The AI Act Proposal: Towards the next transparency fallacy?

Why AI regulation should be based on principles based on how algorithmic discrimination works

Bettina Berendt[1]

I. Introduction

Artificial Intelligence (AI) can entail large benefits as well as risks. The goals of protecting individuals and society and establishing conditions under which citizens find AI "trustworthy" and developers and vendors can produce and sell AI, the ways in which AI works have to be understood better and rules have to be established and enforced to mitigate the risks. This task can only be undertaken in collaboration. Computer scientists are called upon to align data, algorithms, procedures and larger designs with values, 'ethics' and laws. Social scientists are called upon to describe and analyse the plethora of interdependent effects and causes in socio-technical systems involving AI. Philosophers are expected to explain values and ethics. And legal experts and scholars as well as politicians are expected to create the social rules and institutions that support beneficial uses of AI and avoid harmful ones.

This article starts from a computers-and-society perspective and focuses on the action space of lawmaking. It suggests an approach to AI regulation that starts from a critique of the European Union's (EU) proposal for a Regulation commonly known as the AI Act Proposal, published by the EU Commission on 21 April 2021.[2]

---

[1] I thank Laurens Naudts, Geoffrey Rockwell, Pieter Delobelle, Rainer Rehak, Kristen Scott and Koen Vraenckaert for their helpful comments on an earlier version of this work and the participants of the conferences "Rechtliche Rahmenbedingungen für KI in der Schweiz" (Zürich/online, November 2021) as well as "Verbraucherrechtstage 2021" (online, July 2021) for important discussions. I received funding from the German Federal Ministry of Education and Research (BMBF) – Nr. 16DII113 f. I am also indebted to the inputs from the projects VeriLearn (Research Foundation – Flanders (FWO), EOS No. 30992574) and NoBIAS (NoBIAS – H2020-MSCA-ITN-2019 project GA No. 860630).

[2] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM(2021) 206 final 2021/0106 (COD), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

The AI Act Proposal deals with a wide range of phenomena that have been and are being discussed in large bodies of literature, including risks deemed unacceptable, manipulation, and health and safety risks. This paper will concentrate on phenomena related to AI, algorithmic systems and Big Data, discussed under terms such as "discrimination", "fairness" or "bias". I argue that regulation should be based on science and that, viewed from this angle, the AI Act Proposal falls short. In particular, its strong focus on transparency as a measure against bias and discrimination ignores relevant research on how algorithmic discrimination 'works' and how it can be countered.[3]

The AI Act Proposal is being discussed from many angles, in academic, political, and other fora. The present paper was motivated, in particular, by the analysis of *Veale* and *Zuiderveen Borgesius*,[4] but differs from it in its thematic focus (bias and discrimination) and in the approach it takes (the search for principles). Like EDRi et al.,[5] I consider it key to ground AI regulation in the protection of fundamental and human rights.

The article is organized as follows. In Section II, key terms are defined. Section III highlights the role of algorithmic discrimination in the AI Act Proposal and in the General Data Protection Regulation (GDPR) and outlines the relationship between discriminatory effects via the processing of personal data and via AI-based processing. Section IV takes a closer look at how algorithmic discrimination arises; in particular, it emphasizes the cumulative effects of human and machine biases/discrimination in real-life chains of processes, using labour as an example domain. Section V argues that the AI Act Proposal claims that the measures it requires are suitable and that these measures rely centrally on transparency. This assumption is challenged in Section VI, which shows that transparency, even if coupled with data quality, security, human oversight and documentation, is not sufficient for the goal of preventing discrimination (and in this sense, if relied upon as the core requirement, not suitable). Section VII proposes to build on the construction of data protection on a web of principles (of which transparency

---

[3] The Explanatory Memorandum points out that the proposal "is the result of extensive consultation with all major stakeholders" (p. 7 of COM[2021] 206 final 2021/0106 [COD]), "state-of-the-art for many diligent operators", "derived from the Ethics Guidelines of the HLEG, piloted by more than 350 organisations", and "largely consistent with other international recommendations and principles" (p. 13). It also notes that "[t]he precise technical solutions to achieve compliance with those requirements may be provided by standards or by other technical specifications or otherwise be developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the AI system" (p. 13). While this refers to the science and may be indirectly influenced by the science, it stops short of constituting a scientifically based set of norms and rules.

[4] *Veale/Zuiderveen Borgesius*, Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, 2021, 22 (4), 97–112.

[5] EDRi (European Digital Rights) et al., An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement, https://epicenter.works/sites/default/files/political-statement-on-ai-act.pdf, 2021.

is one, but not the only one) to shape AI regulation. Section VIII concludes by formulating legal and computer-science goals for better AI regulation: legal principles and software engineering recommendations that map principles to design strategies and these to design patterns and technologies.

II. Terminology: (Algorithmic) bias and (algorithmic) discrimination

(Unlawful) discrimination consists in making differentiations on the basis of objectionable or illegal grounds, for example on the basis of gender, sexual preference, or ethnic origin.[6] *Algorithmic discrimination (AD)* can be defined as discrimination in contexts that involve (usually digital) computers. *Friedman* and *Nissenbaum* argued, against a then frequent conception of computers as 'more objective' than humans, that computer systems can in fact be biased and can lead to discrimination. They "use the term *bias* to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others".[7]

On the one hand, their definition overlaps with the legal notion: "A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate".[8] On the other hand, this definition allows for any type of grounds of the differential treatment, including for example "effecting a long/resource-intensive computing job in a multi-user computer system". Thus, *Friedman* and *Nissenbaum* make no a priori commitment regarding whether the differentiation is unfair (in a moral sense) and whether it constitutes discrimination (in a legal sense).

AD can amount to direct discrimination, but also and more typically to indirect discrimination.[9]

Discrimination and AD are linked to complex questions of justice, equality, and fairness. In line with the dominant terminology in the current machine-learning literature and to simplify, out of these three only "fairness" will be used here.

Berendt gives an introduction to the computational-legal discussion of these questions.[10] The examples of AD in Section IV concentrate on ethnicity and gen-

---

[6] *Zuiderveen Borgesius*, Strengthening legal protection against discrimination by algorithms and artificial intelligence, The International Journal of Human Rights, 2020, 24 (10), 1572–1593.

[7] *Friedman/Nissenbaum*, Bias in computer systems, ACM Transactions on Information Systems, 14(3), 1996, 330–347, 332.

[8] *Friedman/Nissenbaum* (Fn. 7), 332.

[9] E. g. *Pedreschi/Ruggieri/Turini*, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York City (NY), 2008, 560–568; *Barocas/Selbst*, Big data's disparate impact, California Law Review, 104(3), 2016, 671–732.

[10] *Berendt*, Algorithmic discrimination, in: Comandé (Ed.), Elgar Encyclopedia of Law and Data Science, Cheltenham (UK), 2022.

der that are legally protected attributes in many jurisdictions. The attribute (body) weight will be used in one example to demonstrate how the openness of the AD definition above allows us to also consider a wider range of grounds.[11] All of these attributes are typically personal data, so the question arises whether data protection law might already be sufficient to counter AD.

III. Algorithmic systems and discriminatory effects: GDPR and AI Act Proposal

The AI Act is commonly regarded as being part of a 'family' of recent EU legislation, some of which is already in effect (GDPR) and some of which is currently in various stages of deliberation (Data Governance Act, Digital Services Act, Digital Markets Act, the updated General Product Safety Regulation, and the Product Liability Directive). The GDPR is particularly relevant, since discriminatory effects are often linked to the processing of personal data. Data protection and anti-discrimination are linked and convergent also in further legal[12] and computational[13] ways. In this section, I will survey the ways in which the GDPR and the AI Act Proposal address AD, explain why many but not all discriminatory effects are covered by the GDPR, and highlight core potentials and limitations of the two laws.

The GDPR requires personal data to be processed *in such a way as to* not produce discriminatory effects (Recitals 71, 75 and 85). The AI Act Proposal motivates its regulatory approach, inter alia, with reference to discrimination: *because* certain AI systems can have discriminatory effects, they should be operated only under constraints or be forbidden altogether (Recitals 15, 17, 28, 33, 35–39, 44 and 47). Thus, both laws recognize that AD is a specific and relevant risk of algorithmic systems (usually AI systems working on big personal data and similar software-based systems[14]) and that this risk often arises from the processing of per-

---

[11] Weight is an attribute that some argue should become a legally protected attribute, cf. *Schallenkamp/DeBeaumont/Houy*, Weight-Based Discrimination in the Workplace: Is Legal Protection Necessary?, Employee Responsibilities and Rights Journal, 2012, 24 (4), 251–259; *McCall/Bever*, Current trends in combating weight discrimination in the workplace, https://www.fisherphillips.com/news-insights/current-trends-in-combating-weight-discrimination-in-the-workplace.html, 2020.

[12] *Gellert/de Vries/de Hert/Gutwirth*, A comparative analysis of anti-discrimination and data protection legislations, in: Custers/Calders/Schermer/Zarsky (Eds.), Discrimination and Privacy in the Information Society. Data mining and Profiling in Large Databases, Berlin etc. 2013, 61–89; *Naudts*, How machine learning generates unfair inequalities and how data protection instruments may help in mitigating them, in: Leenes/van Brakel/Gutwirth/de Hert (Eds.), Data Protection and Privacy: The Internet of Bodies, Oxford 2019, 71–92.

[13] *Berendt* (Fn. 10).

[14] This argument follows the German Data Ethics Commission's terminological and semantic proposal to not limit attention to "AI" in any specific technical sense, Datenethikkommission, Opinion of the Data Ethics Commission, https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3, 2019.

sonal data. These claims are in line with a body of literature that has been growing strongly since the mid-1990s.[15]

However, the question arises whether this is enough.

To the extent that the laws talk about discrimination, they implicitly refer to the applicable notion(s) of discrimination in anti-discrimination law. However, neither of the two laws defines discrimination or bias in light of the specific particularities and challenges of AD, and neither stipulates specific principles or measures against them. This, I argue, does not take sufficient advantage of the available science. In the remainder of this article, I will survey pertinent findings and from these derive recommendations for regulation.

At this point, it is important to consider possible reasons for these gaps, for the differences between the gap in the AI Act Proposal from that in the GDPR, and to ask what we could learn from the GDPR for the AI Act Proposal.

Algorithmic systems can (a) involve (the processing of) personal data, (b) violate data protection law, and/or (c) have discriminatory effects. All combinations are possible (except combinations with (b) + not (a), since (b) requires the involvement of personal data), as the following examples show.

i. Discriminatory effects + violating data protection law (or at least data protection principles): This constellation is in the focus of many current debates around the business models of large online platforms. Via the large-scale collection of personal data on their users (what people posted, clicked on, interacted with, what friends they have and how these behave, …), advertising can become highly personalized. Marketers who buy advertising space (and the automated, real-time online bidding algorithms that control the serving of ads) often do so based on the demographics of their target groups – and platforms have such data about each individual user. The data may be true or fictitious (an individual may have an online identity with, say, a self-specified gender that deviates from the one they have or use otherwise); the data may be given explicitly or inferred by a software; and the matching may be done inside and by the platform (probably legally) or the data may be sold to the advertising partners (probably violating GDPR provisions – see the 2022 ruling of the Belgian Data Protection Authority against IAB's Transparency and Consent framework[16]). This may lead, for example, to women systematically being shown ads for lower-paying jobs than men (see Section IV).

---

[15] See *Berendt* (Fn. 10) for a survey and discussion. Seminal articles include *Friedman/Nissenbaum* (Fn. 7), *Pedreschi/Ruggieri/Turini* (Fn. 9) and *Barocas/Selbst* (Fn. 9). Ongoing conference and workshop series contribute to the further development of the field, e. g. https://www.facctconference.org/network.

[16] Decision on the merits 21/2022 of 2 February 2022, Case number: DOS-2019-01377, English translation available at https://www.gegevensbeschermingsautoriteit.be/publications/beslissing-ten-gronde-nr.-21-2022-english.pdf.

Such personalization ("targeting") is increasingly being viewed with suspicion, and large platforms are promising to abandon certain forms of it.[17]

ii. Discriminatory effects + involving personal data but compliant with data protection law and principles: In *Heinz Huber v Germany*[18], the European Court of Justice found that while a registry of personal data kept by German authorities complied with data protection law under specified restrictions, it was discriminatory because it treated non-German EU nationals different from German citizens. Examples more closely linked to AI can be found in Section IV.

iii. Discriminatory effects without involvement of personal data: A machine-learning algorithm that learns a classifier from data that reflect past discriminatory patterns, can replicate these patterns. The learning may happen on personal data (such as microdata) or on anonymous data or even on synthetic data (generated on the basis of statistical distributions on real-life data). The last two types of training data are not considered personal data. The application of the classifier to new individuals, for example in order to decide on allocations of goods and services (such as job ads) or decisions with significant consequences such as recruitment decisions involves personal data (namely those of the new individual).

However, harms may also arise without any involvement of such data in algorithmic processing. Harms can arise through biases, stereotypes and similar that are implicit in text corpora and that can have subtle, unexpected, and still pervasive effects in subsequent data processing steps, as described in more detail in Section IV below.

Both laws are motivated by the goal of avoiding risks, but their starting points differ.

Data protection law is, to a large extent, driven by the insight that structural asymmetries, often power asymmetries, pose risks to fundamental rights. The GDPR emphasises, in various places, that it considers "risks to the fundamental rights and freedoms" – the rights to data protection and privacy in particular, but also all others. This implies that the GDPR is also meant to, and designed to, protect against risks of individual discrimination. The AI Act Proposal, on the other hand, is strongly patterned on product safety laws.[19] Thus, products (now extended to also include AI systems) are conceived of as sources of risks and harms, and a process resulting in a CE mark for an AI system[20] is proposed as a remedy.

---

[17] E. g. The Guardian, Facebook bans ads targeting race, sexual orientation and religion, https://www.theguardian.com/technology/2021/nov/10/facebook-bans-ads-targeting-race-sexual-orientation-and-religion, 2021.

[18] Case C-524/06 Heinz Huber v Bundesrepublik Deutschland [2008] ECR 2008 I-09705.

[19] *Veale/Zuiderveen Borgesius* (Fn. 4).

[20] See the definition in Article 3 (24) AI Act Proposal. A commercial product with a CE mark ("conformité européenne") indicates that the manufacturer or importer affirms the good's conformity with European health, safety, and environmental protection standards. The CE marking is required for goods sold in the European Economic Area (EEA), but is also found on products sold elsewhere that have been manufactured to EEA standards.

Both laws also are aware of societal risks, but both 'lineages' can create blind spots and challenges for protecting against social risks.

Data processing affects not only the individual, but also collectives of various kinds, up until and including the fabric of democracy itself.[21] Thus, increasingly, data protection law has to be interpreted also with a view to protecting against societal harms. However, the GDPR's focus on (individual) fundamental rights presents challenges; for example, the effects of profiling on social groups are difficult to subsume.[22]

The AI Act Proposal explicitly regulates products (AI systems) that lead to "detrimental or unfavorable treatment of certain persons or whole groups thereof" (Article 5 (1) (c)), but it does so in the relatively narrow context of the prohibited AI systems regulated in Article 5. Challenges for dealing adequately with these risks derive from the AI Act Proposal's heritage from product safety, which is typically the protection of an (individual) 'user' of a product against risks that arise from the interaction between said product and said individual, risks that in traditional products are handled by requiring certain modifications to those products. Therefore, the AI Act Proposal requires modifications to these AI systems rather than taking a wider view of the algorithmic systems / information systems / socio-technical systems that the AI is part of and that are often the underlying origins of risks and harms, such that AI primarily 'automates [existing] inequality' (Eubanks, 2018).

Both laws operate in a socio-legal context in which discrimination itself is predominantly viewed as deriving from a locatable decision. The traditional focus on individual human decision-makers who discriminate has led to a corresponding focus on individual machine decision-makers that discriminate[23]. As a result, the law and its enforcement face difficulties in the attempt to protect against structural discrimination. This problem has been recognised with regard to discrimination in the workplace under changing managerial policies, even before the advent of AD.[24]

---

[21] See the German Constitutional Court's Census Judgment of 1983: "Persons who assume, for example, that attendance of an assembly or participation in a citizens' interest group will be officially recorded and that this could expose them to risks will possibly waive exercise of their corresponding fundamental rights (Articles 8 and 9 of the Basic Law). This would not only restrict the possibilities for personal development of those individuals but also be detrimental to the public good since self-determination is an elementary prerequisite for the functioning of a free democratic society predicated on the freedom of action and participation of its members." (Bundesverfassungsgericht [BVerfG], Urteil vom 15.12.1983 – 1 BvR 209/83, 1 BvR 269/83, 1 BvR 362/83, 1 BvR 420/83, 1 BvR 440/83, 1 BvR 484/83, translation at https://freiheitsfoo.de/census-act/, retrieved 2021-12-07).

[22] Taylor/Floridi/van der Sloot (Eds.), Group Privacy: new challenges of data technologies, Dordrecht 2017.

[23] As an example, consider the wording of Article 22 GDPR.

[24] *Bagenstos*, The structural turn and the limits of antidiscrimination law, California Law Review, 94(1), 2006, 1–47.

The present paper starts from a focus on risks for fundamental rights and for society, and from the assumption that data protection's roots in tracing problems back to structural (power) asymmetries can help us re-centre structural asymmetries and structural discrimination when regulating AI. As the discussion in the present section has shown, the GDPR alone is not sufficient to address these issues.

In the next section, we will take a closer look at examples of discrimination that for the most part are not directly linked or linkable to a specific use of an individual's personal data or to a specific decision that can be called discriminatory.

IV. How and why do algorithms discriminate?

AD arises from interactions between data and algorithms and the larger sociotechnical systems they are used in. To illustrate these interactions, I will sketch computational effects with the help of examples from the scientific literature and popular media.

First, we will consider how AD can result from key elements of machine learning – the training data, the algorithms, and category labels – in an idealized linear pipeline of functioning.

When data are biased, so will system representations learned from these data (without corrections). The Twitter bot Tay, designed to learn to tell jokes from interacting with human Twitter users, learned to spew out racial slurs because it was fed with racist input. This can happen very fast – Tay bot was taken offline within a day.[25] In predictive policing, using data on drug-related arrests for training has led to system recommendations to patrol the area in which most arrests were made, leading to more arrests in this area. When these areas are predominantly 'black' neighbourhoods, drug use in 'white' areas becomes under-patrolled and under-reported.[26] In addition, the statistical under-representation of some demographics in training data can lead to more prediction errors. This has been observed for facial recognition algorithms that have led to several innocent black men being arrested.[27]

Machine-learning algorithms need to have some strategy for generalising beyond the data they have seen, i. e. in order to be able to function at all. This can lead to socially biased algorithmic output. For example, recommender algorithms that are based on popularity can lead to people receiving sexist, racist, etc. claims as

---

[25] *Wakefield*, Microsoft chatbot is taught to swear on Twitter, BBC News, 2016, https://www.bbc.com/news/technology-35890188.

[26] *Lum*/*Isaac*, To predict and serve? Significance, 2016, 13(5), 14–19.

[27] *Hill*, Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. The New York Times, https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html, 2020.

query-completion suggestions of what they may be looking for[28] or the recommendation to consult somebody's criminal record when they search for a black-sounding name more often than when searching for a white-sounding name[29]. Algorithms designed to detect and filter out pornographic images have been alleged to compare the percentage of nude-coloured pixels with a threshold, which may lead to disproportionately high (mis)classifications for images of overweight people, which in turn can lead to disproportionate blocking of content and accounts.[30]

The categories to which a predictor maps can be a source of further problems. For example, body scanners at airports are designed to detect 'unexpected' shapes and objects on passengers – which means they have to 'expect' for example bra underwiring on female passengers, or genitals on men, to avoid causing an alarm for every single person. Airport staff therefore inform the body scanner, usually by the press of a button, of the passenger's gender before scanning. This applied binary gender schema can lead to patterns of false alarms and disproportionately burden non-binary people with pat-downs and further questions.[31]

Second, we will consider the combined effects of several steps of machine and human learning. Such combinations can produce strong AD effects. The cumulation can be understood using the notions of allocative and representational harm.

AD can consist in withholding opportunities or resources (*allocative harm)*,[32] or in perpetuating stereotypes and cultural denigration (*representational harm*). In the examples above, many harms are allocative (arrests, account blocking, airport security treatment), while some seem primarily representative (jokes, recommendations). Discrimination often works through chains of such harms, an effect that has been termed "pernicious" or "runaway" feedback loops.[33] Such loops have been described as occurring in environments dominated by one learning algorithm (ibid.) or in multi-algorithmic environments such as the Web.[34] They can

---

[28] UN Women, UN Women ad series reveals widespread sexism, http://www.unwomen.org/en/news/stories/2013/10/women-should-ads, 2013.

[29] *Sweeney*, Discrimination in Online Ad Delivery. Communications of the ACM, 2013, 56(5), 44–54.

[30] *Richman*, This is the impact of Instagram's accidental fat-phobic algorithm. https://www.fastcompany.com/90415917/this-is-the-impact-of-instagrams-accidental-fat-phobic-algorithm, 2019.

[31] *Waldron*/*Medina*, When transgender travelers walk into scanners, invasive searches sometimes wait on the other side, ProPublica, https://www.propublica.org/article/tsa-transgender-travelers-scanners-invasive-searches-often-wait-on-the-other-side, 2019.

[32] *Blodgett*/*Barocas*/*Daumé III*/*Wallach*, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Stroudsburg / PA 2020, 5454–5476; also referred to as *distributive harm*, *Binns*, Fairness in Machine Learning: Lessons from Political Philosophy, in: Proceedings of Machine Learning Research 81, 2018, 149–159.

[33] *O'Neil*, Weapons of Math Destruction, New York 2016; *Ensign*/*Friedler*/*Neville*/*Scheidegger*/*Venkatasubramanian*, Runaway Feedback Loops in Predictive Policing, Proceedings of Machine Learning Research 81, 2018, 160–171.

[34] *Baeza-Yates*, Bias on the web, Communications of the ACM 61(6), 2018, 54–61.

become even more pernicious when many different, independent actors and steps, including non-algorithmic ones, are involved. I will illustrate this with a cumulation scenario of phenomena observed around algorithms involved in people's paths into jobs. Each step is annotated with its predominant type of discrimination, type of harms, and involvement of personal data / data protection violations.

1. People (prospective candidates as well as prospective employers) perceive a world in which high-paying, prestigious jobs tend to be held by men (and less prestigious jobs by women).

This perception may derive from actual observation, but more frequently it is mediated through the texts one reads – but algorithms can mirror and exacerbate imbalances. This is easy to see in machine translation. For example, the German sentence "Sie ist eine gute Ärztin" (she is a good doctor) was translated, by Google Translate[35], to Turkish as "O iyi bir doktor", but then back to German as "Er ist ein guter Arzt" (he is a good doctor). This is correct in the sense that Turkish has no gender pronouns like German or English, but biased in the sense that the machine translation chooses the male interpretation.

This bias is probably due to the much higher frequency of example sentences found in parallel corpora online (= the training data of machine translation algorithms) in which the doctors described were actually male.[36] Google has been called out on such biases and addressed them in its translations to and from English: "O iyi bir doctor" is translated as "he is a good doctor *or* she is a good doctor". However, this improvement is an exception local to English, and it is brittle. For example, "Murat is her son" gets translated as "Murat onun oğlu" and back as "Murat is his son".

[Direct discrimination; representational harm; personal data generally not involved]

2. Women are served ads for different jobs than men. *Datta* et al. found that high-paying jobs tended to be served to men.[37] *Kayser-Bril* showed, in an experimental study, that Facebook served specific job ads mostly to men or mostly to women (e. g. machine learning developer vs. nurse), even if the ads are phrased in gender-neutral ways.[38] *Imana*, *Korolova*, and *Heidemann* found similar bias for *one* job type (delivery driver) mirroring the gender ratios in the respective company that placed the ad.[39]

---

[35] All translations were generated in July 2021.

[36] This interpretation can be substantiated by using a translation engine that shows context sentences, such as context.reverso.net.

[37] *Datta*/*Tschantz*/*Datta*, Automated experiments on ad privacy settings, Proceedings on Privacy Enhancing Technologies 2015 (1), 92–112.

[38] *Kayser-Bril*, Automatisierte Diskriminierung: Facebook verwendet grobe Stereotypen, um die Anzeigenschaltung zu optimieren, https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren/, 2021.

[39] *Imana*/*Korolova*/*Heidemann*, Auditing for discrimination in algorithms delivering job ads, in: Proceedings of The Web Conference 2021 (WWW '21), New York City (NY) 2021, 3767–3778.

[Direct discrimination; allocative harm; personal data are involved, currently not considered a data protection law violation[40]]

3. Ads are phrased in ways that women tend not to apply for a position because "they do not find themselves in it" or because of lower self-esteem.[41]

[Direct discrimination operating through psychological effects; can lead to allocative harm; no personal data involved]

4. Descriptions of jobs and descriptions of people holding them are machine-learned as describing good matches of CVs to jobs. Through the historical over-representation of men in these jobs, the machine may learn to map attributes that are typically found in job applications written by women to the class of not suitable applicants. The algorithm may pick up the explicit gender attribute as a predictor; more frequently though, it will choose 'female wording' (see previous item) or even all-women educational institutions as predictors.[42] Business goals affect what prediction errors the algorithm is designed to avoid, which can lead to further bias.[43] As a result, fewer women will be invited to job interviews, and fewer women will be recruited.

[Indirect discrimination; allocative harm; applicant's personal data involved]

5. The effects in 4. may be exacerbated when modern pre-processing methods for language understanding are used that rely on word embeddings or pre-trained language models. Through the biases embedded in the large corpora on which these intermediate models are learned,[44] associations with skills, characteristics, and features that may influence the subsequent classification in subtle ways, can be learned and may lead to biased outcomes.[45]

[Indirect or direct discrimination; representational harm; personal data may or may not[46] be involved.]

---

[40] Facebook describes targeting by gender as a technically possible and legitimate choice: https://www.facebook.com/business/help/151999381652364 (retrieved 21 Dec. 2021).

[41] *Burell*, Die Diskriminierung steckt oft im Detail, Der Spiegel, 15 June 2021, https://www.spiegel.de/start/stellenanzeigen-werden-oft-fuer-maenner-formuliert-wie-frauen-trotzdem-den-job-bekommen-a-2ff0215c-009c-48b1-b045-a462ca808cd7.

[42] *Dastin*, Amazon scraps secret AI recruiting tool that showed bias against women, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G 2018; *Lauret*, Amazon's sexist AI recruiting tool: how did it go so wrong?, https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e, 2019.

[43] *Lauret* (Fn. 42).

[44] Of the type 'doctors are men, nurses are women', *Bolukbasi*/*Chang*/*Zou*/*Saligrama*/*Kalai*, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, 4349–4357.

[45] Specific complications arise in languages that are more strongly gendered grammatically than English, for example Dutch (*Delobelle*/*Winters*/*Berendt*, RobBERT: a Dutch RoBERTa-based Language Model, in: EMNLP [Findings] 2020: 3255–3265).

[46] Aggregate statements ("women are …" etc.) usually do not involve personal data.

6. Public employment services, a different and independent actor than those in 1.–5., also rely on data analysis. Due to historically grown imbalances on the labour market, machines learn that women are less likely to find (re-)employment fast and therefore assign them a worse risk score/class in a statistically based risk-assessment system. If this classification leads to the job-seeker not receiving funding for (re-)training courses, they may find it harder still to find a new position, and their self-esteem will suffer. Due to the results of classification, this will affect women more often.

Additional effects may result from variable and user interface design. In the system analysed by *Allhutter* et al., only female job-seekers are asked whether they have "care obligations" and whether these are "being taken care of" such that they can be available fully to the labour market.[47] Since this variable is lacking for men, any predictive power of this attribute will likely discriminate against some women.

[Indirect discrimination, regression-based models may also lead to direct discrimination; allocative and representational harm; personal data of the new job seeker involved, whether data protection rights are being violated is one question in an ongoing legal dispute, cf. ZackZack[48]]

7. As a result, fewer women will hold well-paid positions, which will lead to data that show that women do not work in these jobs, and to texts that describe this world, in which "a good doctor" is male. Data and texts are objective and representative.

8. Go to step 1.

Even from this highly simplified sequence alone, it is clear that not only algorithms or data are to blame. It is also obvious that data protection law cannot be called upon to mitigate all of the intermediate nor, *a fortiori*, the cumulative effects. Data are generated and collected in socio-technical systems, and algorithms operate on modelling choices and feature and category definitions that are in themselves value-laden. Decisions are not only made by companies or algorithms, but also by the affected individuals themselves, which exacerbates structural discrimination. In addition, algorithms are the back-end of systems with user interfaces and application-specific additions that were often added with good intentions but without an understanding of how they affect the system outcome as a whole.

Examples of assumptions and choices (which may further drive the feedback loops) are discussed in the analyses of public employment services algorithmic systems of several countries by *Allhutter* et al. and by *Jędrzej*, *Sztandar-Sztander-*

---

[47] *Allhutter*/*Mager*/*Cech*/*Fischer*/*Grill*, Der AMS Algorithmus – Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS), http://epub.oeaw.ac.at/?arp=0x003bfdf3/, Wien 2020.

[48] *ZackZack*, Diskriminierung und fehlende Gesetzeslage: Darum ist der AMS-Algorithmus gefährlich, https://zackzack.at/2021/05/01/diskriminierung-und-fehlende-gesetzeslage-darum-ist-der-ams-algorithmus-gefaehrlich/, 2021.

ska and *Szymielewicz*.[49] An example of assumptions is that the reasons for being unemployed lie principally with the job-seeker themselves (as reflected by the detailed way in which the individual job-seeker is modelled in the prediction model and the scarce representation of labour market factors).

An example of design choices is related to human oversight: While job counsellors can override the system's risk classification of an individual, the system requires an extra justification to be entered in such cases and thereby, especially given counsellors' time constraints, nudges them towards accepting the proposal. In a wide range of contexts, people tend to favour suggestions from automated decision-making systems and ignore contradictory information even if it is correct ("automation bias"). In the Polish public employment service system, case workers accepted the system's suggestion in 99.4 % of cases.[50] Thus, not all forms of human oversight are effective safeguards. The AI Act Proposal stipulates, in Article 14 (4) (b), that design should make users aware of automation bias, but it is unclear whether such awareness would suffice to avoid it.

The complexity of these effects is well-known in an active and growing community of researchers and practitioners, and many approaches to mitigating such discriminatory effects have been and are being developed. Still, the challenges remain, in particular since every real-life system rests on interconnected global and local design decisions, allocation decisions, and representation decisions both by human and machine actors. In addition, the very concepts of what constitutes discrimination are evolving. An overview of current approaches and open questions with a special focus on connections between informatics and legal concerns can be found in *Berendt*.[51]

To what extent does the AI Act Proposal address such complex effects? To answer this question, we need to understand how it claims to counter risks.

V. The AI Act Proposal's claim of proportionality

According to the AI Act Proposal's Explanatory Memorandum, "[t]he proposal […] is proportionate and necessary to achieve its objectives, since it follows a risk-based approach" (p. 7). This statement involves two claims.

First, the proposal weighs infringements on the rights of AI providers against risks and "imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety" (ibid., p. 7). This first claim appears validated by the construction of the law.

---

[49] *Allhutter* et al. (Fn. 47); *Jędrzej/Sztandar-Sztanderska/Szymielewicz*, Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making, https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf, 2015.

[50] *Jędrzej/Sztandar-Sztanderska/Szymielewicz* (Fn. 49).

[51] *Berendt* (Fn. 10).

Second, these "regulatory burdens" on AI providers must at the same time act as "measures" (as the GDPR would call them) for another important class of stakeholders, namely those affected by the algorithmic system. To be proportional, these measures must be suitable, necessary, and proportional in the narrow sense to achieve the goal of protecting these stakeholders' fundamental rights, such as the right to not be discriminated against.

But are they suitable?[52]

Title II's prohibitions of certain AI systems is, under the assumption that enforcement is possible, trivially suitable: a product that does not exist cannot harm anyone.

More interesting are the provisions concerning "high-risk" AI systems in Title III. These systems are allowed. The measures proposed comprise[53], in particular, transparency of various kinds, from and to various actors[54] including documentation and traceability as well as "information […] in relation to possible risks to fundamental rights and discrimination" (Recital 47). They also comprise high-quality data, human oversight, accuracy, robustness, and safety. Procedurally/structurally, a risk management system is required (Article 9).

Transparency is also the central requirement for "certain AI systems" that pose manipulation risks and that are regulated (less stringently than the so-called "high-risk" systems) in Title IV.

No doubt transparency is relevant and useful. This principle has been regarded as a cornerstone of democracy and its checks and balances for a long time, and it is nicely summarized as "sunlight is the best disinfectant".[55]

But is this enough?

VI. Limitations of transparency as a measure against discrimination

All of the examples in Section IV have been well-publicised, and phenomena like the scarce representation of women in high-profile jobs and the gender pay gap are common knowledge and the subject of many openly available datasets, many of which have high data quality and security. Arguably, there is human oversight in the labour sector, for example through employee representation and trade unions. So it seems that discrimination can also thrive under the conditions required by the AI Act Proposal.

An additional problem is that transparency, e. g. in the form of an explanation, "is probably not the remedy you are looking for" – especially when transparency is

---

[52] Since we will argue that the answer to this question is "no", we will not analyse necessity and proportionality *strictu sensu*. These questions can be the subject of future work.

[53] The wording in this paragraph is a minor re-organisation of the terms in the Explanatory Memorandum, pp. 7, 13.

[54] *Veale/Zuiderveen Borgesius* (Fn. 4), 104 ff.

[55] Modified from *Brandeis*, What publicity can do, Harper's Weekly, Dec 20, 1913, 10–13.

ensured by an explanation given ex post and the harm has already been done and is irreversible.[56] *Edwards* and *Veale* have called the overly narrow focus on explanations as an answer to data-processing challenges in the GDPR a "transparency fallacy"; the point of the present section is to argue towards the same conclusion for algorithmic systems and AI.

Power. From a sociological perspective, it can be observed that the knowledge that (and even how) discrimination has happened does not per se change existing power relations or structures.[57] The quest for a commonly acceptable fair solution requires more than just transparency and knowledge, but also power checks through rules of discourse, legal safeguards, and political deliberation in (to the extent possible) domination-free spaces of discourse.

Feedback loops. Bias and discrimination have self-reinforcing dynamics that are well-known in sociology ("cumulative causation"),[58] but that can take on unknown speed, opacity and effectiveness in technologically-enhanced decision making – the so-called "pernicious" or "runaway" feedback loops.[59] Allocative and representational *harms* are constitutive of such loops.

Feedback loops can be a challenge in systems that are fairly simple in the sense of mainly relying on one algorithm.[60]

Architecture. The challenge becomes larger when it is unclear which algorithms have which effects, in a time of IT infrastructures in which even entities perceived as one system, such as an Internet platform, are increasingly composed of huge complex and dynamic web of components.[61] The growing intractability of technical 'decisions' resulting from modern IT infrastructures can be said to extend developments in managerial decisions: as *Bagenstos* has argued, "boundaryless workplaces" in which peer decisions assume an increasingly important role, make it more difficult to apply anti-discrimination law that is patterned on identifying flaws in traditional decision making by superiors.[62] And the challenges become even more untraceable when the discriminatory effects arise in the context of larger sociotechnical systems.[63]

---

[56] *Edwards/Veale*, Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for, Duke Law & Technology Review, 16 (1), 2017, 18–84.

[57] E. g. *D'Ignazio/Klein*, Data Feminism, Cambridge (MA) 2020; *Miceli/Posada/Yang*, Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?, in: Proceedings of the ACM on Human-Computer Interaction 6 (GROUP), 2022, 1–14.

[58] *Myrdal*, An American Dilemma: The Negro Problem and Modern Democracy, New York 1944.

[59] *O'Neil* (Fn. 33); *Ensign* et al. (Fn. 33).

[60] See various examples in *O'Neil* (Fn. 33).

[61] *Gürses/Van Hoboken*, Privacy after the agile turn, in: Polonetsky/Tene/Selinger (Eds.), Cambridge Handbook of Consumer Privacy, Cambridge 2018, 579–601.

[62] *Bagenstos*, The structural turn and the limits of antidiscrimination law, California Law Review, 94(1), 2006, 1–47.

[63] See *McDonald/Barwulor/Mazurek/Schaub/Redmiles*, "It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online, in: 30th USENIX Security

Veale and *Zuiderveen Borgesius* have outlined further problems in the construction of the AI Act Proposal, e. g. for responsibility and accountability, that arise from an overly limited understanding of the boundaries of AI(-involving) systems and their structures and actor roles.[64] In future work, also these effects should be investigated with respect to their possible contribution to discriminatory effects.

Categories. The very categories used to describe discrimination can play an ambivalent role. On the one hand, it is commonly agreed that categories need to be observed in order to trace discrimination.[65] On the other hand, the continued use of a category can also serve to perpetuate ingroup-outgroup boundaries and bias and discrimination resulting from them.[66]

Non-unique concepts of fairness. Ethically, legally and politically, the question is whether some constellation is considered unfair at all. Who defines this, and based on which concepts of fairness?

Computational approaches are confronted with long-standing philosophical problems of the different notions of fairness and the non-commensurability between some of them. In the machine-learning literature, these problems have resurfaced in the form of different fairness metrics for data and predictions/prescriptions and the impossibility to satisfy some of them simultaneously[67] and also in the recognition that the same problem has been observed in previous attempts at formalization.[68] Political and legal scholars have pointed out that different contexts and domains are perceived as requiring different notions of fairness and thus different metrics.[69]

In addition to these considerations, practical measures are needed for involving multiple *stakeholders* (besides the developers and decision-makers) in meaningful

---

Symposium, USENIX Security, Berkeley (CA) 2021, 375–392, for an example of the interactions between loosely coupled internet platforms and the dominance of US-based ethics over the law that is applicable in users' countries. The AD reported in *McDonald* et al. is difficult to counteract also because it falls neither under European nor under German anti-discrimination law (*Pekel*, Airbnbs schwieriger Umgang mit Sexarbeiter:innen, https://netzpolitik.org/2020/diskriminierung-airbnbs-schwieriger-umgang-mit-sexarbeiterinnen/, 2020). And the AI Act Proposal would not subsume this 'trustworthiness score' under its prohibition of social scoring because the operator is a private company.

[64] *Veale*/*Zuiderveen Borgesius* (Fn. 4).

[65] Cf. the arguments about "color blindness" failing, e. g. Neville/Gallardo/Sue (Eds.), The Myth of Racial Color Blindness: Manifestations, Dynamics, and Impact, Washington, D.C. 2016.

[66] See *Bowker*/*Star*, Sorting things out: classification and its consequences, Cambridge (MA) 1999, for a comprehensive study of the effects of categorization.

[67] For a recent overview and clustering, see for example *Barocas*/*Hardt*/*Narayanan*, Fairness and Machine Learning, http://www.fairmlbook.org 2019. For their relation to legal criteria, see *Wachter*/*Mittelstadt*/*Russell*, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, Computer Law & Security Review, 2021, 41, 105567.

[68] *Hutchinson*/*Mitchell*, 50 years of test (un)fairness: Lessons for machine learning, in Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* 2019, New York City (NY) 2019, 49–58.

[69] *Binns* (Fn. 32); *Zuiderveen Borgesius* (Fn. 6).

ways, helping them express their applicable fairness notions, and resolving conflicts.

In sum, it appears that transparency, even if coupled with data quality, security, human oversight and documentation, alone is not suitable to counteract AD.

VII. Transparency is not enough: Lessons learned from data protection

In the light of these observations, we once again turn to data protection for inspiration.

Transparency is an important principle in data protection theory and data protection law. A classic argument was given by the German Constitutional Court in its 1983 Census Judgment: "A social order in which individuals can no longer *ascertain* who knows what about them and when and a legal order that makes this possible would not be compatible with the right to informational self-determination." (emphasis added).

The GDPR lists transparency as one of its fundamental principles (Article 5 (1) (a)), and it also contains requirements of documentation (through the principle of accountability, Article 5 (2)), data quality (Article 5 (1) (d)), IT security (Article 5 (1) (f)), and certain forms of human oversight ("right to obtain human intervention" under the conditions of Article 22 (3)).

However, it has long been recognized that transparency and similar requirements are not enough to protect against risks arising from the processing of personal data. Already in the paragraph immediately following the one cited, the Census Judgment adds that "the fundamental right [to informational self-determination[70]] guarantees in principle the power of individuals to *make their own decisions* as regards the disclosure and use of their personal data." – in other words, to have not only knowledge but also some extent of *control* over data processing concerning them. This has been operationalized further in the privacy protection goal of "intervenability",[71] and it has found its expression in the rights to rectification and erasure, the right to object and the rights associated with automated individual decision-making (Articles 13–22 GDPR).

Privacy and/or data protection guidelines (such as the 1980/2013 OECD Guidelines[72]) and laws (such as the GDPR in Article 5) have recognized a small

---

[70] "[…] the authority of the individual to decide himself, on the basis of the idea of self-determination, when and within what limits information about his private life should be communicated to others".

[71] *Hansen*/*Jensen*/*Rost*, Protection goals for privacy engineering, 2015 IEEE Security and Privacy Workshops, New York City (NY) 2015, 159–166.

[72] OECD, Guidelines on the protection of privacy and transborder flows of personal data. https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm (1980), https://www.oecd.org/digital/ieconomy/privacy-guidelines.htm, (2013).

but powerful set of further principles that derive from a long history of observations of risks and harms caused by the processing of personal data. *Hoepman* has provided a comparative analysis of these principles (including but not limited to the list of seven in Article 5 GDPR), for example the above as "individual participation".[73]

The principles of data minimization and purpose limitation are prime examples that illustrate that transparency alone is not sufficient to protect against abuses and unintended risks of the processing of personal data. They derive from social- and computer-science findings such as 'no data is innocent', contextual integrity as a widespread desideratum of privacy, and the fact that any re-use of data (usually in combination with re-purposing and linking previously unlinked data items) can lead to undesired re-identification and profiling of individuals.[74]

VIII. Conclusions: Legal and computer-science goals

Based on the observations and research results concerning the ways in which AD operates, and how it often operates not only (or even not at all) in isolated datasets, algorithms, or even AI systems, but in larger socio-technical systems, I consider two additions to current practices necessary. First, legal provisions (such as the AI Act) should extend beyond their current focus on transparency (and documentation, data quality, data security and human involvement in their various currently discussed forms). Second, these principles have to be transformed into system design.

1. Principles for protection against bias and discrimination

The following requirements should be turned into additional principles to be embedded into laws. As in the GDPR, these requirements should be accompanied by the requirement to deploy appropriate technical and organizational measures. This effort can draw on the lessons learned with data protection principles, but it goes beyond them because "data protection is […] less contentious than anti-discrimination. Indeed, data protection is about one particular operation (the processing of personal data), the status of which is unproblematic. Discrimination goes a step further because it does not regulate an action as such (e. g., data processing), but a legal consequence of any actions (thus, also including eventually

[73] *Hoepman*, Privacy design strategies, in: ICT Systems Security and Privacy Protection – 29th IFIP TC 11 International Conference, Berlin etc. 2014, 446–459.

[74] Transparency *can contribute* to control and the reduction of power asymmetries (*Naudts/Dewitte/Ausloos*, Meaningful transparency through data rights: A multidimensional analysis, in: Kosta/Leenes/Kamara [Eds.], Research Handbook on EU data protection, Cheltenham [UK] 2022) – the point is that it also may *not* do this.

data processing), which inherently entails operating a (legal) qualification of the facts."[75]

a) Bias and discrimination avoidance should be foundational. This has two aspects: *detection and prevention* (or at least mitigation).

b) Feedback loop interruption: loops should be anticipated, detected, and broken.

c) Harms recognition: the law should protect not only against risks to "health and safety" or "physical and psychological harms to individuals" (the wordings in Articles 5 and 7). These allocative and individual harms need to be supplemented by typical discrimination risks: further allocative harms and group-related representational harms.

Representational harms are difficult to operationalise, not least because one person's or group's discriminatory language is another person's or group's freedom of speech. When such language is created by a human, one can apply legal concepts such as libel, defamation, sedition and incitement to hatred and violence, or hate speech. Legal certainty should be created regarding language generated by machines (especially the responsibility and accountability for linguistic content), and state-of-the-art measures should be deployed to minimize the likelihood of such harms. Assigning responsibility and implementing measures becomes more challenging as language-processing pipelines are becoming more complex and dynamic (e. g. in architectures using pre-trained language models).

d) Fundamental rights orientation: If risk classes of AI systems are to be retained, the classification of a system should not rest on its function or operator, but its impact(s). For example, a chatbot is not per se more manipulative than, e. g., emotion recognition built into a non-chat-bot recommender system. Also, private-sector AI firms may cause "grave socioeconomic consequences [to individuals] similar to the exclusion of state-provided services".[76] The potential for harm should be assessed by a fundamental-rights impact assessment that takes into account the previous points. This principle follows the GDPR's requirement for a data protection impact assessment (Article 35) and the German Data Ethics Commission's treatment of risk classes.[77]

e) Categories awareness: The use of sensitive categories[78] should be justified in an impact assessment by an explanation of how the benefits of this use exceed its

---

[75] *Gellert* et al. (Fn. 12), p. 71.

[76] *Veale/Zuiderveen Borgesius* (Fn. 4), p. 100.

[77] See Datenethikkommission (Fn. 14). Mandatory impact assessments have also been called for by, e. g., ECNL (European Centre for Not-for-Profit Law), ECNL position statement on the EU AI Act, https://ecnl.org/news/ecnl-position-statement-eu-ai-act, 2021 and AlgorithmWatch, Draft AI Act: EU needs to live up to its own ambitions in terms of governance and enforcement, https://algorithmwatch.org/en/eu-ai-act-consultation-submission-2021/, 2021.

[78] These include, but are not limited to the GDPR's "special categories of data". See *Schwartz*, Classifying books, classifying people, https://medium.com/the-bytegeist-blog/classifying-books-classifying-people-302430282a3d, 2018, for why this need not even be limited to personal data.

risks. Such risk-benefit thinking appears to be implied by the claim to proportionality when this claim is interpreted w.r.t. the individuals affected by the AI (see Section V). In a sense, Article 10 (5) AI Act Proposal constitutes an example of Article 9 (2) (g) GDPR: Processing of special categories of personal data shall *not* be prohibited if "processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued …", with bias monitoring a processing that is necessary to minimize or avoid discrimination. This explicitness of proportionality is missing in the AI Act Proposal, and the protection of individual seems to be weaker ("appropriate safeguards" rather than the GDPR Article's "suitable and specific measures to safeguard").[79] The risk of perpetuating pernicious labelling of people through the use of categories (whether from a list of "special categories" or not) should be one of the risks considered in these analyses.

f) Metrics specificity: The concepts and metrics of fairness employed must correspond to the concepts used in the application domain; they should be derived via stakeholder-based and democratic procedures; and their choice and design should be justified in the technology impact assessments.

g) Stakeholder orientation: Stakeholders should be involved in design and processes. Involvement should include rights to transparency and participation. Who counts as a relevant stakeholder and their specific rights should be decided based on, inter alia, current debates around data protection.[80]

As *Vedder* observes, transparency obligations in data protection law have an "obvious addressee": the data subject.[81] But who would be an appropriate addressee for transparency regarding discriminatory risks if no personal data are involved and therefore no such individual can be determined? *Vedder* therefore considers transparency "not the end [but] just a beginning" and calls for a regulatory regime that enables deliberations about the possible impacts on humans. The AI Act Proposal previews public databases, such that for example consumer rights group may act as or on behalf of addressees to obtain information. However, they cannot act to intervene because "affected [individuals and] communities are provided with no mechanism for complaint or judicial redress".[82]

---

[79] The exemption of Article 10 (5) "can only be used in relation to high-risk systems, and only by those systems' providers" (*Veale/Zuiderveen Borgesius* [Fn. 4], 103). It remains to be seen whether this is useful for categories awareness or not.

[80] *Naudts* (Fn. 12).

[81] *Vedder*, Why data protection and transparency are not enough when facing social problems of machine learning in a big data context', in: Bayamlioğlu/Baraliuc/Janssens/Hildebrandt (Eds.), Being profiled: Cogitas, ergo sum. 10 Years of Profiling the European Citizen, Amsterdam 2018, 42–45. The AI Act Proposal, on the other hand, focuses on transparency towards "users", thus neglecting data subjects who are not users (*Sesing/Tschech*, AGG und KI-VO-Entwurf beim Einsatz von Künstlicher Intelligenz, MMR – Zeitschrift für IT-Recht und Recht der Digitalisierung, 2022, 25, 24–30).

[82] *Veale/Zuiderveen Borgesius* (Fn. 4), 112.

2. Software design

To transform the principles into system design, I build on the idea of privacy design patterns.[83] *Hoepman* proposed to map *principles* to *design strategies*, identify *design patterns* as the conceptual implementation of a design strategy, and then choose appropriate concrete *technologies*, in his case PETs (privacy-enhancing technologies). Table 1 shows examples for four key data protection principles.

Berendt and *Preibusch* suggested applying the design-pattern idea also to discrimination avoidance.[84] Table 2 sketches the application to principles a)–c) in Section VIII.1 above. The literature is by now too large to fit a representative selection of concrete methodologies into the table cells. The reader is referred to overviews in books[85] and encyclopedias[86], and they should ideally consult the most recent surveys in this rapidly-evolving field. Further principles (such as d)–g) above) should be transformed into design strategies, design patterns and technologies in future work.

However, designers must not forget that the 'divide and conquer' approach of design patterns cannot guarantee fair systems. As the example described by *Schaar* has illustrated, too much attention to detail in a privacy impact assessment can produce a system that, even if the best privacy-enhancing technologies are used, as a whole violates core values of data protection.[87] Just like transparency is not everything, design guidelines are not everything. A holistic view needs to complement the attention to detail.

---

[83] *Hoepman* (Fn. 73); see for an extended description *Danezis*/*Domingo-Ferrer*/*Hansen*/*Hoepman*/*Le Métayer*/*Tirtea*/*Schiffner*, Privacy and Data Protection by Design – from Policy to Engineering, https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design, 2014.

[84] *Berendt*/*Preibusch*, Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass, Big Data, 5 (2), 2017, 135–152.

[85] *Barocas*/*Hardt*/*Narayanan* (Fn. 67).

[86] *Ruggieri*, Algorithmic fairness, in: Comandé (Ed.), Elgar Encyclopedia of Law and Data Science. Cheltenham (UK) 2022; *Berendt* (Fn. 10).

[87] *Schaar*, Privacy by Design, Identity in the Information Society, 2010, 3(2), 267–274.

Principle → DESIGN STRATEGY	*Design pattern*	*Privacy-enhancing technology*
Data minimisation → MINIMISE	'select before you collect', 'anonymise and use pseudonyms', 'collect anonymous data if possible'	Anonymity metrics and anonymisation techniques, anonymous communication, …
Unlinkability, Purpose limitation → SEPARATE	"No specific design patterns" (*Danezis* et al. [Fn. 83]), but can be realized via database design: partitioning, distribution, k-anonymity, …	Specific methods for these patterns, e. g. *Jiang/Clifton*[88]
Transparency → INFORM	P3P (or rather: improved versions of its basic idea)	e. g. TILT (*Grünewald/Pallas*[89])

Table 1: From legal principles to technology choice: examples from data protection (based on *Danezis* et al. [Fn. 83]; *Hoepman* [Fn. 73]; with additions).

Principle → DESIGN STRATEGY	*Design pattern*	*Privacy-enhancing technology*
Bias and discrimination prevention & detection, Harms recognition → PREVENT-D	*Prevent* bias/discrimination	*Specific methods* (cf. *Ruggieri* [Fn. 86])
→ DEMONSTRATE-ND	*Detect* bias/discrimination	–"–
Feedback loop interruption → BREAK-LOOPS	Use data-collection procedures and machine-learning algorithms that avoid pernicious feedback loops	e. g. reinforcement learning (*Ensign* et al. [Fn. 32])

Table 2: From legal principles to technology choice: examples for the avoidance of algorithmic discrimination in future AI regulation.

[88] *Jiang/Clifton*, A secure distributed framework for achieving k-anonymity. VLDB Journal, 2006, 15(4), 316–333.

[89] *Grünewald/Pallas*, TILT: A GDPR-aligned transparency information language and toolkit for practical privacy engineering, in: ACM Conference on Fairness, Accountability, and Transparency, New York City (NY) 2021, 636–646.