and the societal value of pursuing efficiency through automation and differentiating between people.[22] Allowing interference with these fundamental rights, e.g., by accepting certain AI applications with residual risks of errors or inaccuracies in favour of efficiency arguments, is a normative decision about societal value conflicts. Other relevant value conflicts that are the subject of intense discussion are the right to national or public security, which is allegedly improved by the extensive use of AI-based surveillance technologies, versus the rights to informational self-determination, data protection and privacy as well as human dignity, or the conflict between transparency and people's interest in knowing the decision criteria versus the protection of business and trade secrets.

With the exemption of the inviolable right to human dignity and some other rights ([149]:783), many fundamental rights can be weighed up against each other, though only under very restrictive conditions. In jurisprudence on fundamental rights, this is usually done, if at all, according to the principles of legality, necessity and proportionality (in particular according to Article 52 of the Charter). These principles require not only a legal basis for restrictions of fundamental rights but also judgements on the importance, advantageousness and effectiveness of an application or measure in achieving a legitimate objective. The judgements on necessity and proportionality also require an examination of whether the application or measure is in fact necessary for the objective at all, whether the objective is in the public interest, whether any less intrusive alternatives are available, whether the essence of the impaired right is still respected and whether the restriction of fundamental rights is proportionate to the pursued objectives (e.g., [25]:52f.; [38]:9-12; balancing human rights, even against objectives in the communal interest, is also critizised [148]). In particular, limitations of fundamental rights must not destroy the essence of fundamental rights [149].

As far as risk regulation is concerned, such provisions for balancing of conflicting values might be useful to distinguish legitimate from illegitimate risk impositions and acceptable from unacceptable risks. This also demonstrates how important these decisions are given that such comparisons, weighing and decisions on derogations of fundamental rights are normally made by the highest courts or in legislation. The AI Act, however, assigns these proportionality decisions *de facto* to various actors, in particular, private

[22] This trade-off involves the phenomenon of statistical or rational discrimination with protected attributes or proxies being used to evaluate individuals for the sake of efficiency [87, 130].

actors (see Section "Problematic distribution of normative risk decisions in the AI Act").[23] In this context, it should also be noted that the knowledge base for assessing necessity and proportionality is still weak and constantly developing [131]. For example, it has not yet been fully scientifically proven in any respect that using AI for automated decision-making offers benefits such as improvements in efficiency, accuracy and objectivity compared to human decision-making or compared across different systems, methods and application types. Additionally, there has been and is still little discussion as to whether and to what extent the groups that benefit from the applications on the one hand and those that have to bear the risks on the other are falling apart and to what extent this divide is societally acceptable.

Furthermore, recent AI developments have tended to prolong and intensify the impacts of information and communication technology that has been used for decades (e.g., profiling, scoring, data mining or analytics based on personal data). However, developments in AI can change the tipping points at which existing societal agreements on how to balance value conflicts break down because the proportionality of encroaching on one right at the expense of another no longer holds. In the area of risk regulation, this means that the conditions for determining a risk as an illegitimate encroachment on a fundamental right are changing. This requires further political procedures to regularly define the boundaries between legitimate and illegitimate practices based on constantly evolving AI technologies and their capabilities, i.e. to update the balancing decisions ([61]:70).

Unlike previous data analysis systems, for example, AI systems that form part of biometric systems can infer personality traits, such as character and emotional, psychological or social states or conditions, from seemingly innocuous data such as communications on online social networks or from photos, voice or video recordings (emotional AI, sentiment analysis, affective computing, computational personality assessments) (e.g., [41, 132]:15-16). This has considerable implications for the balancing of conflicting values. IT-based surveillance systems in general often already process the data of a disproportionately large number of people in order to identify only a few, implicitly casting suspicion on innocent people. They draw conclusions from a person's appearance about their identity or personality traits and thus generate external images of essential personality elements, which constitutes a serious infringement of personality rights. It can be difficult for those affected to recognise when these are used at places and events or in services or products, meaning that their individual rights to correction or redress can be virtually impossible to enforce. This uncertainty surrounding the use and the consequences for those affected can give rise to significant chilling effects, for example when using public spaces to exercise the right to freedom of opinion. With the implementation of AI, biometric recognition technology for public or private applications can process, analyse and generate extensive personal profiles, can be discriminatory on account of biased training data and higher error rates for certain groups of persons (e.g., marginalised groups), can be misused to illegitimately infer personality traits that are used for subverting personal decision-making, manipulations, dark patterns, or automated personal persuasions ([38, 133]:12; [134]).[24] In this respect, video surveillance with AI-based facial recognition or gesture recognition technologies has, for example, disproportionately greater potential for encroaching fundamental rights than traditional CCTV. As another example, biometric AI systems used in marketing, financial or employment contexts can potentially infer persons' dependencies on products, services or positions such as jobs, credits, or housing, thereby reinforcing the provider's structural dominance over the persons on the demand side and thus changing the conditions of legality originally established by regulatory frameworks ([41]:64f.). These imbalances can be further accelerated by a loss of consumer and citizen choice brought about by a reduced range of analogue alternatives or changes in digital markets that give rise to greater information asymmetries, deficits in the informed consent approach, market concentration

[23] When it comes to determining discrimination, for example, decisions about the acceptable level of bias or thresholds for illegal disparity are traditionally made by national courts or the Court of Justice of the European Union ([48]:26).

[24] Individual-specific content created by generative AI can be (mis-) used if messages based on profiling of psychological traits are automatically sent in an attempt to effectively influence persons in marketing or political campaigns on a massive scale [135].

through network effects or more pronounced power asymmetries.

Acceptable Risk Levels and Types

Determining acceptable risk levels or the acceptable extent of residual risks that decision-makers can impose on affected individuals in the form of threshold values, criteria, and selected types of risk entails normative and political decisions (in general [12]; [9, 110]: 65, 149-156; [136]:28-30). When determining an acceptable risk level in AI risk regulation, the basic decisions are whether fundamental rights should be encroached at all, and if so, for which purposes, in whose interests and under which conditions. One of the objectives of the AI Act is to protect fundamental rights. On the one hand, the AI Act aims to achieve this by prohibiting certain AI applications considered to be unacceptable. On the other hand, the regulatory requirements for high-risk applications mean that certain degrees of potential restrictions of fundamental rights are regarded as acceptable. This gives rise to a range of normative decisions about the practicalities of risk regulation.

First, the setting of acceptable risk levels in other areas of risk regulation, such as environmental protection or occupational safety and health, can be supported by empirical knowledge about cause-effect relationships, dose-response relationships, and threshold levels for environmental inputs or workplace exposures at which certain types of harm can be caused. When it comes to protecting fundamental rights, defining harms and threshold values for acceptable encroachments of fundamental rights involves normative and political decisions. Scientific knowledge, for example about the causes of AI-based discriminations, can give insights into cause-effect relations but does not help with defining the threshold values for acceptable risks.

Second, risk assessments and risk management are never perfect and can rarely achieve a zero-risk situation;[25] nor will they be with the AI Act. A central normative decision about the acceptable risk level involves defining the scope of high-risk applications and thus classifying all non-prohibited and non-high-risk applications as acceptable. This is based on a decision between Type I or Type II errors. The question here is: "should the regulator lean on the side of individuals with the risk of including non-high risk AI systems in the high risk category (cf., Type I error), or should they lean on the side of AI system providers and fail to consider an AI system as being high risk when that should have been the case (cf., Type II error)" ([137]:28] with reference to [12]:188]).[26]

Third, another peculiarity of the AI Act is that the level of (residual) risk will actually depend significantly on the resources, time and efforts devoted to testing, identifying, analysing, assessing and mitigating the risks of AI applications. These decisions will either be regulated by common specifications or standardisation (see below) or will fall within the margin of discretion of providers or deployers. This includes deciding in particular on the level of accuracy required in risk assessments and risk mitigations and on the efforts that regulatees need to make in order to reach this level, such as decisions about sample size, the intensity of testing and simulation or the types of evidence required. Further normative decisions include defining which types of errors (false positives or false negatives) have to be avoided, the appropriate properties of the data sets used ([38]:33f.), the level of risk of adversarial attacks on AI systems or of the risks of misuse and misinterpretation by human operators. Even if society were to agree on certain metrics, defining thresholds, baselines or target values for acceptable risk levels "within" these metrics would involve normative decisions.

Article 9(5)c of the AI Act stipulates that residual risks shall be communicated to the deployer (Article 13 of the AI Act specifies the transparency and information provisions in detail). This means that possible restrictions of fundamental rights will either become a factor in decisions about purchasing, deploying, or using the system, as the information is understood by buyers, deployers or users of the AI application. Or residual risks even become the matter of

[25] Risk-based regulation requires regulators to take risks ([12]:193).

[26] Gellert ([137]:29f.) argues that, according to the precautionary principle underlying other forms of European risk regulation, the regulatory scope should also include non-high-risk applications.

the relationship between the user and those affected and are communicated and dealt with there, although this raises many questions as to whether they are understood there and whether those affected can assert their rights at all. If deployers or users decide in favour of an AI application nevertheless, they will be left with the legal uncertainty of becoming the possible defendant in a later lawsuit, e.g. under anti-discrimination or data protection law. A further scenario could be that the communication of residual risks will not be understood or will be ignored by deployers or users. They would then unwittingly pass on residual risks to persons affected. In such cases, buyers, deployers or users of AI systems would also face legal uncertainty. In this sense, European legislation is missing an opportunity to define the levels for acceptable risks centrally and thereby to establish legal certainty. This can also erode trust in the regime of the AI Act in general and in particular in CE marking, which is supposed to demonstrate conformity with the provisions of the AI Act.

Article 6(3) of the AI Act allows providers to decide for themselves whether their AI systems pose a significant risk to fundamental rights or not, and thus, whether their systems should be in the high risk category of the regulation. Although details of whether an AI system is deemed to entail no significant risk of harm are given in Article 6(3) and Recital 53 of the AI Act and have to be further clarified by guidelines and delegated acts provided by the European Commission (Article 6(5) and (6) of the AI Act), there is a considerable degree of leeway for providers in determining whether the high risk provisions applies to them or not.

The many legal provisions for self-selection, self-assessment and self-certification are one of the disadvantages of hybrid governance approaches that include regulated self-regulation approaches. Self-assessment and self-certification procedures do not provide legal certainty, because non-compliance can be subsequently ascertained and legally prosecuted by market surveillance authorities (Chapter IX of the AI Act), even when this is initiated as a result of complaints by persons with grounds to consider that the provisions of the AI Act (Article 85 of the AI Act) have been infringed or by other authorities whose task is to protect fundamental rights (Article 77(1) of the AI Act). Normative ambiguities concerning the interpretation of fundamental rights and the acceptable levels of residual risks are especially prone to legal uncertainty.

Fourth, some human and fundamental rights and fundamental values are founded on human dignity, such as the rights to free development of personality, non-discrimination, informational self-determination, data protection or privacy, freedom of expression and political participation, concepts of justice that emphasise the capabilities necessary to lead a decent life and respect for and recognition of the least advantaged. In this context, the meaning of human and fundamental rights derives from the protection of human dignity and its incommensurability and inviolability ([4]:para. 16; [23]:10; [138]). Therefore, risk regulation should involve defining those applications and contexts in which fundamental rights may not be violated, trade-offs must be avoided or residual risks are not acceptable. This includes the goal that fundamental rights should ensure a minimum level of treatment to live a life with dignity and that any toleration of residual impacts can lead to a failure to protect human dignity. The consequence for risk regulation is that the human and fundamental rights approach demands that such residual impacts on human rights should be deemed unacceptable and that further remedial measures should be taken ([139]:524). The AI Act achieves this in part by banning certain AI applications (see above). However, the argument on protection human dignity is particularly relevant for some high-risk AI applications in the AI Act. For instance, the AI Act assigns certain AI applications for law enforcement to the high-risk category (Annex III (6) of the AI Act). In order for those systems to be able to assess the risks of offending or re-offending, the *de facto* toleration of error rates when calculating risk scores for offending or reoffending can thus mean tolerating the risks of infringements of the right to human dignity through false incarceration.[27]

Problematic Distribution of Normative Risk Decisions in the AI Act

According to the hybrid governance approach chosen for the AI Act, normative decisions are spread

[27] Other AI systems for risk assessments of criminal offence are prohibited (Article 5(1)d of the AI Act).

among a number of actors. The European Commission establishes and updates the risk categories and the assignment of AI application types to them, and thereby defines systems with unacceptable or acceptable risks. The Act outlines certain criteria for the European Commission to consider when AI systems are assessed and evaluated as "high risk" (Article 7(2) of the AI Act). However, no further guidelines are provided on how to interpret and operationalise the criteria, leaving it unclear how "adverse impact on fundamental rights" should be operationalised or "significant concerns" are to be dealt with (Article 7(2)e of the AI Act). Additionally, the national supervisory agencies (market surveillance authorities) responsible for *ex post* market surveillance and the AI Office responsible for regulating general-purpose AI systems have to make several normative decisions concerning the criteria used to determine when investigative and regulatory actions should be triggered or when the risks of AI applications are so severe that systems should be corrected, prohibited or withdrawn from the market.

Further interpretations and operationalisations of fundamental rights are required, especially in order for providers to be able to conduct risk management and conformity assessments and inform deployers about risks (see above), for deployers of high-risk AI of high-risk AI systems to assess and determine risks and serious incidents (Articles 26(5) of the AI Act), for certain deployers to conduct fundamental rights impact assessments (Article 27 of the AI Act) or for assessing and understanding the risks of original high risk AI systems and components that are integrated by other providers into their systems (Article 25 of the AI Act). For clarifying the responsibilities along the AI value chain, Article 25 of the AI Act provides obligations to provide information to the new providers and to specify by written agreement the information to be provided, the capabilities, technical access and assistance (Article 25(4) of the AI Act). The written obligation negotiated between the two types of providers lead in effect to a form of private ordering with respect to the realisation or restriction of fundamental rights. As Helberger and Diakopoulos emphasise, however, such private ordering suffers from information asymmetries, unequal negotiation powers and incentives to limit liability at the expense of weaker parties ([140]:5).

According to the New Legislative Framework, the AI Act assigns crucial normative decision-making roles not only to standardisation bodies, but also *de facto* to developers, providers, and deployers. As far as providers of high-risk systems are concerned, conformity with the requirements of the AI Act can be assumed if they adhere to the harmonised standards provided by standardisation organisations (Article 40 of the AI Act). The European Commission can mandate European standardisation organisations[28] to develop harmonised standards ([141]:104f.). Where harmonised standards do not exist or the European Commission deems the harmonised standards insufficient, it can adopt common specifications (Article 41 of the AI Act). In the impact assessment accompanying the AI Act proposal, the Commission noted that "technical standards may not be fully compliant with existing EU legislation (e.g., on data protection or non-discrimination)" ([118]:27; with reference to [142]).[29]

Social rule-setting by standardisation bodies is problematic with respect to fundamental rights. First, under EU law, the delegation of rule-setting powers to private standardisation bodies raises issues of constitutionality, mainly because of the lack of judicial oversight and judicial scrutiny ([141]:104-106; [143]:213-214). Second, standardisation bodies are not legitimate representatives of society. Their procedures are often less transparent, suffer from asymmetries among different stakeholder groups in terms of the resources necessary to participate and often lack a systematic inclusion of stakeholders or persons affected [144–146]. Furthermore, they are subject neither to sufficient democratic control nor the same procedural safeguards and options of public scrutiny and debate as legislation is. Third, according to Veale and Zuiderveen Borgesius they lack experience of interpreting and operationalising fundamental rights ([141]:115).

[28] These are the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC), and the European Telecommunications Standards Institute (ETSI).

[29] Christofi and co-authors [142] show that the standardisation of the Privacy Impact Assessment by ISO (ISO/IEC 29134) deviates from the GDPR's provisions concerning the Data Protection Impact Assessment.

In cases where the expected standardisation does not provide concrete specifications of fundamental rights or specifications arrived at by balancing between fundamental rights, or where this is deliberately left to the discretion of the standards' addressees, such normative decisions fall to the providers or deployers of AI systems, or to a certain extent to the notified bodies which certify conformity with the AI Act and standards. These actors are mostly private companies. This means that either private standardisation bodies, providers, deployers, or notified bodies will be making profound normative decisions about the interpretation of fundamental rights, about the actual quality and scope of the risk management system and the quality management system, either considering the specific risks or not considering them (Articles 9 and 17 of the AI Act), about what risks can be "reasonably mitigated"[30], about an "appropriate balance" between risk management measures and "acceptable" levels of risks and residual risks (Articles 9(4) and (5) of the AI Act), or about the "appropriate" levels of accuracy, robustness and cybersecurity required for the design and development of AI systems (Article 15(1) of the AI Act). This also includes normative decisions not only about which of the metrics or measurements should be applied, but also about the baselines or thresholds of the metrics or measurements, such as the actual error rates to be imposed on persons affected.

When standardisation initiatives focus mainly on process, management or procedural standards (e.g., standards for risk management procedures such as documentation) rather than on standards for concrete levels of acceptable risks, benchmarks etc., standard addressees may tend to concentrate primarily on merely complying with standards.[31] In these cases it is doubtful whether providers or deployers will identify, investigate and assess the potential risks to fundamental rights from the perspective of the persons affected. It is less likely for instance that providers or deployers will consider risks to the common good, potential chilling effects or the distribution of risks among population groups. This includes questions relating to the composition of affected populations, such as the potential impacts on already disadvantaged groups or minorities, whether such groups should be afforded particular protection or whether and how social inequalities that may be perpetuated or aggravated by AI applications should be resolved. Providers or deployers who assess themselves can also be expected to look for ways to reduce risks and optimise existing models and systems only as far as is necessary to achieve the level of risk they themselves have determined to be acceptable rather than considering the option of not using models and systems at all. This may apply in particular to the fundamental choice of whether to use algorithmic predictions for automated decision-making or not and, thus, to make decisions based on human case-by-case assessments of the persons affected.

Under the AI Act, it is standardisation bodies, developers, providers, deployers, users or notified bodies – not the courts or legislative bodies – that make far-reaching normative decisions about the realisation or infringement of fundamental rights when they standardise, design, develop, procure or deploy systems, when they make the settings on systems, or when they frame the learning, use and adaptation of systems. This situation is inappropriate because such sensitive decisions that have a bearing on fundamental rights should normally be made by courts or legislative bodies according to the principles of legality, necessity and proportionality ([38]:29-31, 38). According to these principles (see above), proportionality is generally assessed on the basis of whether the encroachment of a fundamental right is justified on the grounds that a conflicting fundamental right arises as the result of a pressing social need[32] or is in the public interest and thus benefits society as a whole (e.g., national security). In contrast, when such normative balancing decisions are made by developers, providers or deployers, this is done by weighing up violations of fundamental rights against economic interests in terms of profits, thus benefiting only their own interests. This makes a huge difference when

[30] Article 9(3) of the AI Act provides that risks that are to be dealt in the risk management system "shall concern only those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information."

[31] For a critique of similar provisions in the GDPR and the problem of 'box-ticking' exercises, see ([142]:144, 153).

[32] For the difference between proportionality tests in the rights-based and risk-based approaches in data protection law, see ([17]:9-24).

it comes to applying the proportionality test and becomes particularly problematic in cases where the essence of fundamental rights is not clear or needs to be adapted to new socio-technological developments. Furthermore, balancing decisions by developers, providers or deployers can lack investment and legal certainty, either because they discover themselves that the aims can be achieved by alternatives that have a less intrusive impact on fundamental rights and thus choose themselves to refrain from development and deployment of less favourable AI systems (as the aforementioned principles would require) or because use of the application may be stopped through later prohibitions by supervisory authorities or court decisions.

This inappropriateness is all the more relevant given the severity of potential violations of fundamental rights and the above-mentioned larger number of people potentially affected by AI applications and automated decision-making. Such normative choices affect broad segments of the population. Systems with a wide reach and with apparently small residual error rates can easily discriminate large segments of the population, for instance. This effect is especially critical in applications where no alternatives exist for the affected persons, such as in public services, with AI applications used by state authorities, or on concentrated markets. Given this wide reach, it would be more appropriate for major normative decisions of risk regulation to take place within inclusive democratic processes that involve those who are actually or potentially affected and their elected policy-makers (see also [66]).

The situation is particularly inappropriate on account of the normative ambiguities in assessment and evaluative criteria that further increase the scope for discretion in risk assessment and risk management and may promote arbitrariness. When it takes the form of meta-regulation, risk governance generally involves delegating risk-regulatory tasks to diverse governance actors, including private actors or regulatees. The appropriateness of this governance option depends, among other things, on whether objective or understandable knowledge is generated by providers that can be verified by others such as the regulator ([126]:420-422). However, for as long as normative ambiguities prevail in AI regulation, with the result that risk assessments are neither objective nor based on scientific and public consensus, such delegation hardly seems suitable. Risk assessment by providers cannot be regarded as objective scientific endeavours but rather as context-dependent subjective balancing of trade-offs according to their private interests.

Furthermore, science and technology studies on risk perception have shown that the interpretation of facts about risks is influenced by contextual information (e.g., [111]:127f.). Thus, the outcomes of risk assessment and risk management depend on factors specific to the various providers of AI systems or to the risk assessors, such as the individual risk appetite, the willingness to incur costs, experience, knowledge about fundamental rights or insights into or attitudes towards the potentially affected population. This leads to incongruent or fragmented levels of protection of fundamental rights. Fundamental rights are universal in the sense that all people have equal fundamental rights, however.

Conclusions

We have highlighted several normative ambiguities in the risk regulation of AI applications. These necessitate normative choices when it comes to establishing and maintaining a risk regulation scheme. These choices involve deciding on the specific form of the risk-based approach itself, selecting and prioritising risks, choosing the concepts that underlie the interpretations and operationalisations of fundamental rights and values, selecting approaches to aggregating or quantifying risks, balancing value conflicts, updating balancing decisions and determining acceptable risk levels. Since normative choices are normally an integral part of devising and operating a risk regulation scheme, it is a question of who makes these decisions and in whose interests.

Decisions about balancing fundamental rights against each other or against other fundamental values play a central role in determining acceptable risk types and levels and constitute one of the sources of normative ambiguities that are usually prone to conflicts. One of the challenges posed by value conflicts in risk regulation is that value conflicts may not be recognised as such or as a normative matter of risk regulation. Risk trade-offs can lead to residual risks that can actually turn out to be unacceptable in terms of protecting human and fundamental rights. Furthermore, trade-offs might be decided on the basis

of an outdated and inadequate balancing consideration that is institutionalised in secondary legislation. For example, the rapid increase in the capabilities of AI systems to process personal data and make inferences about highly personality-relevant attributes necessitates the constant adjustment of balancing processes. The fact that trade-offs are decided by actors who lack democratic legitimacy is another problem. Such trade-offs may also be based on an inadequate formal foundation (e.g., without any legal basis) or take place covertly despite the requirement for procedures to be conducted in public political processes. A further problem under these conditions is that balancing might be distorted in favour of particular private interests, disregarding public interests such as the overall distribution of risks and benefits.

The hybrid governance approach of the proposed AI Act results in a problematic diffusion of normative decision-making across several actors. Crucial normative decisions will be made not only by the European Commission, national supervisory authorities and market surveillance authorities, but also by standardisation organisations, notified bodies, and/or the providers and deployers of AI systems. The core elements of self-assessment and self-certification require providers or deployers either to make normative choices themselves or to rely on the decisions taken by standardisation organisations. Standardisation organisations, notified bodies, providers and deployers lack the legitimacy, options of democratic control, and competence to decide on encroachments of fundamental rights. This is also inappropriate given the society-wide scope of impacts on fundamental rights that AI and ADM applications can potentially have. AI-based ADM systems can easily affect large portions of a population. The risk-based approach in the form chosen for the AI Act can jeopardise the guarantee of fundamental rights for everyone. This can ultimately undermine trust in the regulatory institutions and their authority (see also [147]).

Decisions about weighing fundamental rights up against their limitations are normally made by the highest courts or legislative bodies. Court decisions may also be an insufficient source for specifying fundamental rights and balancing trade-offs as they relate to the specific situations of the case and it is not always possible to translate them to other contexts. Furthermore, in order to bring about court decisions, there must be a plaintiff who files a suit. Any legal clarification and operationalisation of fundamental rights on the basis of court decisions could thus be fragmented and take years. In contrast, the advantage of regulatory rule-setting by legislation is that it offers clear guidance to developers and providers on legitimate development paths and enable legal security of investments.

Normative decisions inherent to risk regulation should be identified as such and delegated to legitimate, democratic political processes supported by knowledge from systematic research and publicly discussed evidence. However, the AI Act does not explicitly establish the rights of consultation and public participation for stakeholders ([38]:48-54), but only enables the options that stakeholders are included in standardisation procedures, for instance. Furthermore, it remains to be seen in this context whether the AI Act provisions on common specifications (Article 41 of the AI Act), the many delegated acts expected to be adopted by the European Commission (Article 97 of the AI Act), or the Guidelines from the Commission on the concretisation of obligations for providers of high risk AI systems, on the distribution of responsibilities when AI components are integrated into other systems, and on prohibited practices, transparency obligations, etc. (Article 96 of the AI Act) will be used and how democratically the related political processes will be shaped.

Research should especially address the context-specific levels of acceptable risks, further unacceptable, unnecessary and disproportionate restrictions on fundamental rights, and the distribution of benefits and risks – above all any disproportionate risks to vulnerable or already disadvantaged groups. Since risks to fundamental rights are imposed on those affected by AI applications, these normative choices should be negotiated in public discourse (in general [34]:103f.; [110]:136f.). This should be done before decisions about risks are distributed among governance actors, including private actors. This should also be done in order to prevent ambiguities and failures, to provide sufficient guidance to regulators and regulatees and, ultimately, to ensure the legal certainty and legitimacy of the regulated AI applications and their impacts on fundamental rights and societal values. Any "hidden privatisation" of decisions about public values ([19]:34f.) and value conflicts must be avoided.

Acknowledgements We would like to thank the two anonymous reviewers as well as Karen Yeung, Ingrid Schneider, Reinhard Heil, Armin Grunwald, Torsten Fleischer, Isabel Kusche, Lucas Staab, Paul Grünke, Catharina Rudschies, Marc Hauer, Christina Timko, Sylke Wintzer and Chris Cave for their valuable feedback and support.

Research involving Human Participants and/or Animals The research did not involve human participants or animals.

Funding Open Access funding enabled and organized by Projekt DEAL. This article is part of the project "Governance of and by Algorithms" funded by the German Federal Ministry for Education and Research (Grant No. 01IS19020B). The funding is gratefully acknowledged.

Declarations

Consent for publication Since no human participants were involved, informed consent was not necessary.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. European Parliament, (2020) Framework of ethical aspects of artificial intelligence, robotics and related technologies. Resolution of 20 (October 2020) with recommendations to the Commission on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies (2020/2012(INL)). European Parliament, Strasbourg
2. European Commission (2020) White paper on artificial intelligence - a European approach to excellence and trust. COM(2020) 65 final. European Commission, Brussels
3. Council of Europe - CAHAI (2020) Feasibility study, CAHAI (2020)23 Ad hoc committee on artificial intelligence. Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Strasbourg
4. Council of Europe - CAHAI (2021) Possible elements of a legal framework on artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law. CM(2021)173-add, 17 Dec 2021. Council of Europe - Ad hoc Committee on Artificial Intelligence (CAHAI), Strasbourg
5. Council of Europe - European Committee of Ministers (2020) Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems. Council of Europe, Strasbourg
6. German Data Ethics Commission (2019) Opinion of the data ethics commission. Data Ethics Commission of the Federal Government; Federal Ministry of the Interior, Building and Community. Federal Ministry of Justice and Consumer Protection, Berlin
7. Enquete-Kommission Künstliche Intelligenz (2020) Bericht der Enquete-Kommission Künstliche Intelligenz - Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale (Drucksache 19/23700, 28.10.2020). Deutscher Bundestag, Berlin
8. Hood C, Rothstein H, Baldwin R (2001) The government of risk: understanding risk regulation regimes. Oxford University Press, Oxford
9. Renn O (2008) Risk governance. Coping with uncertainty in a complex world. Earthscan, London
10. Black J (2010) The role of risk in regulatory processes. In: Baldwin R, Cave M, Lodge M (eds) The Oxford handbook of regulation. Oxford University Press, Oxford, pp 302–348
11. van der Heijden J (2019) Risk governance and risk-based regulation: A review of the international academic literature, State of the art in regulatory governance research paper series. Victoria University of Wellington, Wellington
12. Black J (2010) Risk-based regulation: Choices, practices and lessons being learnt. In: OECD (ed.): Risk and regulatory policy: Improving the governance of risk, OECD reviews of regulatory reform. Organisation for Economic Co-operation and Development (OECD), Paris, pp 185–236
13. Macenaite M (2017) The "riskification" of European data protection law through a two-fold shift. Eur J Risk Regul 8(3):506–540. https://doi.org/10.1017/err.2017.40
14. Hutter BM (2005) The attractions of risk-based regulation: accounting for the emergence of risk ideas in regulation. ESRC Centre for Analysis of Risk and Regulation, London
15. Rothstein H, Irving P, Walden T, Yearsley R (2006) The risks of risk-based regulation: Insights from the environmental policy domain. Environ Int 32(8):1056–1065. https://doi.org/10.1016/j.envint.2006.06.008
16. Black J, Baldwin R (2010) Really responsive risk-based regulation. Law Pol 32(2):181–213. https://doi.org/10.1111/j.1467-9930.2010.00318.x
17. Gellert R (2020) The risk-based approach to data protection. Oxford University Press, Oxford

18. Zuiderveen Borgesius F (2018) Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe, Directorate General of Democracy, Strasbourg
19. Yeung K (2019) Responsibility and AI. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe study DGI(2019)05. Council of Europe, Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT), Strasbourg
20. Helberger N, Eskens S, van Drunen M, Bastian M, Moeller J (2020) Implications of AI-Driven tools in the media for freedom of expression. Council of Europe, Strasbourg
21. Wagner B (2018) Algorithms and Human Rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications. Council of Europe study DGI(2017)12. Council of Europe, Committee of experts on internet intermediaries (MSI-NET), Strasbourg
22. Mantelero A (2019) Artificial intelligence and data protection: challenges and possible remedies. Consultative committee of the convention for the protection of individuals with regard to automatic processing of personal data (Convention 108), Report on Artificial Intelligence. Council of Europe, Directorate General of Human Rights and Rule of Law, Strasbourg
23. AI HLEG (2019) Ethics guidelines for trustworthy AI. Independent High-Level Expert Group on Artificial Intelligence (AI HLEG), report published by the European Commission, Brussels
24. AI HLEG (2019) Policy and investment recommendations for trustworthy AI. Independent High-level Expert Group on Artificial Intelligence (AI HLEG), report published by the European Commission, Brussels
25. FRA (2020) Getting the future right - Artificial intelligence and fundamental rights. European Union Agency for Fundamental Rights (FRA), Luxembourg
26. Access Now (2018) Human Rights in the Age of Artificial Intelligence. Access Now, Brooklyn
27. Latonero M (2018) Governing artificial intelligence: upholding human rights and dignity. Data & Society Research Institute, New York
28. Raso FA, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L (2018) Artificial intelligence & human rights: Opportunities & risks. Berkman Klein Center for Internet & Society, Cambridge
29. Donahoe E, MacDuffee Metzger M (2019) Artificial intelligence and human rights. J Democr 30(2):115–126. https://doi.org/10.1353/jod.2019.0029
30. Mantelero A, Esposito MS (2021) An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Comput Law Secur Rev 41:105561. https://doi.org/10.1016/j.clsr.2021.105561
31. UN (2018) Report on artificial intelligence technologies and implications for freedom of expression and the information environment. Report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye. Report A/73/348. United Nations, OHCHR; Geneva
32. Bundesregierung (2018) Strategie künstliche Intelligenz der Bundesregierung. Bundesregierung; Berlin
33. Nickel JW (2007) Making sense of human rights. Wiley-Blackwell, Malden
34. Shrader-Frechette K (2005) Flawed attacks on contemporary human rights: Laudan, Sunstein, and the cost-benefit state. Hum Rights Rev 7(1):92–110. https://doi.org/10.1007/s12142-005-1004-1
35. Nemitz P (2018) Constitutional democracy and technology in the age of artificial intelligence. Philos Trans A Math Phys Eng Sci 376(2133):1–14. https://doi.org/10.1098/rsta.2018.0089
36. Folberth A, Jahnel J, Bareis J, Orwat C, Wadephul C (2022) Tackling problems, harvesting benefits - A systematic review of the regulatory debate around AI, KIT Scientific Working Papers No. 197. KIT Scientific Press, Karlsruhe
37. Smuha NA (2020) Beyond a human rights-based approach to AI governance: promise, pitfalls, plea. Philos Technol (34)Suppl.iss. 1:91–104. https://doi.org/10.1007/s13347-020-00403-w
38. Smuha NA, Ahmed-Rengers E, Harkens A, Li W, MacLaren J, Piselli R, Yeung K (2021) How the EU can achieve legally trustworthy AI: A response to the European Commission's Proposal for an Artificial Intelligence Act. University of Birmingham, LEADS Lab, Birmingham
39. Yeung K (2019) Why worry about decision-making by machine? In: Yeung K, Lodge M (eds) Algorithmic regulation. Oxford University Press, Oxford, pp 21–48
40. Gandy OH Jr (2010) Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems. Ethics Info Tech 12(1):1–14. https://doi.org/10.1007/s10676-009-9198-6
41. Orwat C (2020) Risks of discrimination through the use of algorithms. A study compiled with a grant from the Federal Anti-Discrimination Agency. Federal Anti-Discrimination Agency, Berlin
42. Europan Commission (2020) Report on the safety and liability implications of artificial intelligence, the internet of things and robotics. European Commission, Brussels
43. Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2018) Discrimination in the age of algorithms. J Leg Anal 10:113–174. https://doi.org/10.1093/jla/laz001
44. Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc 3(1):1–12. https://doi.org/10.1177/2053951715622512
45. Brkan M, Bonnet G (2020) Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: Of black boxes, white boxes and fata morganas. Eur J Risk Regul 11(1):18–50. https://doi.org/10.1017/err.2020.10
46. Kroll JA (2018) The fallacy of inscrutability. Philos Trans A Math Phys Eng Sci 376(2133):1–14. https://doi.org/10.1098/rsta.2018.0084
47. Hacker P (2018) Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. Common Mark Law Rev 55(4): 1143-1185, https://doi.org/10.54648/cola2018095

48. Wachter S, Mittelstadt B, Russell C (2021) Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 41:105567. https://doi.org/10.1016/j.clsr.2021.105567
49. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E (2021) On the opportunities and risks of foundation models. ArXiv, https://doi.org/10.48550/arXiv.2108.07258
50. Cobbe J, Singh J (2021) Artificial intelligence as a service: Legal responsibilities, liabilities, and policy challenges. Comput Law Secur Rev 42:105573. https://doi.org/10.1016/j.clsr.2021.105573
51. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp 610–623. https://doi.org/10.1145/3442188.3445922
52. Weidinger L, Uesato J, Rauh M, Griffin C, Huang P-S, Mellor J, Glaese A, Cheng M, Balle B, Kasirzadeh A (2022) Taxonomy of Risks posed by Language Models. FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp 214–229. https://doi.org/10.1145/3531146.3533088
53. Novelli C, Casolari F, Hacker P, Spedicato G, Floridi L (2024) Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity. Arxiv, https://doi.org/10.48550/arXiv.2401.07348
54. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2
55. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1. Berkman Klein Center for Internet & Society at Harvard University, Cambridge, MA
56. Rudschies C, Schneider I, Simon J (2021) Value Pluralism in the AI Ethics Debate – Different Actors, Different Priorities. Int J Inf Ethics. 29(3):1–15. https://doi.org/10.29173/irie419
57. Wagner B (2018) Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In: Bayamlioğlu, E, et al. (eds.): Being profiled: Cogitas ergo sum. 10 years of 'Profiling the European Citizen'. Amsterdam University Press, Amsterdam, pp 84–88
58. Deutscher Ethikrat (2018) Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Stellungnahme. Deutscher Ethikrat, Berlin
59. Hildebrandt M (2016) Law as information in the era of data-driven agency. Mod Law Rev 79(1):1–30. https://doi.org/10.1111/1468-2230.12165
60. Ruggiu D (2018) Human rights and emerging technologies. Analysis and perspectives in Europe. Pan Stanford, New York
61. Ruggiu D (2019) Models of anticipation within the responsible research and innovation framework: the two RRI approaches and the challenge of human rights. NanoEthics 13(1):53–78. https://doi.org/10.1007/s11569-019-00337-4
62. Yeung K, Howes A, Pogrebna G (2020) AI Governance by human rights-centred design, deliberation and oversight: An end to ethics washing. In: Dubber M, Pasquale F, Das S (eds) The Oxford Handbook of AI Ethics. Oxford University Press, New York, pp 77–106
63. Götzmann N, Vanclay F, Seier F (2016) Social and human rights impact assessments: What can they learn from each other? Impact Assess Proj Apprais 34(1):14–23. https://doi.org/10.1080/14615517.2015.1096036
64. Johansen IL, Rausand M (2015) Ambiguity in risk assessment. Saf Sci 80:243–251. https://doi.org/10.1016/j.ssci.2015.07.028
65. Stirling A (2008) Science, precaution, and the politics of technological risk. Ann NY Acad Sci 1128(1):95–110. https://doi.org/10.1196/annals.1399.011
66. EGE (2018) Statement on artificial intelligence, robotics and 'autonomous' systems, European Group on Ethics in Science and New Technologies (EGE), European Commission Brussels
67. Jones ML (2017) The right to a human in the loop: Political constructions of computer automation and personhood. Soc Stud Sci 47(2):216–239. https://doi.org/10.1177/0306312717699716
68. Kaminski ME (2019) Binary governance: Lessons from the GDPR'S approach to algorithmic accountability. S Cal L Rev 92(6):1529–1616
69. Mahlmann M (2012) Human dignity and autonomy in modern constitutional orders. In: Rosenfeld M, Sajó A (eds) The Oxford handbook of comparative constitutional law. Oxford University Press, Oxford, pp 1–26
70. Schaber P (2012) Menschenwürde. Reclam, Stuttgart
71. McCrudden C (2008) Human dignity and judicial interpretation of human rights. Eur J Int Law 19(4):655–724. https://doi.org/10.1093/ejil/chn059
72. Becchi P, Mathis K (2019) Handbook of human dignity in Europe. Springer International Publishing, Cham
73. Schaber P (2013) Instrumentalisierung und Menschenwürde. Mentis, Münster
74. Düwell M (2017) Human dignity and the ethics and regulation of technology. In: Brownsword R, Scotford E, Yeung K (eds) The Oxford Handbook of Law, Regulation and Technology. Oxford University Press, Oxford, pp 177–196
75. Teo SA (2023) Human dignity and AI: Mapping the contours and utility of human dignity in addressing challenges presented by AI. Law Innov Technol 15(1):1–39. https://doi.org/10.1080/17579961.2023.2184132
76. Orwat C (2024) Algorithmic Discrimination from the Perspective of Human Dignity. Soc Inc 112: Article 7160. https://doi.org/10.17645/si.7160
77. Scholz P (2019) DSGVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling. In: Simitis, S, Hornung, G, Spiecker genannt Döhmann, I (eds.): Datenschutzrecht. DSGVO und BDSG. Nomos, Baden-Baden
78. de Terwangne C (2022) Privacy and data protection in Europe: Council of Europe's Convention 108+ and the European Union's GDPR. In: Fuster, GG, Van Brakel, R, Hert, Pd (eds.): Research handbook on privacy and data

protection law. Edward Elgar, Cheltenham, Northampton, pp 10-35
79. Brkan M (2019) The essence of the fundamental rights to privacy and data protection: Finding the way through the maze of the CJEU's constitutional reasoning. Ger Law J 20(6):864–883. https://doi.org/10.1017/glj.2019.66
80. Fuster GG, Gutwirth S (2013) Opening up personal data protection: A conceptual controversy. Comput Law Secur Rev 29(5):531–539. https://doi.org/10.1016/j.clsr.2013.07.008
81. Britz G (2010) Informationelle Selbstbestimmung zwischen rechtswissenschaftlicher Grundsatzkritik und Beharren des Bundesverfassungsgerichts. In: Hoffmann-Riem W (ed) Offene Rechtswissenschaft. Mohr Siebeck, Tübingen, pp 561–596
82. Rouvroy A, Poullet Y (2009) The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In: Gutwirth S et al (eds) Reinventing data protection? Springer, Amsterdam, pp 45–76
83. Solove DJ (2006) A taxonomy of privacy. U Pa Law Rev 154(3):477–560. https://doi.org/10.2307/40041279
84. Tavani HT (2008) Informational privacy: Concepts, theories, and controversies. In: Himma KE, Tavani HT (eds) The handbook of information and computer ethics. John Wiley and Sons, Hoboken, NJ, pp 131–164
85. Koops B-J, Newell BC, Timan T, Skorvanek I, Chokrevski T, Galic M (2016) A typology of privacy. U Pa J Int Law 38(2):483–575
86. Drackert S (2014) Die Risiken der Verarbeitung personenbezogener Daten. Eine Untersuchung zu den Grundlagen des Datenschutzrechts. Dunker & Humblot, Berlin
87. Britz G (2008) Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung. Mohr Siebeck, Tübingen
88. Wagner I, Eckhoff D (2018) Technical privacy metrics: a systematic survey. ACM Comput Surv 51(3):1–38. https://doi.org/10.1145/3168389
89. Pohle J (2020) On Measuring fundamental rights protection: can and should data protection law learn from environmental law? In: The Global Constitutionalism and the Internet Working Group (ed.): Don't give up, stay idealistic and try to make the world a better place - liber amicorum for Ingolf Pernice. HIIG, Berlin, pp. 71–79
90. Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 29(5):582–638. https://doi.org/10.1017/s0269888913000039
91. Barocas S, Selbst AD (2016) Big data's disparate impact. Cal L Rev (104)3:671–732. https://doi.org/10.15779/Z38bg31
92. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607
93. Verma S, Rubin J (2018) Fairness definitions explained. 2018 IEEE/ACM Int. Workshop on Software Fairness (FairWare)
94. Suresh H, Guttag JV (2021) A framework for understanding unintended consequences of machine learning. EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization; October 2021
95. Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning. Limitations and opportunities (online book). fairmlbook.org
96. Binns R (2018) Fairness in machine learning: Lessons from political philosophy. Conference on Fairness, Accountability, and Transparency (FAT) 2018
97. Mulligan DK, Kroll JA, Kohli N, Wong RY (2019) This thing called fairness: Disciplinary confusion realizing a value in technology. Proc ACM Hum-Comput Interact 3(Article No. 119):1–36. https://doi.org/10.1145/3359221
98. Hauer MP, Kevekordes J, Haeri MA (2021) Legal perspective on possible fairness measures - Why AI in HR needs help. Comput Law Secur Rev 42:105583. https://doi.org/10.1016/j.clsr.2021.105583
99. Wachter S, Mittelstadt B, Russell C (2020) Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. W Va L Rev 123(3):735–790
100. Lee MSA, Floridi L, Singh J (2021) Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. AI Ethics 1(4):529–544. https://doi.org/10.1007/s43681-021-00067-y
101. Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 5(2):153–163
102. Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent Trade-Offs in the Fair Determination of Risk Scores. Proceedings of Innovations in Theoretical Computer Science (ITCS) 2017
103. Binns R (2020) On the Apparent Conflict Between Individual and Group Fairness. FAT* '20, January 27–30, 2020, Barcelona, Spain
104. Chohlas-Wood A, Coots M, Goel S, Nyarko J (2023) Designing equitable algorithms. Nat Comput Sci 3(7):601–610
105. YVTltk (2018) Assessment of creditworthiness, authority, direct multiple discrimination, gender, language, age, place of residence, financial reasons, conditional fine. Plenary Session (voting), Register number: 216/2017, 21 March 2018. Yhdenvertaisuus- ja tasa-arvolautakunta / National Non-Discrimination and Equality Tribunal of Finland, Finland, Government Publication
106. Snelling J, McMillan J (2017) Equality: Old debates, new technologies. In: Brownsword R, Scotford E, Yeung K (eds) The Oxford handbook of law, regulation and technology. Oxford University Press, Oxford, pp 69–89
107. Hussain W (2018) The common good. In: Zalta, EN (ed.): Stanford encyclopedia of philosophy (Spring 2018 Edition)
108. Jaume-Palasi L (2019) Why we are failing to understand the societal impact of artificial intelligence. Soc Res 86(2):477–498. https://doi.org/10.1353/sor.2019.0023
109. Mantelero A (2016) Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. Comput Law Secur Rev 32(2):238–255. https://doi.org/10.1016/j.clsr.2016.01.014

110. Fischhoff B, Watson SR, Hope C (1984) Defining risk. Pol Sci 17(2):123–139
111. Jasanoff S (1993) Bridging the two cultures of risk analysis. Risk Anal 13(2):123–123. https://doi.org/10.1111/j.1539-6924.1993.tb01057.x
112. Horlick-Jones T (1998) Meaning and contextualisation in risk assessment. Reliab Eng Syst Saf 59(1):79–89. https://doi.org/10.1016/S0951-8320(97)00122-1
113. Jasanoff S (1999) The Songlines of Risk. Environ Values 8(2):135–152. https://doi.org/10.3197/096327199129341761
114. Felt U, Wynne B, Callon M, Gonçalves ME, Jasanoff S, Jepsen M, Joly P-B, Konopasek Z, May S, Neubauer C, Rip A, Siune K, Stirling A, Tallacchini M (2007) Taking European knowledge society seriously. Report of the Expert Group on Science and Governance to the Science, Economy and Society Directorate, Directorate-general for Research, European Commission. Office for Official Publications of the European Communities, Luxembourg
115. Baldwin R, Black J (2016) Driving priorities in risk-based regulation: What's the problem? J Law Soc 43(4):565–595. https://doi.org/10.1111/jols.12003
116. Cranor CF (1997) The Normative Nature of Risk Assessment: Features and Possibilities. Risk Health Saf Environ 8(Spring):123–136
117. Hansson SO (2010) Risk: Objective or Subjective. Facts or Values. J Risk Res 13(2):231–238. https://doi.org/10.1080/13669870903126226
118. European Commission (2021) Impact assessment accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts; 24.4.2021, SWD(2021) 84 final. European Commission, Brussels
119. Liu M, Ning Y, Teixayavong S, Mertens M, Xu J, Ting DSW, Cheng LT, Ong JCL, Teo ZL, Tan TF, RaviChandran N, Wang F, Celi LA, Ong MEH, Liu N (2023) A translational perspective towards clinical AI fairness. NPJ Digit Med (6)1:172. https://doi.org/10.1038/s41746-023-00918-4
120. Ansell C, Baur P (2018) Explaining trends in risk governance: How problem definitions underpin risk regimes. Risk Hazards Crisis Public Policy 9(4):397–430. https://doi.org/10.1002/rhc3.12153
121. Sovrano F, Sapienza S, Palmirani M, Vitali F (2022) Metrics, explainability and the European AI Act proposal. J - Multidiscipl Sci J 5(1):126–138. https://doi.org/10.3390/j5010010
122. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
123. Merry SE (2016) The seductions of quantification: Measuring human rights, gender violence, and sex trafficking. University of Chicago Press, Chicago
124. Bowker GC, Star SL (2000) Sorting things out: Classification and its consequences. MIT Press, Cambridge, London
125. Mennicken A, Espeland WN (2019) What's new with numbers? Sociological approaches to the study of quantification. Annu Rev Sociol 45(1):223–245. https://doi.org/10.1146/annurev-soc-073117-041343
126. Fisher E (2012) Risk and Governance. In: Levi-Faur D (ed) Oxford handbook of governance. Oxford University Press, Oxford, pp 417–428
127. Jasanoff S (2004) The idiom of co-production. In: Jasanoff S (ed) States of knowledge. The co-production of science and political order, Routledge, London, New York, pp 1–12
128. Yeung K, Bygrave LA (2021) Demystifying the modernized European data protection regime: Cross-disciplinary insights from legal and regulatory governance scholarship. Regul Gov 16(1):137–155. https://doi.org/10.1111/rego.12401
129. Fisher E (2010) Risk regulatory concepts and the law. In: OECD (ed): Risk and regulatory policy: Improving the governance of risk, oecd reviews of regulatory reform. OECD, Paris, pp. 45–92
130. Gandy OH Jr (2009) Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage. Ashgate, Farnham, Burlington
131. Nordström M (2021) AI under great uncertainty: implications and decision strategies for public policy. AI Soc 37:1703–1714. https://doi.org/10.1007/s00146-021-01263-4
132. Matz SC, Appel RE, Kosinski M (2019) Privacy in the age of psychological targeting. Curr Opin Psychol 31:116–121. https://doi.org/10.1016/j.copsyc.2019.08.010
133. EDPB/EDPS (2021) Joint opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). European Data Protection Board (EDPB), European Data Protection Supervisor (EDPS), Brussels
134. FRA (2019) Facial recognition technology: fundamental rights considerations in the context of law enforcement. European Union Agency for Fundamental Rights (FRA), Luxembourg
135. Matz SC, Teeny JD, Vaid SS, Peters H, Harari GM, Cerf M (2024) The potential of generative AI for personalized persuasion at scale. Sci Rep (14)1:4692. https://doi.org/10.1038/s41598-024-53755-0
136. Grunwald A (2019) Technology assessment in theory and practice. Routledge, London
137. Gellert R (2021) The role of the risk-based approach in the General Data Protection Regulation and in the European Commission's proposed Artificial Intelligence Act: Business as usual? J Ethics Leg Technol (3)2:15–33. https://doi.org/10.14658/pupj-jelt-2021-2-2
138. Habermas J (2010) The concept of human dignity and the realistic utopia of human rights. Metaphilosophy 41(4):464–480. https://doi.org/10.1111/j.1467-9973.2010.01648.x
139. Mares R (2019) Securing human rights through risk-management methods: Breakthrough or misalignment? Leiden J Int Law 32(3):517–535. https://doi.org/10.1017/S0922156519000244
140. Helberger N, Diakopoulos N (2023) ChatGPT and the AI Act. Internet Policy Review (12)1, https://doi.org/10.14763/2023.1.1682
141. Veale M, Zuiderveen Borgesius F (2021) Demystifying the draft EU Artificial Intelligence Act - Analysing the

good, the bad, and the unclear elements of the proposed approach. Comput Law Rev Int 22(4):97–112. https://doi.org/10.9785/cri-2021-220402
142. Christofi A, Dewitte P, Ducuing C, Valcke P (2020) Erosion by standardisation: Is ISO/IEC 29134:2017 on privacy impact assessment up to (GDPR) standard? In: Tzanou M (ed): Personal data protection and legal developments in the European Union. IGI Global, Hershey, pp 140–167
143. Van Cleynenbreugel P (2021) EU by-design regulation in the algorithmic society. A promising way forward or constitutional nightmare in the making? In: Micklitz H-W et al (eds) Constitutional challenges in the algorithmic society. Cambridge University Press, Cambridge, pp 202–218
144. Feng P (2006) Shaping technical standards. Where are the users? In: Guston DH, Sarewitz DR (eds) Shaping science and technology policy: The next generation of research, science and technology in society. University of Wisconsin Press, Madison, pp 199–216
145. Werle R, Iversen EJ (2006) Promoting legitimacy in technical standardization. Sci Technol Inno Stud 2(2):19–39
146. Wickson F, Forsberg E-M (2015) Standardising responsibility? The significance of interstitial spaces. Sci Eng Ethics 21(5):1159–1180. https://doi.org/10.1007/s11948-014-9602-4
147. Bareis J (2024) The trustification of AI. Disclosing the bridging pillars that tie trust and AI together. Big Data Soc 11(2):1–14. https://doi.org/10.1177/20539517241249430
148. Çalı B (2007) Balancing human rights? Methodological problems with weights, scales and proportions. Hum Rights Q 29(1):251–270. https://doi.org/10.1353/hrq.2007.0002
149. Lenaerts K (2019) Limits on limitations: The essence of fundamental rights in the EU. Ger Law J 20(6):779–793. https://doi.org/10.1017/glj.2019.62