CODE OF ETHICS

for Data-Based Value Creation

CONTEXT

www.data-innovation.org

data innovation alliance

OVERVIEW

The “Code of Ethics for Data-Based Value Creation” consists of the following documents: 1) Overview; 2) Basics; 3) Recommendations; 4) Implementation; 5) Context.

The **CONTEXT** document provides in-depth information on data ethics in general and on the relation of the Code of Ethics to other codes (and includes relevant references). The Code is available in German, English, French, and Italian; the Context document is available only in English.

IMPRINT

The “Code of Ethics for Data-Based Value Creation” was drawn up by the “Data Ethics” expert group of the Swiss Alliance for Data-Intensive Services. Editorial team: Markus Christen, Christoph Heitz, Tom Kleiber, Michele Loi (lead editor). French translation: Jean-Gabriel Piguet. Graphics: Ana Nicolasa Caduff. Status: 2020.

ISBN 978-3-9522703-3-2; www.data-service-alliance.ch/codex

TABLE OF CONTENTS

1. OUR VALUE FRAMEWORK IN COMPARISON TO OTHERS

1.1. COMPARISON TO A GLOBAL REVIEW

How does our code relate to the other existing codes? In order to address this question, we begin by comparing the Code with a recent wide-scope, global review of guidelines on AI; importantly, this review defined AI in sufficiently broad terms that more general guidelines about the use of big data were included (1). This analysis of 84 guideline documents highlights 11 distinct main clusters of values found across the entire corpus (where no single value is found in every document):

1. TRANSPARENCY
2. JUSTICE & FAIRNESS
3. NON-MALEFICENCE
4. RESPONSIBILITY
5. PRIVACY
6. BENEFICENCE
7. FREEDOM & AUTONOMY
8. TRUST
9. SUSTAINABILITY
10. DIGNITY
11. SOLIDARITY

These values can be partially mapped in the six values of our code as follows.

1. HARM AVOIDANCE

The value cluster designated 'harm avoidance' is intended to ensure that the data-based product or service, while creating the intended values for clients and provider, does not harm individuals. An equivalent expression for this concept is 'non-maleficence', used in the aforementioned analysis by Jobin, Ienca and Vayena (1). Sometimes it is impossible to produce benefits for a person or group without harming that person/group (or others) in some way. The difference between benefits and harms ought to be positive (i.e., there should be a net benefit). Non-maleficence is only a prima facie principle. An action may avoid harm and yet be morally wrong. An action may also cause some harm and yet be morally right if it generates a greater and more important benefit. In weighing different harms and benefits, ethics requires paying attention especially to harm when:

1) the collateral harm violates a human right or fundamental right of an individual;
2) it significantly hinders autonomy or justice (see below).

So, it is often impossible to interpret the harm principle without considering rights and fairness. Human rights and fundamental rights are quite commonly mentioned in ethical guidelines for AI documents. Yet, rights are not mentioned as overarching ethical principles in the analysis by Jobin et al. In our framework, we link rights to the justice value cluster.

The 'harm avoidance' value cluster also includes some values identified as independent value-clusters in the analysis by Jobin et al. We describe the value-cluster of 'harm avoidance' as including the following values: security, safety, privacy, trustworthiness, sustainability. Here we offer an a priori explanation of the relations among these values.

By promoting security, safety, and privacy one prevents individual harm. Jobin et al. include security and safety in the non-maleficence value cluster too. Unlike us, they consider privacy a distinct value. Another value mentioned by Jobin et al. is sustainability; we include sustainability in the harm prevention value cluster because a plausible definition of sustainability implies that if a product is sustainable, it does not damage the environment in which it operates (also in the long term).

We also include trustworthiness in this value cluster. Of course, this fails to capture the full extent to which trust, writ large, is a value, as it encompasses only a portion of that concept: this is the idea that for a service to be trustworthy, its use produces no harm for the intended user. Furthermore, if by "trustworthy" it is mean that the product or service reliably achieves the goal for which it has been designed and publicized, users are not induced to make errors and cause harm when they deploy it. Further, trust decoupled from trustworthiness presents a real risk: misplaced trust (e.g., the trust of low-competence users who cannot fully understand a data-driven service or assess its trustworthiness), may harm a user or at least lead her to achieve a suboptimal result. It is, of course, still the case that a relationship of trust is valuable independent of its role in achieving harm avoidance; indeed, trusting relationships are characterized by their being, to some degree, resistant to harm between the trusting parties.

Notice that when providing a service to clients and applying the harm avoidance principle, one must consider the net balance of benefits and harms. The balancing of different values is a general methodological principle. When benefits and harms are distributed to distinct people, considerations of justice are invoked.

2. JUSTICE

Justice is the value that deals with the fair distribution of harms and benefits from cooperation, or from a policy. This value cluster includes discrimination avoidance, fairness, respect for legal and moral rights, and solidarity.

Fairness is particularly salient when

a) some individuals are benefited and other are harmed (or benefited less), through no fault of their own;
b) a measure violates rules agreed in advance by all parties;
c) a measure violates legal rights, human rights, or moral rights (these being rights that people ought to have, according to a reasonable ethical theory, even if they are not yet reflected in positive law).

A product is not fair if it discriminates among individuals on arbitrary grounds. Ethically speaking, what qualifies as arbitrary grounds varies from context to context – specifying what grounds count as arbitrary is itself an ethically sensitive decision. Even sex, a legally protected category, does not count as an arbitrary ground when it is reasonably related to the purposes of a selection (e.g., a film script may call for an actor of a particular sex for a certain role; in this situation applications from actors of the opposite sex are not considered).

In the case of data-driven prediction, identifying a discriminating ground as morally arbitrary can be challenging. The problem is often caused by the automated discovery of features that improve the accuracy of selection. Since they contribute to predictive accuracy, computer scientists and statisticians may regard these features as morally non-arbitrary grounds statistically justifying the unequal treatment of individuals. But now suppose that the same features are also statistical proxies for features protected by anti-discrimination legislation, such as sex (20). In such cases, some would argue that discriminating according to the feature is morally arbitrary, and the case for this moral judgment seems stronger if a) the feature, while correlated with something of value, is not a cause of value in that particular business context (e.g., it does not reliably or robustly cause working ability, creditworthiness, purchasing power, etc.) or b) the fact that the feature is a proxy for a disadvantaged group is sociologically explained by the group's social disadvantage that is deemed unfair or anyway problematic (e.g., if a feature selects for women via selecting for lower incomes: supposing that women, on the whole, have lower incomes than men because of known social phenomena involving unfairness), or in other words reproduces existing injustice.

Some of the most sophisticated analyses in computer science try to sort out problematic from non-problematic uses of proxy features by analysing the nature of the causal path from the protected category (e.g. sex), the proxy variable, and the business-relevant variable (21,22). When a data-driven process leads to indirect discrimination (see below), the nature of the association between the target (relevant) variable and the indirectly selected group should be scrutinized so as to determine whether it is justifiable to use a variable that is both predictive of and a proxy for a variable protected by anti-discrimination legislation to make decisions.

3. AUTONOMY

Let us now consider the autonomy value cluster. The value of autonomy promotes the freedom of the individual to make decisions about his or her life when this does not cause harm to others. It counts against data-driven products that steer individual behaviour, depriving individu-als of control. The autonomy value cluster includes freedom, privacy, dignity, and liberty-related civil and political rights. Without freedom (from the overwhelming authority or power of another person) an individual cannot pursue the good life as he/she sees it (autonomy). With-out privacy, people fear the opinions of others; those lacking privacy are less free to pursue their own conception of the good life. A person cannot have dignity if her capacity to choose and pursue her own conception of the good life is not recognized by those around her. Hence, having some degree of meaningful autonomy, and the concomitant recognition of those qualities that make such autonomy appropriate for persons, is a necessary condition of dignity. Many civil and political rights are designed to protect individual autonomy in particular realms, for example, freedom of conscience, freedom of speech, political freedom, and reproductive freedom.

1.2. HOW DIFFERENT GUIDELINES MAP THE VALUES

From the analysis of other guidelines, it is apparent that there is no unanimous agreement on 1) how many fundamental values exist, 2) what the fundamental ethical values are, or 3) the definitions of such values. Hence, the values, or value clusters, which are used as headings for different practical guidelines, do not provide a suitable compass for locating actual prescriptions in the text or comparing the contents of different guideline documents.

Consider, for example, the justice value cluster, which includes fairness and non-discrimination, and its implications for data science – specifically, the issue of algorithms being especially biased (e.g., more prone to make harmful errors) against certain individuals or social groups. A detailed analysis of 20 sets of guidelines revealed that very similar prescriptions are grouped together by using different labels and a highly heterogeneous value terminology (including bias, fairness, unjust impact, non-discrimination, avoiding discriminatory outcomes, inclusion, harmful stereotypes, submission, marginalization, gender balance, gender equality, barriers, violations of rights, and disparate impact principles). Someone who needs to compare what different guidelines have to say about the value of justice therefore may have to search for a wide variety of value keywords: control (23), fairness (3,10,12,13,17,18), discrimination (5–7,10,16,19,24), bias (4,5,11,23), equality (5,16,25), inclusion (5), trust (25), and robustness (11).

1.3. THE SCOPE OF THE CODE VS. OTHER GUIDELINES

One significant difference between our Code and the large array of other ethical codes available is the scope of its prescriptions and guidelines. The scope of the Code is more limited, compared to some other guidelines, in terms of the stakeholders it addresses. The Code is not written by and for governmental agencies, inter-governmental and supra-national organizations, non-profit organizations, professional associations and scientific societies and non-profit organizations, so unsurprisingly it includes only measures addressing companies delivering and utilizing data-driven products. Some of its prescriptions can also be extended to government bodies involved in bureaucratic decision-making, but that is not the audience for which we wrote these guidelines. The Code does not aim to direct or inform governments with respect to policies requiring the power to enact law. Measures that can be enacted only by legislators are outside the scope of this Code.
Hence, the Code consists of technical solutions and organizational solutions; it requires technical infrastructure usually available to machine learning specialists operating within firms (or hired by firms) and a willingness to plan and adopt innovative organizational solutions. We compared the guidelines of our Code with guidelines addressing different sets of stakeholders and, in spite of the different audiences, we found a significant (though incomplete) overlap.

2. A COMPARISON OF PRACTICAL RECOMMENDATIONS IN OUR AND OTHER GUIDELINES

We shall now discuss the practical recommendations given by our guidelines and other sets of guidelines. Let us begin with the methodology of our analysis. Both the AI Ethics Guidelines Global Inventory, compiled by the NGO AlgorithmWatch (26), and the list compiled by Jobin et al. (1) include just over 80 guidelines. We selected for this study the EU Trustworthy AI and 19 other sets of guidelines (that is, a subset of the guidelines identified by Jobin et al. (1)) in the order they appear on a third-party website listing the entire set given in Jobin et al. (1) (to limit observer bias). This 20-item set includes guidelines that are heterogeneous with respect to both the issuing stakeholder and the stakeholders addressed by the guidelines: companies, think tanks, communities of researchers and practitioners, private sector alliances, professional associations, non-governmental organizations (NGOs), information commissioner offices (ICOs), inter-governmental organizations (IGOs), non-profit organizations (NPOs), charities, miscellaneous entities (e.g., mixed academic, NPO), governmental agencies/organizations, and federations/unions.
In what follows we review the contents of the Code and compare it to both the summary analyses of other guidelines in reviews by other scholars and to the guidelines in the other 20 documents examined in the preparatory study undertaken for this addendum.

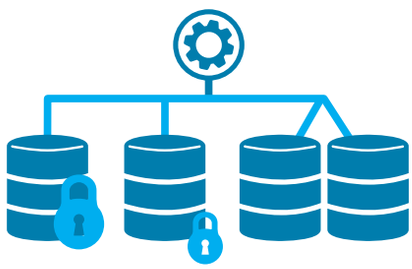
2.1. DATA GENERATION AND ACQUISITION

Most guidelines highlight the value of **transparency** (8, 9, 23) in relation to this phase of the data pipeline. Examples of guidelines for transparency are:

- 'How is the data collected, and from whom? Directly from the consumer or from third party sources about the consumer?'(7)
- '[...]Build out harmonized standards for Data labelling. All companies will benefit from greater transparency requirements around licensed datasets. This will be particularly important for startups/ smaller companies who are not resourced to undergo extensive testing prior to release' (19).
- '[Ask:] "[i]s the data collected in an authorized manner? If from third parties, can they attest to the authorized collection of the data?"'(7)
- 'The aggregation and use of customer data – especially in AI systems – shall always be clear and serve a useful purpose towards our customers' (23).
- 'Organizations, including governments, should immediately explore, test, and implement technologies and policies that let individuals specify their online agent for case-by-case authorization decisions as to who can process what personal data for what purpose' (8).

The Code intentionally avoids overlaps with data protection principles. Since transparency is required by data protection, and we want to avoid overlaps with such law, we focus on the aspects of transparency that are less clearly specified in the law. We also stress the importance of nudges and we did not find the issue of nudges being addressed specifically in the 20 guidelines analysed here. We also stress the importance for each company of demonstrating the ethical quality of its data collection practices. Unlike other guidelines, we stress (in the accountability value cluster) the idea that the **accountability** of data-driven services can only be achieved for a given firm if it is achieved for the entire data ecosystem: each company needs to provide information to the other companies with which it is interdependent.

2.2. DATA STORAGE AND MANAGEMENT



Let us now turn to our prescriptions concerning data storage and management. We found the following overlapping guidelines:

- 'Prohibition on Secret Profiling. No institution shall establish or maintain a secret profile on an individual' (18).
- Guidelines requiring governance processes for all actors involved in the data pipeline (4,6). We did not, however, find any guideline addressing the need for a plan for managing inactive accounts, including those from dead people.
- We did not find recommendations requiring the facilitation of data portability and enabling the reuse of data by customers. The most relevant content was found in a guideline concerning the issue of employee data: 'Workers should have the right to access, manage and control the data AI systems generate, given said systems' power to analyse and utilize that data' (17).
- We found that most guidelines (but not all)[1] at least mentioned the value of privacy, and some also explicitly mention security (23) and cybersecurity obligations (18).

2.3. DATA ANALYSIS AND KNOWLEDGE GENERATION

Let us now turn to another stage in the data pipeline: data analysis and knowledge accumulation. To realize the ethical value of **harm prevention** our code prescribes assessing foreseeable, harmful uses of the knowledge produced via data analysis. Of course, this prescription must be understood as being modest in scope, as assessing all distant consequences of any form of knowledge derived from data is impossible and the ethical duty would not meet any reasonable feasibility test. This part of our Code addresses primarily data scientists – those who carry out this phase in the data science pipeline – and the methodologies that can be implemented in this step of the data pipeline. The emphasis, that is, is on those problems in the use of models that it is possible to foresee and study and predict with the tools of data science, during the process in which models are built and then tested, taught with data (e.g., with machine learning techniques), and optimized.

Indeed, it is now primarily the communities of concerned data scientists who are advocating a broader understanding of their responsibilities, tools, and skills, and providing new tools designed to facilitate the implementation of ethical purposes (such as privacy and fairness 'by design'). For example, consider the principle of accuracy in the FAT-ML guidelines (FAT-ML is a community of practitioners of machine learning) (12), which says:

- 'Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures' (12).

[1] For instance, the FAT-ML principles and implementation recommendations are focused on a different set of problems and do not mention privacy (12).

Some guiding questions in this document include: 'What are realistic worst-case scenarios in terms of how errors might impact society, individuals, and stakeholders?' Examples from other guidelines include:

- 'Responsible Design and Deployment: We recognize our responsibility to integrate principles into the design of AI technologies, beyond compliance with existing laws. [...] As an industry, it is our responsibility to recognize potentials for use and misuse, the implications of such actions, and the responsibility and opportunity to take steps to avoid the reasonably predictable misuse of this technology by committing to ethics by design' (11).
- 'As part of an overall "ethics by design" approach, artificial intelligence systems should be designed and developed responsibly [...] in particular by [...] assessing and documenting the expected impacts on individuals and society...for relevant developments during its entire life cycle' (6).
- Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices' (18).
- 'Leaders, designers, and developers of ML systems are responsible for identifying the potential negative human rights impacts of their systems' (19). (This guideline is also relevant for justice, as the violation of most human rights can be considered a form of injustice.)

The EU Guidelines for Trustworthy AI provide an assessment list related to what we call 'foreseeable misuse', which is linked to autonomy as well as to harm avoidance:

- 'Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?[2]...Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?' (10)

Second, we focus on protecting privacy by preventing de-anonymisation. This is also a concern found in other guidelines. For example, the guidelines by Privacy International and Article 19 emphasize that:

- 'Some applications of AI, in particular uses of machine learning, blur the line between personal and non-personal data [or personally identifiable information (PII) and non-personally identifiable information (non-PII) [in the US], around which data protection and privacy laws around the world are organised. Data that is initially non-personal (non-PII) can become personal data (PII) in a different context or in different points in time, which is a particular risk for sectoral regulations. A similar challenge applies to sensitive personal data. Profiling using machine learning can derive, infer, or predict sensitive information from non-sensitive data, which might undermine additional safeguards for sensitive personal data' (13).

[2] This, unfortunately, is so vague as to be quite useless as a guideline. For instance, does targeted advertising affect human autonomy by interfering with the (end) user's decision process (e.g., to vote for a particular candidate) in an unintended way? Arguably not, but only because the effect of the interference is wholly intended. It is not clear if the fact that the answer to this item on the checklist is 'no' makes that type of interference with human autonomy ethically acceptable for trustworthy AI.

One prescription requires the building of **control** systems to assess the consequences of foreseeable errors in the models. Our prescription involves two design requirements: 1) services should be able to record the information needed to evaluate whether the service is being used appropriately or in an ethically problematic way; 2) services should enable the communication of the relevant feedback to designers, with the aim of improving the design process for the next iteration of service design. Similar suggestions are found in other guidelines:

- The IEEE guidelines and others (e.g., 17) recommend the 'secure storage of sensor and internal state data, comparable to a flight data recorder or black box' (9).
- The FAT-ML guidelines recommend that data scientists '[d]evelop a process by which people can correct errors in input data, training data, or in output decisions' (12). This guideline overlaps with the value of autonomy – it endows the recipients of algorithmic decisions with powers to respond to those decisions, achieving a higher degree of self-direction.

As the last item of harm prevention, our Code includes a prescription to 'consider imposing limits to the software distribution'. The idea here is that certain data-driven products, in particular AIs, may be used for harmful purposes, such as generating fake news (27) or coordinating and launching cybersecurity attacks (28). This prescription is found in one set of the guidelines studied (28).

One of our prescriptions related to the value **justice** concerns the assessment and documentation of indirect discrimination. As already mentioned, many guidelines and declarations claim that AI must be fair – that it should avoid discrimination and harmful biases, especially those that are disadvantageous for specific groups. Indeed, most guidelines fail to draw clear distinctions among these concepts and lack rigorous definitions of the relevant terms. In our guideline, we refer to the European legal (and philosophical) concept of indirect discrimination (20,29–31), which is roughly equivalent to the US legal concept of disparate impact (32). Why such focus on indirect (as opposed to direct discrimination)? The reason is that direct intentional discrimination is easily avoided by big data approaches, and yet it is unclear that the resulting models are unobjectionable from the perspective of the deep moral and political rationales underpinning anti-discrimination law – namely, to ensure fairness and protect vulnerable individuals and communities (31).

The problem is in the nature of statistical decision-making, which, when driven by machine learning, identifies properties as predictive of good or bad outcomes based on observed correlations. These properties can be, in turn, highly correlated with membership in protected groups.

Some arguments against algorithmic bias can be explained as follows: learning to make a prediction via a feature that is merely a proxy for the valuable trait of interest (as opposed to the trait itself, e.g., creditworthiness or productivity) but also a statistical proxy of a protected variable (e.g., gender or race) amounts to learning a human stereotype or prejudice about gender or race. One familiar example prior to the big data and AI era is redlining, the following phenomenon. If postal code is used to predict credit risk, entire areas may be deemed ineligible to borrow, or may receive loans only with very high interest rates. The reason is that, in such areas, the proportion of insolvent credit is higher, which is a statistical signal suggesting that lending to people in those area will be unprofitable. Clearly, postal code is not what lenders are directly interested in; nor is it what causes people to be solvent

or insolvent. It is a mere proxy. But it is also a proxy that in certain societies aligns with race consistently. One could argue that an algorithm that learns the association between credit risk and postal code has merely learned to associate race with creditworthiness through a common statistical proxy for both (postal code).

The example of redlining can be considered a paradigm for the kind of discrimination that is likely to be encountered in the field of big data. Direct discrimination (that is, discrimination in which race or gender variables are factors directly affecting the decision) is likely to be avoided a) because data scientists are typically instructed to avoid using such data and also to avoid violating anti-discrimination law and b) because direct discrimination can easily be avoided, since many other data sources (including those statistically correlated with race) may be available. This enables data scientists to avoid direct discrimination against a protected group even when such direct discrimination would be economically rational in the sense that it would constitute statistical discrimination (i.e., it would contribute to the accuracy of a statistical prediction).[3] The problem is that many statistical proxies for a given variable can be highly correlated with (for instance) ethnicity or gender. Thus, it is not surprising to find that this problem is addressed by the majority of guidelines we have analysed here, even when the terminology to describe these issues is not used consistently.

It should be obvious that mitigating or eliminating indirect discrimination can be a morally controversial and even politically divisive issue. For example, consider the debate in the United States about affirmative action in college and university admissions. Affirmative action can be seen as an attempt to redress the inequality in standardized test scores – a statistical proxy for academic excellence – in the direction of racial equality. Some interpret affirmative action as an attempt to compensate for some kind social inequality – for instance, cultural differences in families and unequal access to test preparation courses and therefore a restoring a form of fairness that the process has lost due to arbitrary social influences.[4] Others regard it as imposing a kind of equality of outcome at the expense of procedural fairness.

The laws of most countries do not provide clear criteria for determining when indirect discrimination (in the sense that entails a violation of anti-discrimination law) occurs (38). Fairness is a vague ethical orientation; people (even reasonable and well-informed people) often disagree about what is fair, and philosophical and political conceptions of social justice are historically some of the most controversial ones. Some interpretations of Marx (39,40) and some libertarian views (41) even

[3] Statistical discrimination is a subset of discrimination and is typically considered rational in economic terms from the point of view of the discriminator – but (mostly) harmful for members of certain groups (33,34). Not all forms of discrimination are statistical, and not all are rational (in economic terms). For example, Gary Becker de-fines taste-based discrimination, which is discrimination that satisfies an intrinsic preference against or in favour of members of a group (qua members of that group) in the absence of any economic rationale for the discrimination. This discrimination may be deemed irrational in the sense that it typically has adverse economic conse-quences not only for members of the group discriminated against but also for the discriminators (35). As algorithms do not have tastes or intrinsic preferences, the kind of discrimination data-driven models are most likely to produce belongs to the statistical variety. This includes discrimination reflecting the tastes and preferences of individ-uals other than the firm's owner or manager: for instance, if customers dislike being served by employees of a given race or gender, or employees from the majority group have a tendency to work less efficiently when paired with individuals from a minority group, such tastes may be reflected in the variables of interest (profit, creditworthiness, etc.) – thus we are back to a case of statistical discrimination, which data-driven methods are likely to produce (unless designed otherwise).

[4] This is not the way in which affirmative action is defended in courts in the US. The winning case for affirmative action has been the one built on the value of the diversity of the student body as a whole for the student body as a whole. The best philosophical/legal formulation of this argument is found in (36,37). This, however, may be consid-ered, not so much a fairness argument, but instead a utilitarian argument treating diverse applicants as means for the education of other students, as opposed to ends in themselves.

entail that social justice is a 'pseudo concept',[5] and other libertarian views argue that justice is fully reducible to ensuring respect for voluntary (i.e., uncoerced) contracts entered after an initial legitimate acquisition of resources (43). People with such views may not accept most of the arguments behind the goal of "fairness in machine learning".

It is no surprise, then, that in the guidelines we examined, there was no consensus and little clarity on the degree to which considerations of social justice should figure into the evaluation criteria intended to determine whether an algorithm is biased or unfairly discriminatory. Clearly, from the point of view of some political doctrines, fairness is pseudo-talk or irrelevant, unless it relates to direct or indirect discrimination explicitly prohibited in the law. Entrepreneurs have the moral right to use their own legitimately acquired resources in any way that pleases them, and they have the right to use even biased systems which are legal. This view enables indirect discrimination to occur, since the legal tort of indirect discrimination is in many jurisdictions significantly harder to assess than the one of direct discrimination (38).

Therefore, the Code relies on the presumption that the value orientation of justice involves, at a minimum, a certain transparency about the ways benefits and harms are allocated by an algorithm, in particular with respect to historically marginalised groups. It also mentions attempts to correct indirectly discriminatory biases by, for instance, using technical methodologies with sound moral and/or legal underpinnings. Any debate about whether and how these are used should be transparent and open to challenges by the public. There is no one-size-fits-all prescription for automating fairness; the discussion in the machine learning literature shows that some intuitively appealing statistical standards of fairness cannot all be realised simultaneously in the most commonly occurring social circumstances (44–46).

Requirements to document and achieve fairness (understood as the fair distribution of advantages and disadvantages produced by the implementation of algorithmic rules) are found in many other guidelines we examined (3,4–7,10-13,16,17,18,19,25). The impossibility of achieving different conceptions of fairness simultaneously, and the contextual sensitivity of the relevant fairness metrics and bias-mitigation goals is also recognized explicitly by some guidelines (those addressing discrimination issues specifically). In some cases, the recommended solution is to engage different stakeholders to obtain some kind of intersubjective agreement about relevant and adequate fairness metrics. Another issue discussed in the Code, related to the value of justice, is the fact that different error rates for different groups may entail that advantages and disadvantages are also distributed differently, statistically speaking, among different groups. This is sometimes addressed in guidelines under the heading of fairness or discrimination, even if it is not exactly the same problem as that of indirect discrimination (indirect discrimination also occurs when the rate of incorrect predictions is the same for all groups). In the computer science literature, inequality in error rates is described as disparate mistreatment (47). Both fairness issues – indirect discrimination and disparate mistreatment – are mentioned in some of the guidelines we examined.

[5] But see Hayek's odd endorsement of Rawls's theory of social justice in (41; for a summary see 42).

For example, guiding questions for both indirect discrimination and disparate mistreatment include:

- 'Have we evaluated the veracity of the data and considered alternative sources? Have we mapped and understood if any particular groups may be at an advantage or disadvantage in the context in which the system is being deployed? Have we calculated the error rates and types for different sub-populations and assessed the potential differential impacts? Have we applied rigorous pre-release trials to ensure that [the ML system] will not amplify biases and error due to any issues with the training data, algorithms, or other elements of system design?' (19)
- 'What is the potential damaging effect of uncertainty / errors to different groups? [...] Calculate the error rates and types (e.g., false positives vs. false negatives) for different sub-populations and assess the potential differential impacts' (12).

Guiding questions for the contextual nature of fairness are:

- 'Have we sufficiently researched and taken into account the norms of the context in which the system is being deployed? Have we identified a definition of fairness that suits the context and application for our product and aligns with the International Declaration of Human Rights? Have we included all the relevant domain experts whose interdisciplinary insights allow us to understand potential sources of bias or unfairness and design ways to counteract them? Have we mapped and understood if any particular groups may be at an advantage or disadvantage in the context in which the system is being deployed?' (19)
- 'In determining the need for applying these disparate impact principles, companies should focus on the groups being protected, the context of use, the nature of the application of the data analytic system and the potential for substantial harm' (7). Note that 'disparate impact' is US legal terminology for indirect discrimination.
- 'Avoidance of unfair bias. [...] The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation' (10).
- 'Fairness: People involved in conceptualizing, developing, and implementing machine learning systems should consider which definition of fairness best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics' (19; italics added).
- 'There are many different ways of defining fairness; people involved in conceptualizing, developing, and implementing machine learning systems should consider which definition best applies to their context and application. [...] Guiding Questions: [...] Have we identified a definition of fairness that suits the context and application for our product and aligns with the International Declaration of Human Rights?' (19)

One aspect of fairness is data diversity. The World Economic Forum recommends that firms:

- 'Determine whether certain data sets fit internally agreed upon standards of "adequate" and "representative" data (looking to both quantitative and qualitative metrics); identify opportunities to expand data collection efforts where contextually appropriate, viable, and possible to do so without violating privacy' (19).

But the concept of 'adequate' data is both vague and problematic. One set of guidelines we examined even claims that using machine learning methods on data should '[e]xclude data that is not relevant to predicting the outcome' (7). One of the important discoveries made from using advanced machine learning techniques, such as neural networks, is that the human ability to determine a priori which data is relevant to predicting an outcome is limited in comparison to the ability of a machine using a general-purpose statistical technique to make the determination. It is therefore unclear whether any data can be ruled out as irrelevant prior to analysing data with machine learning techniques. What that guideline appears to suggest is the exclusion of some data based on a priori ethical (not statistical) grounds – wanting results to be free from the influence of a certain type of data, irrespective of its predictive value (48). The idea of checking the diversity, and/or representativeness, of data inputs, as well as the sources of error (12), is found in different guidelines (12,16,19). Some guidelines explicitly require companies to be transparent about the features of individuals used to make predictions: '[o]rganizations should be prepared to communicate to outside parties the key factors that go into their scores' (7); '[w]orkers should have the right to access, manage and control the data AI systems generate' (17).

Let us now turn to the value orientation of **autonomy.** We interpret this value at this stage of the data pipeline as the act of enhancing the agency of the people affected by the algorithms. This can be achieved by engaging stakeholders in a discussion about the values and goals assumed by the model. This idea is also found in other guidelines; for example, the World Economic Forum white paper on discrimination by machine decisions states:

- 'The development and design of ML applications must actively seek a diversity of input, especially of the norms and values of specific populations affected by the output of AI systems' (19).
- 'Talk to people who are familiar with the subtle social context in which you are deploying. For example, you should consider whether the following aspects of people's identities will have impacts on their equitable access to and results from your system: Race, Sex, Gender identity, Ability status, Socio-economic status, Education level, Religion, Country of origin' (12).

Stakeholder engagement in all its different forms is invoked by the TENETS guidelines by Partnership for AI (15), which mentions some form of stakeholder engagement in principles 2, 3, 4, 5, and 8 in TENETS. Stakeholder engagement is expected to play a role in relation to each of the following.

- Providing feedback on the focus of ethical inquiry, i.e., are companies and/or auditors identifying all the relevant risks and vulnerabilities? (15,23)
- The need for open, interdisciplinary research (15), in particular to identify and bring together different skills and competences (3).
- Collecting the interdisciplinary competences required to identify potential biases and forms of discrimination (11,13,16). According to some guidelines, input from stakeholders should be sought in order to 'identify the entire range of data types necessary to adequately train an [sic] ML in a given context' and 'understand how to appropriately source the data needed' (19).

– Knowing the norms and values of the data subjects or populations affected by AI-driven decisions (3,19).
– Identifying the different stakeholders impacted by AI research (15).
– Identifying domain-specific concerns (11,15,19).
– Avoiding fears and confusions regarding AI (9).
– Promoting new forms of governance that include 'various stakeholders' such as 'civil society, government, private sector or academia and the technical community' (14,15).

To sum up: stakeholder engagement can be viewed not only as a way to promote the ethical orientation of autonomy, but also as a way to realize in practice the requirements of the procedural values of control, transparency, and accountability, to which we now turn.

Controlling the knowledge generation process is related to the capacity to offer (scientifically grounded) justifications for believing (and thus, conscientiously using) the model's output. These justifications include not only estimates of its accuracy but also estimates of its robustness (that is, the capacity of the model to maintain accuracy in the face of minor perturbations in the context of use).

To enhance accuracy and robustness, the Code requires that those responsible for data analysis and knowledge generation 1) justify the appropriateness of the algorithmic techniques they use, and 2) ensure the traceability of the process that starts with using specific datasets (that may be marred by errors, biases, etc.) for the training of machine learning models.

Similar requirements are found in other guidelines. We may refer to this activity as documentation. We have found guidelines prescribing documentation of the training 'goal'[6], the algorithm used (7), the accuracy of a model (10,12), and the reliability and reproducibility of its decisions (10). In the Trustworthy AI guidelines the goal of this documentation activity is called 'traceability', and it is required as an element of transparency (10). In the IEEE document, roughly the same activity is considered an aspect of designing for accountability:

– 'Manufacturers/operators/ owners of A/IS should register key, high-level parameters, including: Intended use; Training data/training environment (if applicable); Sensors/real world data sources; Algorithms; Process graphs; Model features (at various levels); User interfaces; Actuators/outputs; Optimization goal/loss function/reward function' (9).

Let us now turn to **transparency.** The first requirement of transparency for the knowledge derived from data analysis concerns the intelligibility of the predictions or decisions reached by a model. In other words, transparency is achieved by providing some kind of explanation. There is no one-size-fit-all definition of what an explanation should be. Most plausibly the relevant concept of explanation is contextual: an explanation describes the relation between a model and its predictions or decision in a way that is understandable at least by data scientists and people who are expected to use the model (for informing decision-making, perhaps), and, ideally, the people affected by it. The discussion about the intelligibility of AI is a vast inter-disciplinary one, involving academia, industry, and other stakeholders,

[5] We use the vague term 'goal' here to indicate different elements in the training process that are all directly related to the purpose for which a system is trained. Formally, one may identify the goal as the target variable, or in machine learning terms, the true label, as in the prescription 'define what the software is intended to predict' (7). It may also be referred to in terms of 'optimization goal/loss function/reward function' (9).

and it cuts across the disciplines of computer science, mathematics, philosophy of science, law, and others – so it cannot be adequately summarised here. It will suffice to say that we can distinguish at least two quite different conceptions of the goals of explanation or intelligibility.

On the first conception, one makes a system intelligible by clarifying its purposes (e.g., to develop adequate ability to correctly classify cat and dog image) and providing sufficient statistical evidence that the purposes will be achieved in the real world setting in which the model will be used. This type of intelligibility overlaps substantially with what we have defined above as documentation, an activity realising the procedural value of control. We see this supported in guidelines that require documentation of one or more of the following:

– The algorithm's generic goal and purpose (10), in the sense of 'intended use' (9);
– the algorithm's mathematical goals, in the sense of its 'optimization goal/loss function/reward function' (9);
– the features used to train an algorithm/make decisions (3);
– the weightings of these features (if known) (3);
– the algorithm type, the extent of its opacity (3,12);
– the different performance metrics (3,12);
– the procedure of validation (3,12).

On this conception, you make the logic of a model intelligible by revealing the purposes you wanted to achieve with it and the methods you used to achieve your purposes (e.g., techniques such as machine learning). Similar ideas of intelligibility are found in the document by Women Leading in AI (3).

The other conception of intelligibility focuses on a) building interpretable models and b) interpreting the decisions of models. The first idea – interpretable AI – involves imposing constraints on the model so that it is known in advance that it will be easy to interpret by end users (49). Th e second approach, instead, requires different techniques from the first conception: for instance, reproducing the input-output relations in a black box model with an intelligible model delivering similar patterns of outputs (50), or employing methods to assess the veracity of counterfactual claims about the black-box model (51), such as 'the decision would have been Y if the value of your feature X had been K instead of J'. The psychological effects of these and other kinds of explanations have been tested empirically, including in a fictional scenario in which algorithms are used to make a promotion decision (52), and they seem to enhance the perception of the fairness of employing the algorithm. Others have proposed built-in (in-model) explainers, which are produced by algorithmic techniques from the inner workings of the model (53–56). Some guidelines appear to have endorsed using such techniques, for example:

– the proposal to include a 'why did you do that' button in AIs interacting with humans (9);
– the idea that '[t]he data provided by the black box could also assist robots in explaining their actions in language human users can understand' (17);
– the idea that '[i]n some cases it may be appropriate to develop an automated explanation for each decision' (12).

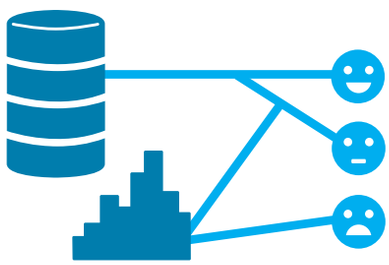
We have not included an explicit endorsement of any of these methods in our Code, because the debate about their validity is on-going and the suitability of any method should be assessed with up-to-date information at all times.

Another requirement of transparency in the Code requires that the limitations, possibility of harmful uses, indirect discrimination features, effects of predictive errors, and measures taken to prevent or mitigate these are both documented and communicated appropriately to end users. Similar requirements are found in the guidelines we examined. For example, the FAT-ML guidelines include this mandate: 'Determine how to communicate the uncertainty/margin of error for each decision' (12).

In our Code, **accountability** does not refer primarily to the activity of documenting the algorithm, but rather, to creating roles and/or assigning responsibilities within organizations to ensure that the ethical orientations and procedural values above find realization in practice. Converging examples in other guidelines include the following:

– 'Establishing demonstrable governance processes for all relevant actors, such as relying on trusted third parties or the setting up of independent ethics committees' (6).
– 'Guiding question: Who will have the power to decide on necessary changes to the algorithmic system during design stage, pre-launch, and post-launch? Initial Steps to Take: Determine and designate a person who will be responsible for the social impact of the algorithm.' (12).

2.4 DEPLOYMENT OF A DATA-BASED PRODUCT OR SERVICE



For **harm avoidance,** our Code includes recommendations for a) preventing the dissemination of potentially harmful products, and, as a necessary monitoring step of the success of this measure, b) examining actual misuse if possible, in particular c) with respect to the risk of de-anonymisation or the inference of sensitive traits from unproblematic personal data. Similar ideas are found in the guidelines examined, for example:

- 'Responsible Design and Deployment: We recognize our responsibility to integrate principles into the design of AI technologies, beyond compliance with existing laws. [...] As an industry, it is our responsibility to recognize potentials for use and misuse, the implications of such actions, and the responsibility and opportunity to take steps to avoid the reasonably predictable misuse of this technology by committing to ethics by design' (11).
- 'Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?' (10)
- 'As part of an overall "ethics by design" approach, artificial intelligence systems should be designed and developed responsibly [...] in particular by: [...] b. assessing and documenting the expected impacts on individuals and society [...] for relevant developments during its entire life cycle' (6).
- 'The capacity of an AI agent to act autonomously, and to adapt its behavior over time without human direction, calls for [...] ongoing monitoring' (4).
- 'Which detection and response mechanisms did you establish to assess whether something could go wrong? Did you verify how your system behaves in unexpected situations and environments?' (10)
- 'Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? [...] Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? [...] Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)? Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? [...] Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?' (10)

Concerning **justice,** our Code focuses on discovering, assessing, and correcting indirectly discriminatory effects resulting from the deployment of the model in practice. This is different from assessing the discriminatory qualities of the models based on test data (as in the previous step of the data pipeline). There may be specific problems (e.g., ethnic groups suffering from a disparate rate of misclassification errors) that are only discovered during model deployment, not during training or testing with historical data. Such disparate harmful effects of using models, which should be prevented, can be both unintended and considered unfair and avoidable on balance. In addition to unfair indirect discrimination and disparate mistreatment (e.g., unequal false negatives), we address in the Code the problem of unintentional stigmatisation and harmful stereotypes, and the exclusion of vulnerable groups.

We also found similar recommendations in the guidelines we reviewed. For example, some addressed indirect discrimination:

– 'Adopt and maintain policies and procedures reasonably designed to collect information sufficient to conduct assessments that would detect any significant disparate impacts, including, if necessary, collecting sensitive information such as race, gender, ethnicity, and religion or constructing accurate proxies for such sensitive information' (7).
– 'Guiding questions: Have we outlined an ongoing system for evaluating fairness throughout the life cycle of our product? Do we have an escalation/emergency procedure to correct unforeseen cases of unfairness when we uncover them?' (19)

Others addressed possible collateral effects of stigmatisation and negative stereotypes:

– 'Support efforts to promote trust in the development and adoption of AI systems with particular attention to countering harmful stereotypes and fostering gender equality. Foster initiatives that promote safety and transparency, and provide guidance on human intervention in AI decision-making processes' (25).

Finally, some addressed social inclusion/exclusion:

– 'Accessibility and universal design. Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards' (10).

Concerning the value of **autonomy,** our Code focuses on ensuring that the knowledge of domain experts and users, especially negative feedback, is considered in the design of the processes, producing a virtuous feedback loop leading to the removal of any problems discovered. We also find similar guidelines concerning this point, for example:

– The FAT-ML guidelines recommend that data scientists '[d]evelop a process by which people can correct errors in input data, training data, or in output decisions' and also '[a]llow data subjects visibility into the data you store about them and access to a process in order to change it.' They also require making 'contact information available so that if there are issues it's clear to users how to proceed' (12).
– The UNI Global Union writes that '[w]orkers should have the right to access, manage and control the data AI systems generate, given said systems' power to analyse and utilize that data' (17).

Control in the application of a data-driven process consists, according to our Code, in working out appropriate ethical design principles and checklists consistent with ethical priorities of this Code and of the company, which should then guide the internal development of a data-driven product for internal use (when the product has been designed by an internal R&D department) or the discussions about adopting a model under consideration for purchase from an external developer. It is based on the idea that ethical principles by themselves do not make organisations more ethical (57). The main idea here is to identify ethics goals that may become salient in the deployment of a service ahead of developing the service, and to use these ethics goals as parameters for the design of the service and as tests for the

service when finally in use. A few guidelines address the design of organisational structures intended to improve the data-pipeline process within companies from an ethical perspective.

- 'Engage stakeholders and domain experts in participatory manner. Best identify what types of considerations should be made for an ML model being applied in a particular domain (industry, geography, population, etc.) to design a fair and contextually appropriate ML model' (19).
- 'Countries need to take proactive steps towards the inclusion of women in the coding and the design of machine learning and AI technologies. The low involvement and marginal inclusion of women in the coding and design of AI and machine learning technologies is leading to a variety of problems, including the replication of stereotypes, such as the submissive role of voice-powered virtual assistants, overwhelmingly represented by women' (5).
- 'Foster initiatives that promote safety and transparency, and provide guidance on human intervention in AI decision- making processes' (25).

According to our Code, transparency at this stage of the data pipeline consists in providing adequate information about compliance with ethical recommendations (such as the ones included in this code) to other stakeholders. The Code maintains that this is a necessary step for the maintenance of ethical data practices at the level of the entire ecosystem. Ideally, a company's claims concerning the ethical qualities of its products and procedures should be externally verifiable, at least by contracted auditors if not by the wider public. Examples of the same general idea are found in the guidelines we examined:

- 'Organizations should publicly describe the model governance programs they have in place to detect and remedy any possible discriminatory effects of the data and models they use, including the standards they use to determine whether and how to modify algorithms to be fairer' (17).
- 'Guiding Questions: Have we openly disclosed what aspects of the decision-making are algorithmic? How much of our data sources have we made transparent? Have we provided detailed documentation, technically suitable APIs, and permissive use of terms to allow third parties to provide and review the behavior of our system? Can you provide for public auditing (i.e., probing, understanding, reviewing of system behavior) or is there sensitive information that would necessitate auditing by a designated third party? How will you facilitate public or third-party auditing without opening the system to unwarranted manipulation?' (19)
- 'Initial Steps to Take: Document and make available an API that allows third parties to query the algorithmic system and assess its response. Make sure that if data is needed to properly audit your algorithm, such as in the case of a machine-learning algorithm, that sample (e.g., training) data is made available. Make sure your terms of service allow the research community to perform automated public audits. Have a plan for communication with outside parties that may be interested in auditing your algorithm, such as the research and development community' (12).

In our Code, the recommendations relating to **accountability** are intended to ensure that the company's top management can take responsibility for the end product in an appropriate manner. Some guidelines we reviewed seemed preoccupied with avoiding the use of algorithmic decisions as a smokescreen behind which to hide human decisions. For example:

– 'For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights: A/IS should always be subordinate to human judgment and control (9). 'A/IS' stands for Autonomous and Intelligent System. Ensuring Accountability Principle: Legal accountability has to be ensured when human agency is replaced by decisions of AI agents' (4).
– 'Identification Obligation. The true operator of an AI system must be made known to the public' (18).

We believe that this is only one special case of failure of accountability, and that accountability should be understood more generally – that is, not in the narrow sense of ensuring meaningful human control (59) (important in the case of autonomous AI but not for many other data-driven services), but in the broader sense of ensuring ethical governance of the overall decision process. This is also a concern found in other guidelines:

– 'Continued attention and vigilance, as well as accountability, for the potential effects and consequences of, artificial intelligence systems should be ensured, in particular by…establishing demonstrable governance processes for all relevant actors' (6).
– 'Recommendation Accountability 3.4. Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS' (9).
– 'Guiding Questions: Who is responsible if users are harmed by this product? What will the reporting process and process for recourse be? Who will have the power to decide on necessary changes to the algorithmic system during design, pre-launch, and post-launch? Initial Steps to Take: Determine and designate a person who will be responsible for the social impact of the algorithm. Develop a plan for what to do if the project has unintended consequences. This may be part of a maintenance plan and should involve post-launch monitoring plans. Develop a sunset plan for the system to manage algorithm or data risks after the product is no longer in active development' (12).

3. REFERENCES

3.1 CITED LITERATURE

1. A. Jobin, M. Ienca, and E. Vayena, “Artificial Intelligence: The global landscape of ethics guidelines,” Nature Machine Intelligence, vol. 1, no. 9, p. 389–99, September 2019. [Online.] Available: http://www.nature.com/articles/s42256-019-0088-2
2. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Arxiv. 2019 Sep;1(9):389–99. Available: https://arxiv.org/abs/1906.11668
3. Women Leading in AI, “10 Principles of Responsible AI,”. Available: http://womenleadinginai.org/wp-content/uploads/2019/02/WLiAI-Report-2019.pdf.
4. Internet Society, “Artificial Intelligence and Machine Learning: Policy paper,” 2017. Available: https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf.
5. W20, “Artificial Intelligence: Open questions about gender inclusion,” 2018. Available: http://webfoundation.org/docs/2018/06/AI-Gender.pdf.
6. ICDPPC. “Declaration on Ethics and Data Protection in Artificial Intelligence,” 2018. Available: https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf.
7. Software & Information Industry Association (SIIA), Public Policy Division, “Ethical Principles for Artificial Intelligence and Data Analytics,” 2017. Available: http://www.siia.net/Portals/0/pdf/Policy/Ethical%20Principles%20for%20Artificial%20Intelligence%20and%20Data%20Analytics%20SIIA%20Issue%20Brief.pdf?ver=2017-11-06-160346-990.
8. Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, Version 1,” 2019. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents.
9. Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, Version 2,” 2017. [Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.
10. Independent High-Level Expert Group on Artificial Intelligence Set Up by The European Commission, Ethics Guidelines for Trustworthy AI, European Commission - Digital Single Market, 2019. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
11. Information Technology Industry Council (ITI), “ITI AI Policy Principles,” 2017. Available: https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.
12. Fairness, Accountability, and Transparency in Machine Learning (FATML). “Principles for Accountable Algorithms and a Social Impact Statement for Algorithms,” 2016. [Online]. Available: https://www.fatml.org/resources/principles-for-accountable-algorithms.
13. Privacy International & Article 19, “Privacy and Freedom of Expression in the Age of Artificial Intelligence,” 2018. [Available: https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf.

14. COMEST/UNESCO, Report of COMEST on Robotics Ethics, 2017. Available: https://unesdoc.unesco.org/ark:/48223/pf0000253952.
15. Partnership on AI, "Tenets," 2016. [Online]. Available: https://www.partnershiponai.org/tenets/.
16. Access Now; Amnesty International, "The Toronto Declaration: Protecting the right to equality and nondiscrimination in machine learning systems," 2018. [Online]. Available: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf.
17. UNI Global Union, "Top 10 Principles for Ethical Artificial Intelligence," 2017. [Online]. Available: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.
18. The Public Voice, "Universal Guidelines for Artificial Intelligence," 2018. [Online]. Available: https://epic.org/international/AIGuidleinesDRAFT20180910.pdf.
19. WEF, Global Future Council on Human Rights 2016-2018, "White Paper: How to Prevent Discriminatory Outcomes in Machine Learning," 2018. [Online]. Available: http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.
20. B. Custers, T. Calders, B. Schermer, and T. Zarsky, Editors. Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases. 2013 edition. New York: Springer, 2012. 370 p.
21. S. Chiappa and T.P.S. Gillam, "Path-Specific Counterfactual Fairness," ArXiv180208139 Stat, 2018 Feb 22. [Online] Available: http://arxiv.org/abs/1802.08139.
22. M.J. Kusner, J.R. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," ArXiv170306856 Cs Stat, 2017 Mar 20. [Online]. Available: http://arxiv.org/abs/1703.06856.
23. Deutsche Telekom, "AI Guidelines," 2018. Available: https://www.telekom.com/resource/blob/532446/f32ea4f5726ff3ed3902e97dd945fa14/dl-180710-ki-leitlinien-en-data.pdf.
24. Mission Villani, "For a Meaningful Artificial Intelligence. Towards a French and European strategy," 2018. Available: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.
25. Leaders of the G7, "Charlevoix Common Vision for the Future of Artificial Intelligence," 2018. [Online]. Available: https://www.mofa.go.jp/files/000373837.pdf.
26. "AI Ethics Guidelines Global Inventory," AlgorithmWatch, n.d. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/. [Accessed: 2020 Jan 15].
27. K. Hao, "OpenAI has released the largest version yet of its fake-news-spewing AI," MIT Technology Review, Aug. 29, 2019. Available: https://www.technologyreview.com/s/614237/openai-released-its-fake-news-ai-gpt-2/
28. Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI, "The Malicious Use of Artificial Intelligence: Forecasting, prevention, and mitigation," 2018. Available: https://arxiv.org/pdf/1802.07228.pdf.
29. K. Lippert-Rasmussen, Editor, The Routledge Handbook of the Ethics of Discrimination. New York: Routledge, 2017. 676 p.

30. K. Lippert-Rasmussen, Born Free and Equal? A philosophical inquiry into the nature of discrimination. Oxford: Oxford University Press, 2014.
31. T. Calders and I. Žliobait?, "Why Unbiased Computational Processes can Lead to Discriminative Decision Procedures," in Discrimination and Privacy in the Information Society, B. Custers, T. Calders, T. Zarsky, Editors. Dordrecht: Springer, 2013, pp. 43-57.
32. S. Barocas and A.D. Selbst, "Big Data's Disparate Impact," Social Science Research Network, Report No. ID 2477899, 2015 August. Available from: http://papers.ssrn.com/abstract=2477899.
33. K. Arrow, "Some Models of Racial Discrimination in the Labor Market," Rand Corporation, 1971. Available: https://www.rand.org/pubs/research_memoranda/RM6253.html.
34. S. Maitzen, "The Ethics of Statistical Discrimination," Social Theory and Practice, vol. 17, no. 1, p. 23–45, 1991.
35. G.S. Becker, The Economics of Discrimination. Chicago: University of Chicago Press, 1957.
36. R. Dworkin, "Affirmative Action: Is it fair?" Journal of Blacks in Higher Education, vol. 28, p. 79–88, 2000.
37. R.M. Dworkin, Sovereign Virtue: The theory and practice of equality. Cambridge, MA: Harvard University Press, 2000.
38. S. Wachter, B. Mittelstadt, and C. Russell. "Why Fairness Cannot Be Automated: Bridging the gap between EU non-discrimination law and AI," Social Science Research Network, Report No. ID 3547922, 2020 March. Available: https://papers.ssrn.com/abstract=3547922.
39. A. E. Buchanan, Marx and justice: the radical critique of liberalism. Totowa, N.J: Rowman and Littlefield, 1982.
40. A. E. Buchanan, "Marx, Morality, and History: An assessment of recent analytical work on Marx," Ethics, vol. 98, no. 1, p. 104–136, 1987.
41. F. A. Hayek, Law, Legislation and Liberty, Volume 2: The mirage of social justice. Chicago: University of Chicago Press, 2012.
42. "Hayek & Rawls," Evatt Foundation, 2007. [Online]. Available: https://evatt.org.au/papers/hayek-rawls.html.
43. R. Nozick, Anarchy, State, and Utopia. New York: Basic Books, 1974.
44. J. Kleinberg, S. Mullainathan, M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," ArXiv160905807 Cs Stat, 2016 Sep 19. Available: http://arxiv.org/abs/1609.05807.
45. R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, "Fairness in Criminal Justice Risk Assessments: The state of the art," Sociological Methods and Research, 2018 Jul 2. [Online first.] Available: https://doi.org/10.1177/0049124118782533
46. H. Heidari, M. Loi, K.P. Gummadi, A. Krause, "A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity" in Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019. pp. 181–190. Available: http://doi.acm.org/10.1145/3287560.3287584. (FAT– '19).
47. M.B. Zafar, I. Valera, M.G. Rodriguez, K.P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact: Learning classification without disparate mistreatment," ArXiv161008452 Cs Stat, p. 1171–80, 2017.

48. S. Wang and M. Gupta, "Deontological Ethics by Monotonicity Shape Constraints," ArXiv200111990 Cs Stat, 2020 Mar 12.. Available: http://arxiv.org/abs/2001.11990.
49. Z.C. Lipton, "The Mythos of Model Interpretability," ArXiv160603490 Cs Stat, 2016 Jun 10. Available: http://arxiv.org/abs/1606.03490.
50. M.T. Ribeiro, S. Singh, C. Guestrin, "'Why Should I Trust You?': Explaining the predictions of any classifier," ArXiv160204938 Cs Stat, 2016 Feb 16. Available: http://arxiv.org/abs/1602.04938.
51. S. Wachter, B. Mittelstadt, C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated decisions and the GDPR," Harvard Journal of Law and Technology, 2017. Available https://ssrn.com/abstract=3063289
52. R. Binns, M.V. Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, "'It's Reducing a Human Being to a Percentage': Perceptions of justice in algorithmic decisions," SocArXiv, 2018 Jan 31. Available: https://osf.io/preprints/socarxiv/9wqxr/.
53. C. Molnar, "Interpretable Machine Learning," 2019. Available: https://christophm.github.io/interpretable-ml-book/
54. O. Li, H. Liu, C. Chen, C. Rudin, "Deep Learning for Case-based Reasoning Through Prototypes: A neural network that explains its predictions," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018. Available: https://arxiv.org/abs/1710.04806
55. D.V. Carvalho, E.M. Pereira, J.S. Cardoso, "Machine Learning Interpretability: A survey on methods and metrics," Electronics, vol. 8, no. 8, p. 832, 2019. Available: https://www.mdpi.com/2079-9292/8/8/832
56. C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, "This Looks Like That: Deep learning for interpretable image recognition," in Advances in Neural Information Processing Systems. 2019, p. 8928–8939. Available: https://papers.nips.cc/paper/9095-this-looks-like-that-deep-learning-for-interpretable-image-recognition.pdf
57. B. Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," Nature Machine Intelligence, p. 1-7, 2019. Available: https://www.nature.com/articles/s42256-019-0114-4.pdf
58. C.R. Beitz, The Idea of Human Rights. New York: Oxford University Press, 2009.
59. F. Santoni de Sio and J. Van den Hoven, "Meaningful Human Control over Autonomous Systems: A philosophical account," Front Robot AI, 2018;5. Available: https://www.frontiersin.org/articles/10.3389/frobt.2018.00015/full.

3.2 ESSENTIAL BIBLIOGRAPHY (BY TOPIC)

ACCOUNTABILITY (FOR ALGORITHMS)

M. Wieringa, "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, Jan. 2020, pp. 1–18, doi: 10.1145/3351095.3372833.

AUTONOMY AND FREEDOM

A. Sen, "Well-being, agency and freedom: the Dewey lectures 1984," The journal of philosophy, vol. 82, no. 4, pp. 169–221, 1985.

DISCRIMINATION (IN GENERAL)

K. Arrow, "Some Models of Racial Discrimination in the Labor Market," Rand Corporation, 1971. Available: https://www.rand.org/pubs/research_memoranda/RM6253.html.

G.S. Becker, The Economics of Discrimination. Chicago: University of Chicago Press, 1957.

R. Dworkin, "Affirmative Action: Is it fair?" Journal of Blacks in Higher Education, vol. 28, p. 79–88, 2000.

K. Lippert-Rasmussen, Editor, The Routledge Handbook of the Ethics of Discrimination. New York: Routledge, 2017. 676 p.

K. Lippert-Rasmussen, Born Free and Equal? A philosophical inquiry into the nature of discrimination. Oxford: Oxford University Press, 2014.

S. Maitzen, "The Ethics of Statistical Discrimination," Social Theory and Practice, vol. 17, no. 1, p. 23–45, 1991.

DISCRIMINATION AND FAIRNESS IN MACHINE LEARNING

S. Barocas and A.D. Selbst, "Big Data's Disparate Impact," Social Science Research Network, Report No. ID 2477899, 2015 August. Available from: http://papers.ssrn.com/abstract=2477899.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, "Fairness in Criminal Justice Risk Assessments: The state of the art," Sociological Methods and Research, 2018 Jul 2. [Online first.] Available: https://doi.org/10.1177/0049124118782533

T. Calders and I. Žliobait?, "Why Unbiased Computational Processes can Lead to Discriminative Decision Procedures," in Discrimination and Privacy in the Information Society, B. Custers, T. Calders, T. Zarsky, Editors. Dordrecht: Springer, 2013, pp. 43-57.

H. Heidari, M. Loi, K.P. Gummadi, A. Krause, "A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity" in Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019. pp. 181–190. Available: http://doi.acm.org/10.1145/3287560.3287584. (FAT– '19).

J. Kleinberg, S. Mullainathan, M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," ArXiv160905807 Cs Stat, 2016 Sep 19. Available: http://arxiv.org/abs/1609.05807.

M.B. Zafar, I. Valera, M.G. Rodriguez, K.P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact: Learning classification without disparate mistreatment," ArXiv161008452 Cs Stat, p. 1171–80, 2017.

INTERPRETABILITY, INTELLIGIBILITY, AND EXPLAINABILITY IN MACHINE LEARNING

D.V. Carvalho, E.M. Pereira, J.S. Cardoso, "Machine Learning Interpretability: A survey on methods and metrics," Electronics, vol. 8, no. 8, p. 832, 2019. Available: https://www.mdpi.com/2079-9292/8/8/832

C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, "This Looks Like That: Deep learning for interpretable image recognition," in Advances in Neural Information Processing Systems. 2019, p. 8928–8939. Available: https://papers.nips.cc/paper/9095-this-looks-like-that-deep-learning-for-interpretable-image-recognition.pdf

Z.C. Lipton, "The Mythos of Model Interpretability," ArXiv160603490 Cs Stat, 2016 Jun 10. Available: http://arxiv.org/abs/1606.03490.

O. Li, H. Liu, C. Chen, C. Rudin, "Deep Learning for Case-based Reasoning Through Prototypes: A neural network that explains its predictions," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018. Available: https://arxiv.org/abs/1710.04806

T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3278331, Nov. 2018. Accessed: Nov. 13, 2018. [Online]. Available: https://papers.ssrn.com/abstract=3278331.

C. Molnar, "Interpretable Machine Learning," 2019. Available: https://christophm.github.io/interpretable-ml-book/

M.T. Ribeiro, S. Singh, C. Guestrin, "'Why Should I Trust You?': Explaining the predictions of any classifier," ArXiv160204938 Cs Stat, 2016 Feb 16. Available: http://arxiv.org/abs/1602.04938.

JUSTICE

A. E. Buchanan, Marx and justice: the radical critique of liberalism. Totowa, N.J: Rowman and Littlefield, 1982.

A. E. Buchanan, "Marx, Morality, and History: An assessment of recent analytical work on Marx," Ethics, vol. 98, no. 1, p. 104–136, 1987.

F. A. Hayek, Law, Legislation and Liberty, Volume 2: The mirage of social justice. Chicago: University of Chicago Press, 2012.

R. Nozick, Anarchy, State, and Utopia. New York: Basic Books, 1974.

R.M. Dworkin, Sovereign Virtue: The theory and practice of equality. Cambridge, MA: Harvard University Press, 2000.

J. Rawls, A Theory of Justice. 2nd ed. Cambridge, MA: Harvard University Press; 1999.

PRIVACY AND GROUP PRIVACY

A. Acquisti, “Privacy in electronic commerce and the economics of immediate gratification,” in Proceedings of the 5th ACM conference on Electronic commerce, 2004, pp. 21–29, Accessed: Jun. 23, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=988777.

E. J. Bloustein, Individual and Group Privacy, 2nd ed. New Brunswick, N.J., U.S.A.: Routledge, 2003.

G. Danezis et al., “Privacy and data protection by design-from policy to engineering,” arXiv preprint arXiv:1501.03726, 2015.

C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Foundations and Trends® in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.

W. Hartzog, Privacy’s Blueprint: The Battle to Control the Design of New Technologies. Cambridge, Massachusetts: Harvard University Press, 2018.

M. Loi and M. Christen, “Two Concepts of Group Privacy,” Philos. Technol., May 2019, doi: 10.1007/s13347-019-00351-0.

H. Nissenbaum, “Privacy as contextual integrity,” Washington law review, vol. 79, no. 1, 2004, Accessed: Jun. 05, 2015. [Online].

H. Nissenbaum, Privacy in context: Technology, policy, and the integrity of social life. Stanford: Stanford University Press, 2009.

L. Taylor, L. Floridi, and B. Van der Sloot, Eds., Group Privacy: New Challenges of Data Technologies. Cham: Springer, 2016.