Towards a multi-stakeholder value-based assessment framework for algorithmic systems

Mireia Yurrita
m.yurritasemperena@tudelft.nl
Human Information Communication Design
Delft University of Technology
The Netherlands

Dave Murray-Rust
D.S.Murray-Rust@tudelft.nl
Human Information Communication Design
Delft University of Technology
The Netherlands

Agathe Balayn
a.m.a.balayn@tudelft.nl
Web Information Systems
Delft University of Technology
The Netherlands

Alessandro Bozzon
A.Bozzon@tudelft.nl
Knowledge and Intelligence Design
Delft University of Technology
The Netherlands

ABSTRACT

In an effort to regulate Machine Learning-driven (ML) systems, current auditing processes mostly focus on detecting harmful algorithmic biases. While these strategies have proven to be impactful, some values outlined in documents dealing with ethics in ML-driven systems are still underrepresented in auditing processes. Such *unaddressed* values mainly deal with contextual factors that cannot be easily quantified. In this paper, we develop a value-based assessment framework that is not limited to bias auditing and that covers prominent ethical principles for algorithmic systems. Our framework presents a circular arrangement of values with two bipolar dimensions that make common motivations and potential tensions explicit. In order to operationalize these high-level principles, values are then broken down into specific criteria and their manifestations. However, some of these value-specific criteria are mutually exclusive and require negotiation. As opposed to some other auditing frameworks that merely rely on ML researchers' and practitioners' input, we argue that it is necessary to include stakeholders that present diverse standpoints to systematically negotiate and consolidate value and criteria tensions. To that end, we map stakeholders with different insight needs, and assign tailored means for communicating value manifestations to them. We, therefore, contribute to current ML auditing practices with an assessment framework that visualizes closeness and tensions between values and we give guidelines on how to operationalize them, while opening up the evaluation and deliberation process to a wide range of stakeholders.

CCS CONCEPTS

• **General and reference → Evaluation**; • **Human-centered computing → Human computer interaction (HCI)**; • **Social and professional topics → User characteristics**.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533118

KEYWORDS

values, ML development and deployment pipeline, algorithm assessment, multi-stakeholder

ACM Reference Format:
Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3531146.3533118

1 INTRODUCTION

In recent years, it has become clear that algorithmic systems might encode harmful biases and might lead to unfair outcomes [132, 159]. The dangers of using Machine Learning (ML) in Computer Vision (CV) [31] or Natural Language Processing (NLP) [8, 21, 46, 112, 153], for assessing recidivism [170], for candidate screening [146] and for recommending content on social media platforms [99, 142, 148, 192] have been pinpointed. The origins of harmful algorithmic bias[1] might be diverse [159, 168]. Just to mention a few, representativeness issues, play a key role in disparate algorithmic performance [3, 34]. The way in which data is collected [21, 144] and labelled [42, 53, 144] is a major menace to data soundness. Beyond the data generation process, aggregation, learning, evaluation and deployment biases have been identified throughout the ML pipeline [168]. In response to harmful algorithmic bias, current auditing processes [2] [2, 151, 181] have provided numerous useful bias detection techniques [10, 18, 19, 57, 70, 90, 184, 185, 188].

However, harmful algorithmic behavior is not limited to biases encoded in the ML life cycle [159]. The lack of social and cultural context in the mathematical representation of socio-technical systems [124, 159] or the omission of changing practices and long-term effects of the deployed systems [29, 45, 96, 109] are also some

[1]Following the approach adopted by Shen et al. [159], we will distinguish between harmful algorithmic biases and harmful algorithmic behaviors, since not all harmful algorithmic behaviors originate from biases and not all algorithmic biases are necessarily harmful [28].

[2]We will use the term *auditing* processes to refer to external audits, where third parties only have access to model outputs [152]. We will use the term *assessment* processes to refer to an evaluation process that is applied "throughout the development process and that enables proactive ethical intervention methods" [147]. We will not use the term *Internal Audit* defined by Raji and Smart [147] to avoid erroneous inferences that would limit the stakeholders of our framework to the employees of an organization.

problematic aspects that are hardly considered in current auditing processes. Such processes mostly consist of quantitative analysis for assessing the conformance of those systems to applicable standards [93], rather than additionally gaining insights into their contextual implications [147, 159]. Furthermore, these auditing approaches solely rely on ML researchers, and practitioners, who can fail to detect issues that arise from context-dependent unanticipated circumstances during usage time [159].

In this paper, we argue that: firstly, assessment processes for algorithmic systems should go beyond bias auditing and take into account additional high-level values [3] that are outlined in Artificial Intelligence (AI) ethics documents [39, 41, 60, 74, 92, 94, 95, 127, 138, 157, 164]. Contestability, for example, has been identified as a key value of algorithmic systems, but there is still little guidance on what contestability requires [121]. In order to provide a good coverage of values that deal with principled algorithmic behavior, we develop a value-based assessment framework, where contextual conditions are considered along with quantifiable measurements. We organize such values in a circular layout with two bipolar dimensions. As claimed by Friedman et al. [64], values do not exist in isolation. They are situated in a delicate balance and touching one value might have implications in another value [64]. This means that value interactions need to be taken into account when making choices about value prioritization and situating algorithmic systems in a space of trade-offs [16]. The circularity of our framework makes such interactions explicit and facilitates the identification of common motivations and tensions among values.

Secondly, an assessment process should give tangible guidelines for the operationalization [4] of values, so as to eventually put ethics into practice following a context-aware approach [160]. To this end, each value in our framework is broken down into criteria manifested through quantifiable indicators, process-oriented practices or signifiers[5]. These value-specific criteria and their manifestations can be used either as a checklist if our framework is applied for evaluating a system that is already developed, or for promoting such values if it is being used during design time.

Thirdly, assessment processes should allow critical reflection on algorithmic systems and engage in conflictual plurality[6]. Inevitable value tensions inherent in the nature of socio-technical systems [77] require spaces for ethical discussions [160], that can benefit from the insights of multiple stakeholders beyond ML practitioners [16, 159]. To enable fruitful multi-stakeholder discussions [116], we map and match value-specific communication means with different stakeholders. We, therefore, contribute with:

- A review of prominent high-level values in AI ethics and translation into specific criteria through the:
  - Design of an assessment framework that facilitates the identification of common motivations and tensions among values encoded in ML-driven systems.
  - Definition of guidelines to deal with the complex middle ground between abstract values and concrete system specifications.
- Translation of value-specific criteria into manifestations that are understandable for diverse stakeholders through the:
  - Review of available means to communicate value manifestations to different stakeholders based on their insight needs and nature of knowledge.
  - Definition of steps to introduce those communication means into multi-stakeholder deliberation dynamics.

The remainder of the paper is organized as follows: in section 2, we analyze related work for documenting and auditing ML systems. We also introduce the theoretical basis of our framework. Section 3 describes and justifies the selected values, criteria and manifestations and their arrangement in our framework. Section 4 maps the stakeholders involved in the algorithm evaluation process and reviews the available means for communicating system-specific information to them. In sections 3 and 4, we illustrate the necessary steps for navigating our framework through an example in the area of life insurance application. We discuss our approach, its implications, and future lines of work in section 5, and we conclude this paper in section 6.

2 BACKGROUND AND RELATED WORK

In this section, we survey current practices for documenting and auditing technical specifications of algorithmic systems. We also provide the theoretical basis of our framework.

2.1 Background

2.1.1 Standardized documentation. In order to facilitate the audit of ML-driven systems, it is important that technical specifications are collected and documented in a standardized way. So far, ML system documentation practices are limited to datasets and models.

Documenting datasets. Recent studies in documentation practices for ML datasets claim the need for greater data transparency [91]. Since the quality of the prediction made by the ML system highly depends on the way the data has been collected, the need for setting rigorous practices (as it is the case in other areas of knowledge, such as social sciences or humanities [69]) has been highlighted [144]. Likewise, the choice of what data to collect and how to collect this data is in itself a value-laden decision [48, 155]. To standardize documentation for ML datasets and make data-related decisions more transparent for other practitioners, various methodologies have been suggested in the last years, "Datasheets" [68] and "Dataset Nutrition Labels" [87], for instance. For NLP techniques, "Data Statements" are regarded as a dataset characterization approach that helps developers anticipate biases in language technology and understand how these can be better deployed [20].

Documenting models. In addition to documenting datasets, the importance of disclosing the technical characteristics of ML models has also been emphasized. A good example of model documentation practices are the "Model Cards" [129].

[3] We will adopt the definition of *values* used in philosophy of science, following Birhane et al. [26]. Values of an entity are, thus, defined as properties that are desirable for that kind of entity.

[4] Our strategy follows the definition by Shahin et al. [158], where "operationalizing values" is defined as the process of identifying values and translating them into concrete system specifications that can be implemented.

[5] We adopt the definition given by Don Norman in his 2013 edition of "The Design of Everyday Things". Signifiers are perceivable cues of an affordance, affordances being "the relationship between the properties of an object and the capabilities of the agent that determine how the object could be possibly used". In this paper, the "object" in question is the ML-driven system.

[6] We understand *conflictuality* as a solution for dealing with the "figure of alterity". Unlike *conflict*, it represents a method for linking opposing views and opening out onto the exercise of thinking [65]

2.1.2 Auditing techniques. Various methodologies and tools for incorporating auditing tasks into the Machine Learning workflow have been suggested. Aequitas [151] is an open source toolkit to detect traces of bias in models. The toolkit designed by Saleiro et al. [151] facilitates the creation of equitable algorithmic decision-making systems where data scientists and policymakers can easily use Aequitas for model selection, evaluation and approval. Wilson et al. [181] described a framework that helps ensure fairness in socio-technical systems, and used it for auditing the model of the startup *pymetrics*. Adler et al. [2] studied auditing techniques for black-box models to discover whether proxy variables linked to sensitive attributes indirectly influence the predictions of the system. The end-to-end "Internal Audit Framework" suggested by Raji and Smart [147] is of special interest for justifying the need of setting specific guidelines to enable multi-stakeholder deliberation in assessment processes. It consists of five main stages where the need for stakeholder diversity is highlighted, e.g. the scoping stage calls for covering a "critical range of viewpoints" to review the ethical implications of the system use case.

2.1.3 Motivation. While standardized documentation practices [20, 68, 87, 129] and audits [2, 151, 181] have been influential methodologies for dealing with harmful algorithmic bias, their scope is limited to performing quantitative analysis over data and model outputs so as to ensure compliance with applicable standards [93]. Such an approach does not deal with additional ethical values which cannot be easily quantified [116] and that are essential for ensuring desirable algorithmic behavior. One could argue that "Datasheets" [68] and "ModelCards" [129] already devote a section to the description of ethical considerations of datasets and models. Yet, there are no specific guidelines on how to identify ethical issues. As Shklovski et al. [160] discovered, technical people both in industry and academia struggle to identify what an ethical issue entails. To address this caveat, as part of our value-based framework, we give tangible guidelines for putting ethics into practice [132, 160]. We operationalize each high-level value into actionable value criteria and their manifestations. One could also argue that Raji and Smart [147] already included an Ethics Review as part of their end-to-end internal audit framework. Indeed, they exemplified such a review by describing ethical considerations and potential mitigation strategies against bias and privacy threats for a smile detection system. However, this review does not address most of the values that are referred in AI ethics documents. We fill in this gap by offering a good coverage of values to examine, including those that normally go unnoticed in current documenting and auditing practices.

2.2 Accounting for human values in the assessment of algorithmic systems

Our ML assessment framework identifies and arranges values encoded in algorithmic systems by covering prominent principles in AI ethics and organizing them in a circular structure.

2.2.1 Addressing human values in technology. For the definition of our value-based framework, we followed other theoretically grounded approaches, such as Value Sensitive Design (VSD) [64]. VSD represents a pioneering endeavour where human values are proactively considered throughout the process of technology design [47]. Just as VSD does with interactive systems, we address the need to account for human values during the design, implementation, use, and evaluation [47] of algorithmic systems. To this end, we select and define values involved in ML-driven systems, and we identify stakeholders that will be in contact with such systems and whose standpoints need to be considered. Our approach resonates with conceptual investigations described in VSD literature [47].

The circular nature of our framework is inspired by Schwartz's Theory of Basic Human Values [156]. This theory identifies individual value priorities based on ten basic personal values. Values are arranged in a circular form and categorized in four quadrants. These quadrants are located in two bipolar dimensions, which visualize "oppositions between competing values". In addition, adjacency between values denotes a common motivation, which results in these values forming a circular continuum. The advantage of adopting a circular arrangement, like the one suggested by Schwartz, for ML-driven systems is that value commonalities and trade-offs can be easily identified thanks to their positioning. Considering the struggles of technical people when addressing ethical issues [160], an explicit representation of value interactions will facilitate the analysis of trade-offs and decision-making about value prioritization.

2.2.2 Ethical principles for ML-driven systems. The values considered in our assessment framework cover prominent principles outlined in AI ethics. In the last five years, many institutions have studied and defined high-level principles that AI systems should follow [60]. As a matter of fact, documents that aim at guiding the "ethical development, deployment and governance of AI" are converging into a common set of principles [131, 132]. However, high-level principles are far from being actionable [132] and it is necessary to provide answers on how to proceed [4]. Efforts for going from "what" to "how" [7] include the review carried out by Morley et al. [132], where available tools for operationalizing ethical principles were examined. Similarly, the AI Ethics Impact (AIEI) Group designed a framework for rating the presence of ethical principles in AI systems, getting inspiration from energy efficiency labels [3].

Our value-based framework differs from previous applied ethics frameworks [3, 132] in various ways. Firstly, we arrange values in a circular form, which makes it easier to navigate common motivations and trade-offs between values. Although such common motivations and trade-offs can be inferred from current AI ethics documents, we make them explicit by arranging values in a geometrically meaningful way. This is especially useful for identifying overlaps between values that are adjacent to each other and for detecting potential value tensions that need to be negotiated and consolidated. Secondly, we do not limit our ethics framework to a mere checklist. We follow Shklovski et al. [160] and combine the enumeration of tangible and actionable value manifestations with the generation of an open space for ethical debate. As opposed to the deterministic approach adopted by the AIEI group [3], we map communication means for facilitating ethical reflections of algorithmic systems and for addressing ethical issues in practice

[7] Expression used by Morley et al. [132] to refer to the operationalization of ethical principles in AI. The 'what' refers to the ethical principles themselves, whereas the 'how' refers to the act of putting such principles into practice.

[160]. Thirdly, as opposed to previous applied ethics frameworks [3, 132], we embrace diversity in ethical reflections and deal with the complexities that arise from plurality. In order to facilitate multi-stakeholder discussions, we match available communication means for addressing different value manifestations with stakeholders that present different insight needs.

3 DESIGN OF OUR VALUE-BASED FRAMEWORK

In this section, we describe the composition of our value-based framework and justify its arrangement. We provide the definition of each of the selected values and the derived criteria and manifestations.

3.1 Methodology for reviewing values, criteria and manifestations in ML-driven systems

To design our framework, we analysed documents outlining high-level ethical principles that ML systems should follow. Our starting point was the review performed by Fjeld et al. [60], where principles coming from governments, inter-governmental organizations, multiple stakeholders, the private sector, and the civil society were examined. In their review, Fjeld et al. identify nine key themes, some of which overlap with the values outlined in our framework. The identification of prominent high-level values was also complemented with other reviews [3, 26, 85, 132, 163]. To identify the criteria that define the fulfilment of prominent high-level values, we navigated the visual representation provided by Fjeld et al. [60] and accessed the documents that offer a higher coverage of the value in question. For instance, for the value of *privacy*, one of our main references has been the GDPR [38].

We went from criteria to value manifestations through an extensive exploration of available value-specific reviews that identify such manifestations. For instance, for the value of *security* Xiong et al. [183] presented a thorough study of mechanisms used for securing the ML pipeline against external threats. For *explainability*, Barredo-Arrieta et al. [13] put together more than four hundred references and mapped strategies in the field of Explainable Artificial Intelligence [13]. We partly rely on such reviews for identifying value manifestations because our contribution lies in covering and putting together a set of values and their manifestations in ML-driven systems to end up with a "health-check" for assessing algorithmic systems, rather than rediscovering such value manifestations ourselves. Similarly, for the values of *performance* and *fairness*, we only included the main value manifestations that represent the basis for any other derived metrics. That is to say, just as Verma et al. [177] did, we outline the main quantifiable indicators (false positives, false negatives etc) used for measuring *performance* and *fairness*, but we are aware that many other metrics that derive from these ones can be insightful for specific contexts. Dealing with such compound metrics is out of the scope of this work.

3.2 Assessment of algorithmic systems through a circular value-based framework

Our resulting ML assessment framework arranges values in a circular form (figure 1). Adjacency between values denotes a common motivation and oppositions between competing values are represented through two bipolar dimensions. For instance, adjacency between *privacy* and *security* denotes a common objective towards the protection of sensitive information [60, 150] and resilience to external threats [127]. The trade-off between *privacy* and *explainability*, on the other hand, is made explicit by their opposing positioning in our circular framework. High-level values are then broken down into specific criteria and their manifestations, as indicated in figure 2. Criteria defining a specific value ultimately represent a set of questions to be asked as part of the assessment process to ensure the fulfillment of the value in question —if the framework is being applied before deployment— or the promotion of a specific value —if the framework is being applied during design time—. These sets of criteria are not unique and exclusive to one value. For instance, when defining the criteria for *privacy* we refer to "data protection", which is also involved in *security* in the form of "resilience to attacks". These overlaps are precisely what we want to highlight and make explicit thanks to the circularity of our framework and adjacency between values.

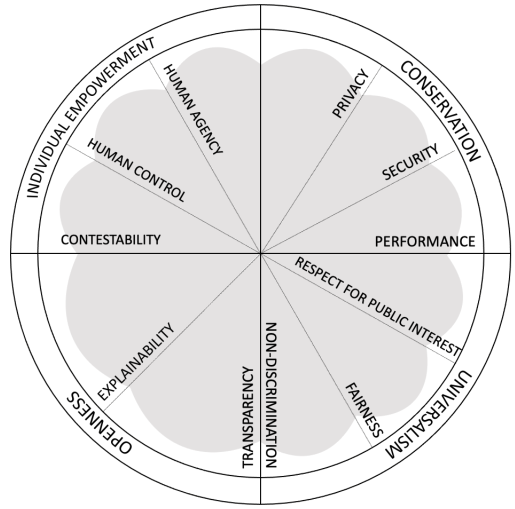

Figure 1: Graphic representation of our circular value-based assessment framework. Oppositions between competing values are illustrated through the arrangement of those values in bipolar dimensions and common motivations through adjacency between values, which form a circular continuum.

Manifestations are classified in three groups depending on their nature: (1) *Quantifiable indicators* are specific measurable parameters that numerically manifest the (lack of) adequacy in the standards set for a criterion (magenta). (2) *Process-oriented practices* are actions and mechanisms implemented during the ML development or deployment process that advocate for a certain value (olive). (3) *Signifiers* [8] are files and reports that describe the relationship between the properties of the algorithmic system and humans that determine how that system can be used (orange). There is a many-to-many relationship between criteria and manifestations.

[8] Check footnote 5

In the next subsections, we present opposing value categories in pairs. We explain in detail the value of *privacy*, the criteria that define it and their manifestations. For the **rest of values**, we give a short definition in the main text, and detail their **criteria and manifestations in appendix A**[9].

Motivating example. For illustrative purposes, we guide the reader through each stage of our framework with a hypothetical yet plausible use case. Consider a team of researchers is developing an ML-driven system for automating life insurance application processes. The system shall accept or refuse the request of a life insurance based on the following data: physiological information of the candidate, details about their employment, insurance history, and individual and family medical history. As part of a new wave of ethical finance companies, the team would like to ensure that their work is ethically grounded. However, it is not clear what that means in practice. From looking at prominent literature, they can develop a sense that the model should be fair and unbiased, and potentially that there should be some level of human control or intervention possible. Yet, they cannot be sure if they have a good set of representative values covered, and they do not know how to go about communicating the way that their model embodies those values. They are now reading through our multi-stakeholder value-based assessment framework.

3.3 Conservation vs Openness

The first dimension of our value-based framework captures the conflict between *conservation* and *openness*. Values included within the *conservation* category emphasize the necessity of ML-driven systems to preserve confidentiality with regards to information, as well as, the need for the system to preserve adequate robustness when it comes to performance. On the contrary, the category of *openness* encompasses values that advocate for making system components and specifications more accessible to the public.

3.3.1 *Conservation.*

Privacy, security and performance uphold confidentiality and robustness within ML systems [60].

Privacy. The defining goal of *privacy* is the need for ML-driven systems to respect individual's informational confidentiality [3, 60] as part of their user rights [26]. When applying this value to the ML development pipeline, data processing itself should integrate privacy standards [3, 60, 132], so that there is no possibility of identifying sensitive information about individuals [85, 178]. Furthermore, the need to provide humans with agency over their data is emphasized [60]. Based on these definitions, we identified six main criteria for the fulfillment of *privacy* within ML systems (table 1). (1) Consent for data usage [3, 38, 60]: data subjects should be appropriately informed when their data is being used and their explicit approval is needed. (2) Implementation of data protection mechanisms [3, 60, 61]: during the development of algorithmic systems, resources should be devoted to making user data management secure and confidential. (3) Users having control over their data and ability to restrict its processing [38, 60]: users should be able to limit the way their personal data is being used. (4) Users having the right to rectify [3, 38, 60]: users should be able to modify their data at any time. (5) Users having the right to erase their data [3, 38, 60]: this criterion refers to the right that users have to be forgotten. (6) Users having right to access their data [38, 140]: this right empowers users to have agency over their data. These criteria manifest in various ways. Signifiers include: a written declaration of consent [38], detailed descriptions of the collected data, how data is handled, how long it will be kept and the purpose of collecting that data [125]. These signifiers are necessary for users to fully understand what sharing their data entails. Process-oriented practices include the obfuscation of data [3] and forms and submission mechanisms to object data collection and make complaints [27].

Security. Definitions characterizing *security* highlight the need for ML systems to be (1) resilient to potential maleficent attacks [60, 132] and to present a (2) predictable [3, 39, 60] and (3) robust [3] behavior at any time. This includes implementing mechanisms to protect user privacy, such as strategies that ensure that inferences about an individual cannot be made by interrogating the model [85, 127, 178]. Following the survey performed by Xiong et al. [183], different methodologies that aim at protecting algorithmic systems against external threats (process-oriented practices) have been classified into two main groups. The first group consists of defence methods against integrity threats at two different stages of the ML pipeline: during training time [23, 43, 72] and during prediction time [23, 73, 122, 141]. The second group aim at defending the ML system against privacy threats, namely membership inference attacks [55, 97, 135, 162, 187].

Performance. The value of *performance* is defined by the (1) correctness of predictions [39, 60], along with the (2-5) resources necessary to reach such predictions [3, 26, 110]. The conditions under which systems are evaluated will have a direct impact on the "appropriateness score" that these systems will obtain in the form of a quantifiable indicator [52]. In other words, if the level of performance is solely measured in terms of accuracy, regardless of the needed data, prerequisites will be inherently favoring big "data-hungry" [113] models. As far as the measurement of performance is concerned, this is mainly done through quantifiable indicators, either referring to the preciseness of the results [129, 180] or to the estimated consumption of environmental resources [15, 44, 66, 67, 123].

3.3.2 *Openness.*

Transparency and explainability advocate for making system components and specifications accessible.

Transparency. Documents providing high-level principles for AI define *transparency* as the property that enables traceability and monitoring of algorithmic systems [60, 132]. *Transparency* relates to the right to information [60] and requires that data or algorithms present some level of accessibility [164]. That is to say, data and models should present some level of (1) interpretability [26, 164], so as to (2) enable human oversight [60, 132]. Those data and models should also be (3) accessible [3, 60, 164], as a step towards achieving (4) traceability [132] and (5) reproducibility [26]. Manifestations of such criteria emerge mostly in the form of documentation detailing technical aspects of the algorithmic system (considered signifiers in our framework) [3, 20, 34, 68, 69, 129, 132, 164]. Process-oriented practices mostly focus on giving open access to data and algorithms

[9] Additional material can be found in the appendices of this paper or on the companion page https://mireiayurrita.github.io/valuebasedframework/

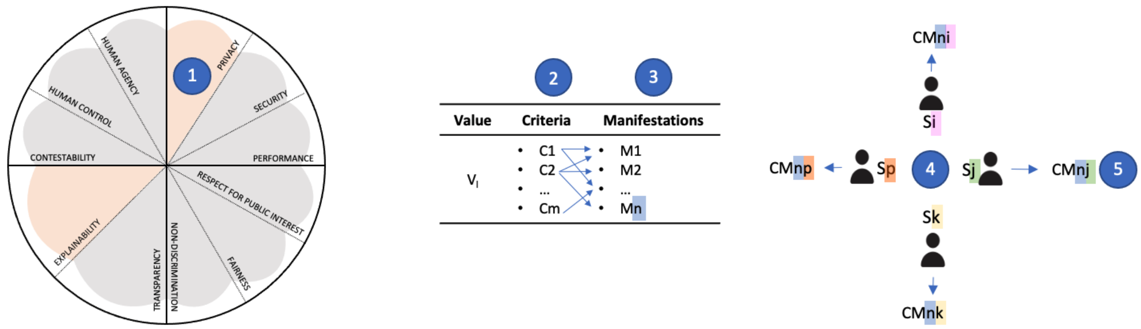

Figure 2: Workflow for operationalizing high-level values and for enabling multi-stakeholder assessment of algorithmic systems. This workflow represents the methodology that we followed for structuring our framework and the steps that researchers and practitioners should take to make use of it. (1) Select and discuss project-specific values (V), (2) Decide on criteria (C) for embodying those values, (3) Select the manifestations (M) that enact value-specific criteria, (4) Map relevant stakeholders (S) to enable ethical reflection of value and criteria tensions, (5) Match adequate communication means (CM) to stakeholders.

Value	Criteria	Manifestations
Privacy	• **Consent for data usage** [3, 38, 60] • **Data protection** [3, 60, 61] • **Control over data / ability to restrict processing** [38, 60] • **Right to rectification** [3, 38, 60] • **Right to erase the data** [3, 38, 60] • **Right of access by data subject, data agency** [38, 140]	• Written declaration of consent [38] • Description of what data is collected [125] • Description of how data is handled [125] • Purpose statement of data collection [125] • Statement of how long the data is kept [125] • Form and submission mechanisms to object data collection and to make complaints [27] • Obfuscation of data [3]

Table 1: Illustration of how to move from values, to criteria and their manifestations with an example for *privacy*. The rest of the values, criteria and manifestations are detailed in appendix A.

[3, 26, 60, 164], regularly reporting key information about the system [60] and notifying users whenever they are being subject to or interacting with an algorithmic system [60].

Explainability. Explainable Artificial Intelligence (XAI) is formed by a set of techniques that allow a wide range of stakeholders to understand why or how a decision was reached by an algorithmic system [61, 164]. *Explainability* is, thus, conceived as an interface that translates reasoning mechanisms of the system into formats that are (1) comprehensible [13, 26, 39, 60–62, 138, 164]. In addition, strategies for making black-box algorithms more interpretable facilitate their (2) monitoring [132] and, therefore, make them (3) suitable for evaluation [60, 132]. XAI techniques (process-oriented practices) are very diverse in nature. As claimed by Vera Liao et al. [117] and Barredo-Arrieta et al. [13], *explainability* methodologies are usually classified by the scope of the explanation, complexity of the model, model specificity and the stage of the ML pipeline where such a strategy is to be used. For our framework, we will consider that explainable models can be either (a) interpretable by design or they can be (b) explained by additional *post-hoc* explanations [13].

3.4 Universalism vs Individual Empowerment

The second dimension captures the conflict between *universalism* and *individual empowerment*. Values included within the *individual empowerment* category emphasize the defense of the decision subjects' interests. These principles advocate for giving decision subjects the means to oppose to the conclusion reached and uphold the need for putting humans in the loop. Values within the *universalism* category emphasize the need to equalize system behavior to *all* and to ensure that such a system adheres to the interests of society as a whole, beyond the interests of a few individuals.

3.4.1 Universalism. Respect for public interest, fairness and non-discrimination uphold the need to ensure equitable and socially acceptable system behavior for *all*.

Respect for public interest. The value of *respect for public interest* deals with the (1) appropriateness of developing algorithmic systems for a certain purpose within a specific context. As Keyes et al. [104] claimed, making ML-driven systems fairer, more transparent and more accountable is insufficient if we ignore the purpose of developing and implementing these systems in a certain context in the very first place [107, 171]. Algorithmic systems should, therefore, (2) be beneficial to society and humanity as a whole

[60–62, 132], respect law [26] and be aligned with human norms [60]. This involves giving a clear justification of the purpose and benefits of building such a system [1, 34, 104, 132], so that the deployment of the system in question upholds public-spirited goals [60]. Universalism aims at protecting the welfare of *all*, both people and nature [110]. AI systems' (3) negative impacts on environment should, therefore, be considered and valued [3, 21]. To this end, process-oriented practices include the creation of diverse and inclusive forums for discussion [14, 60], whereas signifiers include the qualitative measurement of social and environmental impact [21, 132, 147].

Fairness. The value of *fairness* represents a complex concept that accepts multiple definitions [16, 103], some of which cannot be satisfied simultaneously [79, 85, 103]. Overall, we will understand *fairness* in terms of parity in output [51] and equal treatment [3] among individuals. When addressing more specific definitions of *fairness* (1-8), we will adopt the approach followed by Verma et al. [177], which was also echoed by Mehrabi et al. [126] (for a detailed enumeration and explanation of each of the definitions, the reader is encouraged to check appendix A). ML techniques generally conceive fairness in terms of statistical metrics [75] and observe whether specific quantifiable indicators are above or below the thresholds set for a certain application. Even if error rates were equal across groups for a certain application, if those rates are too high, the system could still be considered unfair [79]. This means that for our value-based framework we outline the quantifiable indicators that are normally used for manifesting fairness-related criteria, but we do not determine the threshold for these indicators to be considered good enough for a specific application. Similarly, the quantifiable indicators relate to the output of the system, rather than the outcome that these outputs lead to.

Non-discrimination. The value of *non-discrimination*, as defined in our framework, deals with algorithmic systems not being socially biased [26] and ensuring that equal accessibility is provided to all individuals [132]. This means that (1) quality and integrity of data should be evaluated and ensured [60, 69, 85, 132, 144] in order to prevent "socially constructed biases, inaccuracies, errors, and mistakes" [132] from being present in the data. Processes that safeguard inclusive data generation [3, 34, 69, 132] and analysis procedures for identifying potential biases in data and for assessing its quality [60, 69, 85, 132, 144] are strategies that avoid social stereotypes being codified, maintained and amplified [85]. Furthermore, non-discriminatory systems should (2) ensure diversity and inclusiveness in the design process [39, 60, 132]. From a process-oriented perspective, participants involved in the development process should, thus, present diverse profiles [3, 60, 115, 189]. Finally, giving (3) equal access to the technology [3, 26, 60, 132] avoids the growth of inequalities as a consequence of deploying AI systems [60].

3.4.2 Individual empowerment. Contestability, human control and human agency address the politics behind algorithmic systems [12, 182] and deal with the issues caused by power imbalances [26, 34, 100, 121, 173].

Contestability. The value of *contestability* is defined as the value that ensures that users have the necessary information to (1) enable argumentation against conclusions reached by algorithmic systems [6, 17, 39, 60, 100, 114, 121, 164]. This involves (2) empowering citizens [17, 39, 100] to investigate and influence AI [100], as part of a broader regulatory approach [121]. As a matter of fact, *contestability* has been identified as a "critical aspect of future public decision-making systems" [6]. This implies that, from a documentation perspective (signifiers), users should be made aware of who determines what constitutes a contestable decision, who is accountable for it and who can contest a decision. This last point is particularly necessary to determine whether (legal) representatives of decision subjects can act on their behalf. The review mechanism in place and the workflow of contestations [121] are policy-related details that users should also be informed about. From a process-oriented standpoint, mechanisms for users to ask questions and to record disagreements should also be put in place [86, 130].

Human Control. The value of *Human control* addresses the influence that data-driven technologies have over humans and that leads to a reduction of human agency, power and control [143]. Algorithmic systems should be controllable [26] and (1) subject to user and collective influence [26, 114]. They should also be (2) subject to human review [60]. Governance mechanisms that ensure human oversight of automated decisions are, thus, necessary to maintain control and influence over such systems [132]. It should be possible to (3) choose how and even whether (in the very first place) to delegate a decision to an automated system [60]. From a development perspective, levels of human discretion should be established [39, 127] and the ability to override decisions made by a system [39] ought to be set up by design. Once the system is deployed, it should be continuously monitored to enable adequate intervention when necessary [39, 40, 60].

Human Agency. The value of *human agency* deals with the risks of algorithmic systems displacing human autonomy [39, 60]. As claimed by Cila et al. [37], algorithmic systems may displace human agency in governance processes and may undermine human autonomy. ML-driven systems advocating for human agency should, therefore, (1) respect human autonomy [39, 60, 132] and (2) citizens' power to decide [26, 39]. In addition, (3) decision subjects should be able to opt out of an automated decision [39, 60]. The manifestations of such criteria involve giving knowledge and tools to users to comprehend and interact with AI systems [39] (signifier) and, from a process-oriented perspective, providing strategies for users to self-assess the systems [39].

Selecting values, criteria and manifestations for our example use case. Returning to the hypothetical insurance modelling team from our motivating example (section 3.2), they decided to apply our value-based framework before launching their system. They quickly realised that they need to consider more values than those outlined in current auditing processes. For example transparency, non-discrimination, supporting human agency and the public good. They also discovered a range of methods for enacting those values: from data handling processes that ensure anonymity and meaningful consent around the model, to models of fairness appropriate to their case.

Although we cover prominent ethical principles in AI and the assessment of the algorithmic system might include all of them, here we focus on a subset of those values for illustrative purposes. We imagine that the researchers developing the algorithmic life insurance application system want to focus on *explainability* and *privacy* (fig 2). We assume that they are dealing with a blackbox algorithm that is not interpretable by design. The team needs to examine whether the algorithmic system and the decision reached are understandable. Additionally, the deployed XAI methods should enable traceability and evaluation of the system. As far as the *explainability* manifestations are concerned, since they are dealing with a blackbox algorithm, they need to deploy adequate post-hoc explanations. When it comes to *privacy*, the data used for training and testing the algorithmic model should have been obtained through the explicit approval of the decision subjects. These subjects should have been informed about the nature and purpose of the data that is collected, the way this data is handled and stored. Decision subjects should also have agency and control over their data. Additionally, data protection mechanisms should have been implemented to make sure that there is no possibility of identifying sensitive (in this case medical) data about the subjects. These two values that the team needs to advocate for, represent some trade-offs: XAI methods uphold interpretability of algorithmic systems and some of them even rely on comparing data instances at inference time with those used for training the system. This would directly violate the subjects' right to have their data protected and confidentiality ensured.

4 TOWARDS A MULTI-STAKEHOLDER CRITICAL REFLECTION OF ALGORITHMIC SYSTEMS

Since there are value trade-offs, like the one outlined in our example use case, and certain value-specific criteria are mutually exclusive, we follow the claim made by Raji and Smart [147], and advocate for standpoint diversity. This implies involving a wide range of stakeholders in the negotiation process [160] to discuss and critically reflect on the degree to which each of the values should be promoted in detriment of the other one and how the prioritization process should take place. These stakeholders will possess different types of knowledge and will present different insight needs. In this section we map those stakeholders and match them with the most suitable communication means.

4.1 Methodology for identifying relevant stakeholders and communication means

To identify relevant stakeholders, we follow the stakeholder characterization of Suresh et al. [167]. They classified stakeholders in a two dimensional matrix, where one dimension captured the nature of the knowledge of the stakeholders (formal, instrumental or personal) and the second one identified the context in which that knowledge manifests (Machine Learning, data domain, and the general milieu). Formal knowledge entails a deep understanding of the theories of a certain domain. Instrumental knowledge refers to the capability of applying formal knowledge in one of the three contexts. Personal knowledge is acquired by the participation of the subject in a specific context. The two dimensional-matrix classification results in nine different stakeholder profiles. To facilitate the process of mapping the stakeholders to tailored communication means, we narrow those stakeholders down into four categories [10].

We then proceed to identify the means to communicate system-related information to different stakeholders. We searched such means using arXiv and Google search, so as to cover the state of the art in terms of research papers and open source toolkits. Each search referred to specific value criteria and manifestations, although many of the found means address more than one value. This review does not intend to be exhaustive. We expect novel research to address value manifestations that still present scarce resources in our framework. Hence our review is just a snapshot of some of the available communication means until January 2022, but we host the latest version on an online repository [11] and is open to anyone's contribution. We aim at creating a living document that will keep growing and that will address current research gaps as time goes by.

4.2 Mapping stakeholders

We characterize four main stakeholders in our framework: (1) The development team: they have the formal, instrumental and personal knowledge in the domain of ML [167]. They want to ensure and improve product efficiency and research new functionalities [13]. (2) Auditing team: they have the formal and instrumental knowledge of the general milieu, meaning that they are aware of the social theories behind AI, and are able to evaluate technical specifications of ML systems. They aim at verifying model compliance with legislation [13] (3) Data domain experts: they have the theoretical (formal) and instrumental knowledge of the application context (healthcare, economics etc.). They look forward to gaining scientific or domain-specific knowledge [13, 167], trust the model [13, 167] and act based on the model output [167]. And (4) Data subjects: they have the personal knowledge of the data domain in which the AI is being applied and the general milieu. They aim at understanding their situation [13], verifying that the decision is fair [13], contesting the decision (if needed) [167] and understanding how their data is being used [167].

Mapping stakeholders in our example use case. Going back to our example, once *explainability* and *privacy* have been broken down into specific criteria manifestations, the team needs to map the stakeholders who will take part in the assessment process (fig 2). Based on the mapping presented in appendix B, the development team represents the stakeholders who have the knowledge of the math behind the system. An external auditing team will join the discussion to make sure that the model is aligned with current legislation. Since the algorithmic life insurance application system deals with medical data, the data domain experts will be represented by a medical team and a life insurance expert. Decision subjects will be laypeople who seek to understand and verify their situation with regards to data usage and the decision reached by the system.

[10]This reduced classification is backed up by the framework employed by Barredo-Arrieta et al. [13] when identifying the explainability needs of various stakeholders.
[11]https://github.com/mireiayurrita/valuebasedframework

4.3 Mapping tailored communication means

We then examine each of the reviewed means and identify their typology, the value manifestations that they cover and the stakeholders that can make use of it, as illustrated in table 2 for privacy dashboards. The objective of mapping value manifestations, stakeholder profiles and communication means is that of enabling a fruitful and informed discussion among stakeholders. We classify these means in three categories: (1) *Descriptive documents* (red), (2) *Design strategies* (blue), and (3) *Ready-to-use tools* (green). Appendix C summarizes the rest of the communication means and maps them to value manifestations and stakeholders for whom such methods are suitable.

The stakeholders assigned to a specific communication means are based on the audience addressed by the original authors of such methodologies. In some cases, the characterization of the intended audience was not as granular as our stakeholder mapping and the authors merely differed experts in ML from non-experts. Based on the nature of knowledge that we assigned to each of the mentioned stakeholders in section 4.1, we considered that the development and auditing teams are able to understand technically formulated system details (experts) whereas data domain experts and decision subjects would require more accessible communication means (non-experts). Similarly, some of the communication means identified for *explainability* are suitable for any stakeholder, but the original authors formulated the post-hoc explanations with varying degrees of complexity, which should be taken into account when trying to deploy such strategies. If the target audience are data subjects, we echo van Berkel et al. [174] and Cheng et al. [35] and recommend to limit presentation complexity and to instruct participants throughout the session.

It should be noted that this mapping process represents a first step to making a wide range of stakeholders with different backgrounds understand each other. We are aware that communicating system-related information in a tailored way does not directly lead to the resolution of value trade-offs, and that design strategies are necessary for facilitating such conversations [80]. In any case, the exercise of resolving value tensions should be a communicative process, rather than a simple explanation [139]. However, the means used for communicating specifications of the system will play a key role in the dynamics that will take place in those sessions.

Assigning communication means to each stakeholder in our example use case. The life insurance researchers are now looking into appropriate methods for communicating values to different stakeholders (fig 2), so that they can develop a comprehensive plan that ensures both compliance and communication of values.

Based on the mapping presented in appendix C, for the value of *explainability* and its manifestations in the form of post-hoc explanations, the team can use various design strategies and tools as part of their assessment process. To facilitate the navigation of the available communication means, they first examine appendix C to locate the type of means (tool, strategy, or documentation), values and stakeholders they are interested in. Once they select the codes associated to each communication means, they check the selected communication means to see whether the value manifestations in question are addressed and to explore the details related to those means. If the team working on the life insurance case prefers a ready-to-use tool over the description of design strategies for assessing *explainability*, they can use InterpretML [136] and especially the DiCE [133] functionality, (code [AC]) with the development and auditing teams to evaluate counterfactual examples. These counterfactual examples tell how input features should change in order for the output of the system to be different. That is to say, how the individual applying for life insurance should be different, physically, or when it comes to insurance or medical history, for them to accept the application (if the original output was a refusal). However, this tool might not be suitable for non-experts who are not familiar with ML-related concepts. In the life insurance use case, the medical and insurance team and the decision subjects should receive a description of how the output changes if a feature is perturbed, absent or present adapted to their insight needs. This can be done by describing the answers to the questions "Why, Why not and How to be that" for a certain output [117] (code [P]). As for *privacy* manifestations, the development and auditing teams can examine data collection and storage specifications through the Datasheet [68] associated to the dataset in question (code [K]). Special attention should be paid to the "Collection" and "Preprocessing/cleaning/labelling" sections. For decision subjects, iconsets [38, 88, 125, 149] (code [A]) and privacy dashboards [56, 58, 59, 84, 191] (code [B]) are means for them to explore how their data is being used. It should be noted that the cell that intersects between data domain experts and *privacy* is blank. Based on the characterization of stakeholders that we provided, data privacy-related matters are not directly linked to the purpose that data domain experts show when willing to explore algorithmic systems. This is translated into scarcity of methodologies related to *privacy* manifestations that directly address data domain experts.

5 DISCUSSION AND FUTURE WORK

We discuss important aspects of our framework below.

Design choices for creating a value-based framework. We aim at examining values that characterize ML-driven systems rather than the organizations responsible for these systems. Hence, we did not integrate accountability or responsibility as a value *per se* in our framework. We are aware that algorithms cannot be held responsible for the potential harm that they might cause [30, 85], and that in order to effectively deploy such systems, there is an urgent call for accountability [6, 186]. Likewise, we are aware of the need for rigorous frameworks that support accountability [91] and we consider that the act of conceiving an assessment framework itself answers to the need to evaluate and audit algorithmic systems [60]. Nevertheless, we did not explicitly highlight the profiles of the people accountable for the system. We decided to follow Zhu et al. [190] and considered accountability as a governance issue. We do, however, believe that entities up the chain of command should be held accountable for the potential harm caused by algorithmic systems [3, 60, 85, 121, 154]. It should also be noted that values and criteria presented in this paper might not be unique [160]. We acknowledge current discussions in VSD about the shortcomings of pre-selecting values [47] and, hence, do not claim universality. Extension and modification of values is possible in our framework, but are subject to respecting continuity and opposition between values. Similarly, criteria and manifestations can be extended and

	Means	Value	Manifestations	Stakeholder DT	AT	DE	DS	Application	Visual elements
[B]	Privacy dashboards	Privacy	• Description of what and why data is collected • Description of how data is handled				✓	Agnostic	• Timelines • Bar charts • Maps • Network graphs
		Human Agency	• Self-assessment of the system						
		Transparency	• Disclosure of properties of data						

Table 2: Illustration of how we mapped communication means with values, manifestations and stakeholders (DT = Development Team; AT = Auditing Team; DE = Data Domain Experts; DS = Decision Subjects). Privacy dashboards are tools (green) that allow users to interactively assess the collection and usage of their data. The rest of the reviewed communication means are characterized in appendix C.

subsets could be included to create situationally-specific versions of the framework. Since the aim of our framework is to encourage critical reflection [63, 172] and we identified some value manifestations that require additional communication means, we particularly encourage those context-specific adaptations to happen. Under no circumstances should the scarcity of communication means for certain values identified in our framework represent an excuse to justify inaction or to ignore such values.

Context dependence and consistency. As echoed by Liscio et al. [119], in order to translate values into system requirements [145, 175], to reason about conflicting values [5, 134] and to communicate them to different stakeholders [176], it is necessary to situate these values within a context. The prioritization of values depends on the application context of such systems [3]. In this paper, we showed an example of how the framework could be applied to a particular use case. However, considering the differences between value alignments and tensions that may arise due to context dependence, the validity and consistency of our framework is still to be tested. Future work needs to validate our framework across scenarios [71, 174] through user studies or synthetic experiments [165].

Need for standardization. To systematically review and revisit value priorities and tensions among different stakeholders, our framework should be part of a broader evaluation workflow [116], such as the one suggested by Raji and Smart [147]. Besides, practices from software engineering such as the Values Dashboard [137] could be adopted [161, 169]. This dashboard promotes awareness of values and aims at triggering discussions among stakeholders. It claims to be beneficial in each phase of the software development process, from inception to release, and establishes strategies, such as Timelines or Issues, that are already common practice on software development platforms like Github.

Implications of our work. Our multi-stakeholder value-based framework facilitates the unveiling of assumptions that encode political and social values made by developers [147]. By bringing together a wide range of stakeholders to evaluate and discuss value manifestations, one can anticipate and remedy harmful algorithmic behaviors before deploying a system. Besides, we provide researchers and industry practitioners with a good coverage of values to evaluate their systems and the association of such values to actionable value manifestations. This contributes greatly to the adoption of ethical approaches by practically-minded people [132]. For researchers, we provide them with an easy-to-navigate mapping of value manifestations, stakeholders and communication means. Our framework also visually illustrates research gaps that need to be addressed. Blank spaces in appendix C or values with a scarce number of associated communication means directly refer to valuable research opportunities. For instance, for the value of *fairness*, a great deal of effort has been devoted to designing ready-to-use tools for stakeholders with a deep understanding of ML (developers and auditing teams). However, means for addressing *fairness* manifestations and communicating them to decision subjects have not received the same attention. For industry practitioners, we gathered ready-to-use open source toolkits that can be directly applied to their own use cases. Moreover, since we host this mapping on an online repository [12] open to future contributions, we hope that the number of tools addressing each of the identified value manifestations will grow and that the benefits of designing such a framework will be even more tangible in the future.

6 CONCLUSIONS

In this paper, we designed a value-based framework for assessing algorithmic systems from a multi-stakeholder perspective. This provides investigators of algorithmic systems with an actionable set of criteria manifestations to operationalize high-level ethical principles. We arranged eleven prominent values of ML-driven

[12] Check footnote 11

systems in a circular composition, so that common motivations and trade-offs can be easily identified.

We then broke down each of these values into a set of criteria and their correspondent manifestations in the form of quantifiable indicators, process-oriented practices, and signifiers. In addition, we examined available tools for communicating those value manifestations to different stakeholders based on the nature of their knowledge and their insight needs. This should enable to bring a wide range of stakeholders together to systematically assess values encoded in a system and facilitate value- and ethics-related discussions among them. This work completes previous studies that claim the need for incorporating a diverse range of stakeholders and viewpoints in the ML workflow, so that conflicting priorities and value tensions can be reviewed, negotiated and consolidated.

ACKNOWLEDGMENTS

We thank Kars Alfrink for valuable feedback on an earlier version of this paper and our colleagues at the DCODE Network for insightful discussions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260. https://doi.org/10.1145/3351095.3372871

[2] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54 (2018), 95–122. Issue 1. https://doi.org/10.1007/s10115-017-1116-3

[3] AI Ethics Impact Group (AIEIG). 2020. From Principles to Practice An interdisciplinary framework to operationalise AI ethics. https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf

[4] Evgeni Aizenberg and Jeroen van den Hoven. 2020. Designing for human rights in AI. *Big Data & Society* 7 (7 2020), 2053951720949566. Issue 2. https://doi.org/10.1177/2053951720949566 doi: 10.1177/2053951720949566.

[5] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. 2020. Elessar: Ethics in Norm-Aware Agents. *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 16–24.

[6] Kars Alfrink, T. Turel, A. I. Keller, N. Doorn, and G. W. Kortuem. 2020. Contestable City Algorithms. International Conference on Machine Learning Workshop.

[7] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks. *Proceedings of the 25th International Conference on Intelligent User Interfaces.* https://doi.org/10.1145/3377325.3377519

[8] Thayer Alshaabi, David Rushing Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. 2021. The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *EPJ Data Science* 10 (2021), 15. Issue 1. https://doi.org/10.1140/epjds/s13688-021-00271-0

[9] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/2702123.2702509

[10] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 289–295. https://doi.org/10.1145/3306618.3314243

[11] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3411764.3445736

[12] Sherry R Arnstein. 2019. A Ladder of Citizen Participation. *Journal of the American Planning Association* 85 (1 2019), 24–34. Issue 1. https://doi.org/10.1080/01944363.2018.1559388 doi: 10.1080/01944363.2018.1559388.

[13] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (6 2020), 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012

[14] Mission assigned by the French Prime Minister. 2019. For a Meaningful Artificial Intelligence: Toward a French and European Strategy. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

[15] Mahmoud Assran, Joshua Romoff, Nicolas Ballas, Joelle Pineau, and Michael Rabbat. 2019. Gossip-based Actor-Learner Architectures for Deep Reinforcement Learning. (6 2019).

[16] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. 2021. Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems. (3 2021).

[17] Agathe Balayn and Seda Gürses. 2021. Beyond Debiasing: Regulating AI and its inequalities. https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/

[18] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 116–128. https://doi.org/10.1145/3442188.3445875

[19] R K E Bellamy, K Dey, M Hind, S C Hoffman, S Houde, K Kannan, P Lohia, J Martino, S Mehta, A Mojsilović, S Nagar, K Natesan Ramamurthy, J Richards, D Saha, P Sattigeri, M Singh, K R Varshney, and Y Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63 (2019), 4:1–4:15. Issue 4/5. https://doi.org/10.1147/JRD.2019.2942287

[20] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (12 2018). https://doi.org/10.1162/tacl_a_00041

[21] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[22] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. (3 2017).

[23] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (12 2018), 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

[24] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. (1 2018). https://doi.org/10.1145/3173574.3173951

[25] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Issue MSR-TR-2020-32. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[26] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The Values Encoded in Machine Learning Research. (6 2021).

[27] Alice Namuli Blazevic, Patrick Mugalula, and Andrew Wandera. 2021. Towards Operationalizing the Data Protection and Privacy Act 2020: Understanding the Draft Data Protection and Privacy Regulations, 2020. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3776353

[28] Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. (5 2020).

[29] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 150–159. https://doi.org/10.1145/3287560.3287583

[30] Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25 (9 2017), 273–291. Issue 3. https://doi.org/10.1007/s10506-017-9214-9

[31] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Sorelle A Friedler and Christo Wilson (Eds.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 81, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[32] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces.* https://doi.org/10.1145/3301275.3302289
[33] Simeon C Calvert, Daniël D Heikoop, Giulio Mecacci, and Bart Van Arem. 2019. A human centric framework for the analysis of automated driving systems based on meaningful human control. *Theoretical Issues in Ergonomics Science* 21 (2019), 478–506. Issue 4. https://doi.org/10.1080/1463922X.2019.1697390
[34] Kyla Chasalow and Karen Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 77–89. https://doi.org/10.1145/3442188.3445872
[35] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789
[36] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. (10 2016).
[37] Nazli Cila, Gabriele Ferri, Martijn de Waal, Inte Gloerich, and Tara Karpinski. 2020. The Blockchain and the Commons: Dilemmas in the Design of Local Platforms. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376660
[38] European Commission. 2018. 2018 reform of EU data protection rules. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf
[39] European Commission. 2019. Ethics guidelines for trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
[40] Telia Company. 2019. Guiding Principles on Trusted AI Ethics. https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf
[41] Technology Council for Science and Innovation Japanese Cabinet Office. 2019. Social Principles of Human-Centric Artificial Intelligence. https://www8.cao.go.jp/cstp/english/humancentricai.pdf
[42] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The Politics of Training Sets for Machine Learning.
[43] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. 2008. Casting out Demons: Sanitizing Training Data for Anomaly Sensors. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 81–95. https://doi.org/10.1109/SP.2008.11
[44] Steven Dalton, Iuri Frosio, and Michael Garland. 2019. Accelerating Reinforcement Learning through GPU Atari Emulation. (7 2019).
[45] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534. https://doi.org/10.1145/3351095.3372878
[46] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online.* https://doi.org/10.18653/v1/W19-3504
[47] Janet Davis and Lisa P. Nathan. 2015. Value Sensitive Design: Applications, Adaptations, and Critiques. , 11-40 pages. https://doi.org/10.1007/978-94-007-6970-0_3
[48] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. (7 2020). https://arxiv.org/abs/2007.07399
[49] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. (2 2018).
[50] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73. https://doi.org/10.1145/3278721.3278729
[51] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. (1 2019). https://doi.org/10.1145/3301275.3302310
[52] Ravit Dotan and Smitha Milli. 2019. Value-laden Disciplinary Shifts in Machine Learning. (12 2019).
[53] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9 (10 2021), 48–59. Issue 1. https://ojs.aaai.org/index.php/HCOMP/article/view/18939
[54] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. (4 2011).
[55] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. , 265-284 pages. https://doi.org/10.1007/11681878_14
[56] Julia Earp and Jessica Staddon. 2016. "I had no idea this was a thing". *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*, 79–86. https://doi.org/10.1145/3046055.3046062
[57] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2020. FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9 (2020), 197–213. Issue 2. https://doi.org/10.1007/s41060-019-00181-5
[58] Florian M. Farke, David G. Balash, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. 2021. Are Privacy Dashboards Good for End Users? Evaluating User Perceptions and Reactions to Google's My Activity (Extended Version). (5 2021).
[59] Simone Fischer-Hübner, Julio Angulo, Farzaneh Karegar, and Tobias Pulls. 2016. Transparency, Privacy and Trust – Technology for Tracking and Controlling My Data Disclosures: Does This Work? , 3-14 pages. https://doi.org/10.1007/978-3-319-41354-9_1
[60] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3518482
[61] Luciano Floridi. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32 (6 2019). Issue 2. https://doi.org/10.1007/s13347-019-00354-x
[62] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (12 2018). Issue 4. https://doi.org/10.1007/s11023-018-9482-5
[63] Christopher Frauenberger, Marjo Rauhala, and Geraldine Fitzpatrick. 2016. In-Action Ethics: Table 1. *Interacting with Computers* (6 2016). https://doi.org/10.1093/iwc/iww024
[64] Batya Friedman, David G. Hendry, and Alan Borning. 2017. A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction* 11 (2017), 63–125. Issue 2. https://doi.org/10.1561/1100000015
[65] Georges Gaillard. 2016. La conflictualité : une modalité de lien où s'arrime la destructivité humaine. *Connexions* 106 (2016), 71. Issue 2. https://doi.org/10.3917/cnx.106.0071
[66] Yanjie Gao, Yu Liu, Hongyu Zhang, Zhengxian Li, Yonghao Zhu, Haoxiang Lin, and Mao Yang. 2020. Estimating GPU memory consumption of deep learning models. *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 1342–1352. https://doi.org/10.1145/3368089.3417050
[67] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. 2019. Estimation of energy consumption in machine learning. *J. Parallel and Distrib. Comput.* 134 (12 2019), 75–88. https://doi.org/10.1016/j.jpdc.2019.07.007
[68] Timnit Gebru, Google Jamie Morgenstern, Briana Vecchione, and Jennifer Wortman Vaughan. 2020. Datasheets for Datasets.
[69] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336. https://doi.org/10.1145/3351095.3372862
[70] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, and Klaus Mueller. 2020. Measuring Social Biases of Crowd Workers using Counterfactual Queries. (4 2020).
[71] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3 (11 2021), e745–e750. Issue 11. https://doi.org/10.1016/S2589-7500(21)00208-9
[72] Amir Globerson and Sam Roweis. 2006. Nightmare at test time. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 353–360. https://doi.org/10.1145/1143844.1143889
[73] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. (12 2014).
[74] Google. 2018. AI at Google: Our Principles. https://www.blog.google/technology/ai/ai-principles/
[75] Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML).*
[76] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning.
[77] Christopher Groves. 2015. Logic of Choice or Logic of Care? Uncertainty, Technological Mediation and Responsible Innovation. *NanoEthics* 9 (12 2015), 321–333. Issue 3. https://doi.org/10.1007/s11569-015-0238-x
[78] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

[79] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 392–402. https://doi.org/10.1145/3351095.3372831

[80] Katrina Heijne and Han van der Meer. 2019. *Road Map for Creative Problem Solving Techniques Organizing and facilitating group sessions.* Boom Uitgevers Amsterdam.

[81] Drew Hemment, Ruth Aylett, Vaishak Belle, Dave Murray-Rust, Ewa Luger, Jane Hillston, Michael Rovatsos, and Frank Broz. 2019. Experiential AI. *AI Matters* 5 (4 2019), 25–31. Issue 1. https://doi.org/10.1145/3320254.3320264

[82] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. (1 2020).

[83] Clément Henin and Daniel Le Métayer. 2021. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY* (7 2021). https://doi.org/10.1007/s00146-021-01251-8

[84] Eelco Herder and Olaf van Maaren. 2020. Privacy Dashboards: The Impact of the Type of Personal Data and User Control on Trust and Perceived Risk. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 169–174. https://doi.org/10.1145/3386392.3399557

[85] C{é}sar Hidalgo, Diana Orghian, Jordi Albo-Canals, Filipa de Almeida, and Natalia Martin. 2021. *How Humans Judge Machines.* MIT Press. https://hal.archives-ouvertes.fr/hal-03058652

[86] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing Contestability. *Proceedings of the 2017 Conference on Designing Interactive Systems.* https://doi.org/10.1145/3064663.3064703

[87] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. (5 2018).

[88] Leif-Erik Holtz, Katharina Nocun, and Marit Hansen. 2011. Towards Displaying Privacy Information with Icons. , 338-348 pages. https://doi.org/10.1007/978-3-642-20769-3_27

[89] Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for Natural Language Failures with the AI Playbook. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3411764.3445735

[90] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300637

[91] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575. https://doi.org/10.1145/3442188.3445918

[92] IBM. 2019. IBM Everyday Ethics for AI. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

[93] IEEE. 2008. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* (2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[94] China Electronics Standardization Institute. 2018. Original CSET Translation of "Artificial Intelligence Standardization White Paper". https://cset.georgetown.edu/research/artificial-intelligence-standardization-white-paper/

[95] Access Now Amnesty International. 2018. Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

[96] Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An Agent-Based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 412–423. https://doi.org/10.1145/3442188.3445904

[97] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. (9 2019).

[98] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. 2021. EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence. (2 2021). https://arxiv.org/abs/2102.02437

[99] Jonas Kaiser and Adrian Rauchfleisch. 2020. Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube's Channel Recommendations in the United States and Germany. *Social Media + Society* 6 (10 2020), 2056305120969914. Issue 4. https://doi.org/10.1177/2056305120969914 doi: 10.1177/2056305120969914.

[100] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583 (2020). Issue 7815. https://doi.org/10.1038/d41586-020-02003-2

[101] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2020. On the Effectiveness of Regularization Against Membership Inference Attacks. (6 2020).

[102] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness, Jennifer Dy and Andreas Krause (Eds.). *Proceedings of the 35th International Conference on Machine Learning* 80, 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html

[103] Michael Kearns and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* Oxford University Press, Inc.

[104] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290607.3310433

[105] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. (9 2016).

[106] Daniel Kluttz, Nitin Kohli, and Deirdre K. Mulligan. 2018. Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems. *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3311894

[107] kobi leins, Jey Han Lau, and Timothy Baldwin. 2020. Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* https://doi.org/10.18653/v1/2020.acl-main.261

[108] TD Krafft and K Zweig. 2019. Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse. *Ein Regulierungsvorschlag* (2019).

[109] Lenneke Kuijer and Elisa Giaccardi. 2018. Co-Performance: Conceptualizing the Role of Artificial Agency in the Design of Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173699

[110] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2018. POTs: Protective Optimization Technologies. (6 2018). https://doi.org/10.1145/3351095.3372853

[111] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness, I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (Eds.). *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[112] Claire Larsonneur. 2021. Intelligence artificielle ET/OU diversité linguistique : les paradoxes du traitement automatique des langues. http://www.hybrid.univ-paris8.fr/lodel/index.php?id=1542

[113] Douglass B. Lee. 1973. Requiem for Large-Scale Models. *Journal of the American Institute of Planners* 39 (5 1973), 163–178. Issue 3. https://doi.org/10.1080/01944367308977851

[114] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1035–1048. https://doi.org/10.1145/2998181.2998230

[115] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3 (11 2019). Issue CSCW. https://doi.org/10.1145/3359283

[116] Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk Identification Questionnaire for Unintended Bias in Machine Learning Development Lifecycle. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3777093

[117] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. (1 2020). https://doi.org/10.1145/3313831.3376590

[118] Q. Vera Liao, Milena Pribić, Jaesik Han, Sarah Miller, and Daby Sow. 2021. Question-Driven Design Process for Explainable AI User Experiences. (4 2021).

[119] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 799–808.

[120] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

[121] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. (2 2021). https://doi.org/10.1145/3449180

[122] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. 2015. A Unified Gradient Regularization Family for Adversarial Examples. *2015 IEEE International Conference on Data Mining*, 301–309. https://doi.org/10.1109/ICDM.2015.84

[123] N. Mahendran. 2021. Analysis of memory consumption by neural networks based on hyperparameters. (10 2021).

[124] Donald Martin, Jr Google Vinodkumar Prabhakaran Google Jill Kuhlberg, and Andrew S Smart Google William Isaac DeepMind. 2020. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. (2020).

[125] M. Mehldau. 2007. Iconset for data-privacy declarations v 0.1. https://netzpolitik.org/wp-upload/data-privacy-icons-v01.pdf
[126] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54 (7 2021). Issue 6. https://doi.org/10.1145/3457607
[127] Microsoft. 2018. AI Principles. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6
[128] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3411764.3445096
[129] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model Cards for Model Reporting. (10 2018). https://doi.org/10.1145/3287560.3287596
[130] Tanushree Mitra. 2021. Provocation: Contestability in Large-Scale Interactive {NLP} Systems. *Proceedings of the First Workshop on Bridging Human{–}Computer Interaction and Natural Language Processing*, 96–100.
[131] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1 (2019), 501–507. Issue 11. https://doi.org/10.1038/s42256-019-0114-4
[132] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics* 26 (2020), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5
[133] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2019. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. (5 2019). https://doi.org/10.1145/3351095.3372850
[134] Pradeep K Murukannaiah and Munindar P Singh. 2014. Xipho: Extending Tropos to Engineer Context-Aware Personal Agents. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 309–316.
[135] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine Learning with Membership Privacy using Adversarial Regularization. (7 2018).
[136] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. (9 2019).
[137] Arif Nurwidyantoro, Mojtaba Shahin, Michel Chaudron, Waqar Hussain, Harsha Perera, Rifat Ara Shams, and Jon Whittle. 2021. Towards a Human Values Dashboard for Software Development: An Exploratory Study. (7 2021).
[138] OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0406
[139] Kieron O'Hara. 2020. Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review* 39 (11 2020), 105474. https://doi.org/10.1016/J.CLSR.2020.105474
[140] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (first edition ed.). IEEE.
[141] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *2016 IEEE Symposium on Security and Privacy (SP)*, 582–597. https://doi.org/10.1109/SP.2016.41
[142] Lorenza Parisi and Francesca Comunello. 2020. Dating in the time of "relational filter bubbles": exploring imaginaries, perceptions and tactics of Italian dating app users. *The Communication Review* 23 (2020), 66–89. Issue 1. https://doi.org/10.1080/10714421.2019.1704111
[143] Reema Patel. 2021. Reboot AI with human values. *Nature* 598 (10 2021). Issue 7879. https://doi.org/10.1038/d41586-021-02693-2
[144] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. (12 2020). http://arxiv.org/abs/2012.05345
[145] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn Jonker. 2012. Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology* 14 (12 2012), 285–303. Issue 4. https://doi.org/10.1007/s10676-011-9282-6
[146] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828
[147] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873
[148] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virg{́ı}lio A F Almeida, and Wagner Meira. 2020. Auditing Radicalization Pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879
[149] Arianna Rossi and Monica Palmirani. 2017. A Visualization Approach for Adaptive Consent in the European Data Protection Framework. *2017 Conference for E-Democracy and Open Government (CeDEM)*, 159–170. https://doi.org/10.1109/CeDEM.2017.23
[150] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36 (12 2015). Issue 4. https://doi.org/10.1609/aimag.v36i4.2577
[151] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit.
[152] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry 22.*
[153] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (2019), 1668–1678. https://doi.org/10.18653/V1/P19-1163
[154] Enrique Schaerer, Richard Kelley, and Monica Nicolescu. 2009. Robots as animals: A framework for liability and responsibility in human-robot interactions. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication.* https://doi.org/10.1109/ROMAN.2009.5326244
[155] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 317* (2021). https://doi.org/10.1145/3476058
[156] Shalom H. Schwartz. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture* 2 (12 2012). Issue 1. https://doi.org/10.9707/2307-0919.1116
[157] National Science and United States Executive Office of the President Technology Council Committee on Technology. 2016. Preparing for the Future of Artificial Intelligence. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
[158] Mojtaba Shahin, Waqar Hussain, Arif Nurwidyantoro, Harsha Perera, Rifat Shams, John Grundy, and Jon Whittle. 2021. Operationalizing Human Values in Software Engineering: A Survey. (8 2021).
[159] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. (5 2021). https://doi.org/10.1145/3479577
[160] Irina Shklovski and Carolina Némethy. 2022. Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society* (1 2022), 1–17. https://doi.org/10.1080/1369118X.2021.2014547
[161] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Trans. Interact. Intell. Syst.* 10 (10 2020). Issue 4. https://doi.org/10.1145/3419764
[162] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership Inference Attacks against Machine Learning Models. (10 2016).
[163] Dasha Simons. 2019. Design for fairness in AI: Cooking a fair AI Dish. http://resolver.tudelft.nl/uuid:5a116c17-ce0a-4236-b283-da6b8545628c
[164] The Royal Society. 2019. Explainable AI: the basics. https://doi.org/10.1177/1461444816676645
[165] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* https://doi.org/10.1145/3351095.3372870
[166] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2459–2468. https://doi.org/10.1145/3292500.3330664
[167] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445088
[168] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305
[169] Sarah Thew and Alistair Sutcliffe. 2018. Value-based requirements engineering: method and experience. *Requirements Engineering* 23 (11 2018). Issue 4. https://doi.org/10.1007/s00766-017-0273-y
[170] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 83–92. https://doi.org/10.1145/3322640.3326705
[171] Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the Ethical Limits of Natural Language Processing on Legal Text. (5 2021).

[172] Funda Ustek-Spilda, Alison Powell, and Selena Nemorin. 2019. Engaging with ethics in Internet of Things: Imaginaries in the social milieu of technology developers. *Big Data & Society* 6 (7 2019), 205395171987946. Issue 2. https://doi.org/10.1177/2053951719879468

[173] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants". *Proceedings of the ACM on Human-Computer Interaction* 4 (10 2020), 1–22. Issue CSCW2. https://doi.org/10.1145/3415238

[174] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445365

[175] Ibo van de Poel. 2013. Translating Values into Design Requirements. , 253-266 pages. https://doi.org/10.1007/978-94-007-7762-0_20

[176] W. Fred van Raaij and Theo M.M. Verhallen. 1994. Domain-specific Market Segmentation. *European Journal of Marketing* 28 (10 1994), 49–66. Issue 10. https://doi.org/10.1108/03090569410075786

[177] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

[178] Sandra Wachter and Brent Mittelstadt. 2019. Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review* 2 (2019), 494–620.

[179] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376813

[180] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. (7 2019). https://doi.org/10.1109/TVCG.2019.2934619

[181] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. https://doi.org/10.1145/3442188.3445928

[182] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109 (1980), 121–136. Issue 1. http://www.jstor.org/stable/20024652

[183] Pulei Xiong, Scott Buffett, Shahrear Iqbal, Philippe Lamontagne, Mohammad Mamun, and Heather Molyneaux. 2021. Towards a Robust and Trustworthy Machine Learning System Development. (1 2021).

[184] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. *2018 IEEE International Conference on Big Data (Big Data)*, 570–575. https://doi.org/10.1109/BigData.2018.8622525

[185] An Yan and Bill Howe. 2020. Fairness-Aware Demand Prediction for New Mobility. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (4 2020), 1079–1087. Issue 01. https://doi.org/10.1609/aaai.v34i01.5458

[186] Vahid Yazdanpanah, Enrico Gerding, Sebastian Stein, Mehdi Dastani, Catholijn M Jonker, and Timothy Norman. 2021. Responsibility Research for Trustworthy Autonomous Systems. *20th International Conference on Autonomous Agents and Multiagent Systems (03/05/21 - 07/05/21)*, 57–62. https://eprints.soton.ac.uk/447511/

[187] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, and Reza Shokri. 2021. Enhanced Membership Inference Attacks against Machine Learning Models. (11 2021).

[188] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

[189] Angela Zhou, David Madras, Inioluwa Raji Raji, Bogdan Kulynych, Smitha Mili, and Richard Zemel. [n. d.]. Call for participation: Participatory Approaches to Machine Learning. https://participatoryml.github.io/

[190] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. 2021. AI and Ethics – Operationalising Responsible AI. (5 2021).

[191] Christian Zimmermann, Rafael Accorsi, and Gunter Muller. 2014. Privacy Dashboards: Reconciling Data-Driven Business Models and Privacy. *2014 Ninth International Conference on Availability, Reliability and Security*, 152–157. https://doi.org/10.1109/ARES.2014.27

[192] Arkaitz Zubiaga, Bo Wang, Maria Liakata, and Rob Procter. 2019. Political Homophily in Independence Movements: Analyzing and Classifying Social Media Users by National Identity. *IEEE Intelligent Systems* 34 (2019), 34–42. Issue 6. https://doi.org/10.1109/MIS.2019.2958393

A FROM VALUES TO SPECIFIC CRITERIA MANIFESTATIONS

	Value	Criteria	Manifestations
	Privacy	(1) **Consent for data usage** [3, 38, 60] (2) **Data protection** [3, 60, 61] (3) **Control over data / ability to restrict processing** [38, 60] (4) **Right to rectification** [3, 38, 60] (5) **Right to erase the data** [3, 38, 60] (6) **Right of access by data subject, data agency** [38, 140]	• Written declaration of consent [38] • Description of what data is collected [125] • Description of how data is handled [125] • Purpose statement of data collection [125] • Statement of how long the data is kept [125] • Form and submission mechanisms to object data collection and to make complaints [27] • Obfuscation of data [3]
Conservation	Security	(1) **Resilience to attacks**: protection of privacy [85, 127, 178], vulnerabilities, fallback plans [3, 60, 74, 132] (2) **Predictability** [3, 39, 60] (3) **Robustness / reliability**: prevent manipulation [3]	AGAINST INTEGRITY THREATS [183]: • Training time [183] Ex.: – Data sanitization [13] [23, 43] – Robust learning [14] [23, 72] • Prediction time [183] – Model enhancement [23, 73, 122, 141] Ex.: ∗ Adversarial Learning [15] ∗ Gradient masking [16] ∗ Defensive Distillation [17] AGAINST PRIVACY THREATS [183]: • Mitigation techniques [135]: – Restrict prediction vector to top k classes [18] [162] – Coarsen the precision of the prediction vector [19] [162] – Increase entropy of the prediction vector [20] [162] – Use regularization [21] [101, 162] • Differential privacy mechanisms [135]: – Differential privacy [22] [55, 187]. Ex.: ∗ Adversarial regularization [23] [135] ∗ MemGuard [24] [97]

[13] It ensures data soundness by identifying abnormal input samples and by removing them [183].
[14] It ensures that algorithms are trained on statistically robust datasets, with little sensitivity to outliers [183].
[15] Adversarial samples are introduced to the training set [183].
[16] Input gradients are modified to enhance model robustness [183].
[17] The dimensionality of the network is reduced [183].
[18] Applicable when the number of classes is very large. Even if the model only outputs the most likely k classes, it will still be useful [162].
[19] It consists in rounding the classification probabilities down [162].
[20] Modification of the softmax layer (in neural networks) to increase its normalizing temperature [162].
[21] Technique to avoid overfitting in ML that penalizes large parameters by adding a regularization factor λ to the loss function [162].
[22] It prevents any adversary from distinguishing the predictions of a model when its training dataset is used compared to when other dataset is used [187]
[23] Membership privacy is modeled as a min-max optimization problem, where a model is trained to achieve minimum loss of accuracy and maximum robustness against the strongest inference attack [135].
[24] Noise is added to the confidence vector of the attacker so as to mislead the attacker's classifier [97]

	Value	Criteria	Manifestations
Conservation	Performance	(1) **Correctness of predictions** [26, 39, 60] (2) **Memory efficiency** [3, 26] (3) **Training efficiency** [26] (4) **Energy efficiency** [3, 26] (5) **Data efficiency** [26]	• Accuracy (for classification, sum of true positive and true negative rates) [129, 180] • False Positive and False Negative rates [129, 180] • False Discovery and Omission Rate [129] • Mean and median error [180] • R2 score [25] • Precision and recall rates [180] • Area under ROC curve (AUC) [25] • Estimation of energy consumption through [67]: – performance counters – simulation – instruction- or architecture-level estimations – real-time estimation • Estimation of GPU memory consumption [66, 123] • Wall-clock training time [15, 44]
Universalism	Respect for public interest	(1) **Desirability of technology** [1, 34, 104] (2) **Benefit to society** [60–62, 132] (3) **Environmental impact** [3, 21]	• Diverse and inclusive forum for discussion [14, 60] • Measure of social and environmental impact [21, 132, 147]
	Fairness	(1) **Individual fairness** [25] [13, 54, 111, 126] (2) **Demographic parity** [26] [13, 54, 79, 85, 102, 111, 126, 166, 177] (3) **Conditional Statistical parity** [27] [126, 177] (4) **Equality of opportunity** [28] [78, 126, 174] (5) **Equalized odds** [29] [126] (6) **Treatment equality** [30] [22, 126] (7) **Test fairness** [31] [36, 126, 177] (8) **Procedural fairness** [32] [76, 111, 126]	• Accuracy across groups (for classification, sum of true positive and true negative rates) [36, 79, 105, 132] • False positive and negative rates across groups [36, 105, 126, 151, 179] • False discovery and omission rates across groups [129, 151] • Pinned AUC [50, 129] • Debiasing algorithms [19] • Election of protected classes based on user considerations [76]

[25] Similar individuals should be treated in a similar way. Diverging definitions state that: two individuals that are similar with respect to a common metric should receive the same outcome (*fairness through awareness*); or any protected attribute should not be used when making a decision (*fairness through unawareness*); or the outcome obtained by an individual should be the same if this individual belonged to a counterfactual world or group (*counterfactual fairness*) [126].
[26] The probability of getting a positive outcome should be the same whether the individual belongs to a protected group or not [126].
[27] Given a set of factors L, individuals belonging to the protected or unprotected group should have the same probability of getting a positive outcome [126].
[28] The probability for a person from class A (positive class) of getting a positive outcome, which should be the same regardless of the group (protected group or not) that the individual belongs to [126].
[29] The probability for a person from class A (positive class) of getting a positive outcome and the probability for a person from class B (negative class) of getting a negative outcome should be the same [126].
[30] The ratio of false positives and negatives has to be the same for both groups [126].
[31] For any probability score S, the probability of correctly belonging to the positive class should be the same for both the protected and unprotected group [126].
[32] It deals with the fairness of the decision-making process that leads to the outcome in question [76].

	Value	Criteria	Manifestations
	Non-discrimination	(1) **Quality and integrity of data** [60, 69, 85, 132, 144] (2) **Inclusiveness in design** [39, 60, 132] (3) **Accessibility** [3, 26, 60, 132]	• Inclusive data generation process [3, 34, 69, 132] • Analysis of data for potential biases, data quality assessment [3, 60, 68, 85, 126] • Diversity of participant in development process [3, 60, 115, 189] • Access to code and technology to all [3, 26, 60, 132]
Openness	Transparency	(1) **Interpretability of data and models** [26, 164] (2) **Enabling human oversight of operations** [60, 132] (3) **Accessibility of data and algorithm** [3, 60, 164] (4) **Traceability** [132] (5) **Reproducibility** [26]	• Description of data generation process [3, 20, 34, 68, 69, 132] • Disclosure of origin and properties of models and data [3, 129, 164] • Open access to data and algorithm [3, 26, 60, 164] • Notification of usage/interaction [60] • Regular reporting [60]
Openness	Explainability	(1) **Ability to understand AI systems and the decision reached** [26, 39, 61, 62, 138, 164] (2) **Traceability** [132] (3) **Enable evaluation** [60, 132]	• Interpretability by design [13] • Post-hoc explanations [13]
Individual empowerment	Contestability	(1) **Enable argumentation / negotiation against a decision** [6, 17, 39, 60, 100, 114, 121, 164] (2) **Citizen empowerment** [17, 39, 100]	• Information of who determines and what constitutes a contestable decision and who is accountable [121] • Determination of who can contest the decision (subject or representative) [121] • Indication of type of review in place [121] • Information regarding the contestability workflow [121] • Mechanisms for users to ask questions and record disagreements with system behavior [86, 130]
Individual empowerment	Human Control	(1) **User/collective influence** [26, 114] (2) **Human review of automated decision** [60] (3) **Choice of how and whether to delegate** [60]	• Continuous monitoring of system to intervene [39, 40, 60] • Establishment levels of human discretion during the use of the system [39, 127] • Ability to override the decision made by a system [39]
Individual empowerment	Human agency	(1) **Respect for human autonomy** [39, 60, 132] (2) **Power to decide. Ability to make informed autonomous decision** [26, 39] (3) **Ability to opt out of an automated decision** [39, 60]	• Give knowledge and tools to comprehend and interact with AI system [39] • Opportunity to self-assess the system [39]

Table 3: Summary of the specific criteria that relate to each value considered in our ML assessment framework. These criteria are then translated into specific manifestations in the form of signifiers (orange), process-oriented practices (olive) or quantifiable indicators (magenta).

B MAPPING STAKEHOLDERS

Stakeholder	**Mapping** [167]	**Nature of knowledge**	**Purpose of insight**
Development team	ML, Formal + Instrumental + Personal	• "Knowledge of the math behind the architecture" [167] • "Stakeholder involved in an ex-ante impact assessment of the automatic decision system"[83]	• Ensure/improve product efficiency and debug [13] • Research new functionalities [13]
Auditing team	Milieu, Formal + Instrumental	• "Familiarity with broader ML-enabled systems" [167] • "Experts who intervene wither upstream or downstream" [83]	• Verify model compliance with legislation [13]
Data domain experts	Data domain, Formal + Instrumental	• "Theories relevant to the data domain" [167] • "Professional involved in the operational phase of the automatic decision system" [83]	• Gain scientific or domain-specific knowledge [13, 167] • Trust the model [13, 167] • Act based on the output [167]
Decision subjects	Data domain + Milieu, Personal	• "Lived experience and cultural knowledge" [167] • "Layperson affected by the outcomes of the automatic decision system" [83]	• Understand their situation [13] • Verify fair decision [13] • Contest decision [167] • Understand how one's data is being used [167]

Table 4: Description of potential stakeholders that can be brought together as part of our value-based framework. These stakeholders have been mapped following the two dimensional criteria (type of knowledge —formal, instrumental or personal— and contexts in which this knowledge manifests —ML, data domain, milieu—) outlined by Suresh et al. [167]. The nature of their knowledge and the purpose of gaining insight for each of them have also been defined.

C TAILORED COMMUNICATION OF SYSTEM-RELATED INFORMATION

		Development team	Auditing team	Data Domain experts	Decision subjects
Conservation	Privacy	[K]	[K]		[A] [B]
	Security	[K] [W] [AB]	[K] [W]		
	Performance	[F] [G] [H] [Y] [Z] [AE]	[G] [H] [Y] [Z] [AE]	[I] [J]	[J]
Universalism	Respect for public interest	[E] [AE]	[E] [AE]	[E]	[C] [D]
	Fairness	[G] [H] [K] [W] [X] [Y] [Z] [AD]	[G] [H] [K] [W] [X] [Y] [Z] [AD]	[I] [J]	[J]
	Non-discrimination	[H] [K] [X] [Y] [AD]	[H][K] [X] [Y] [AD]	[J] [L]	[J] [L]
Openness	Transparency	[H] [K] [M]	[H][K] [M]	[I] [J] [L] [M]	[B] [J] [L] [M]
	Explainability	[M] [N] [O] [Q] [AC] [AD] [P]	[M] [N] [O] [Q] [AC] [AD] [P]	[J] [M] [N] [O] [Q] [P]	[J] [M] [N] [O] [Q] [R] [S] [P]
Individual empowerment	Contestability	[U]	[U]	[T] [U]	[T] [AF]
	Human Control	[V]	[V]	[T] [V]	[C] [T] [V]
	Human Agency			[T]	[T] [B] [AA]

Table 5: Mapping of available means for transmitting value-specific manifestations to different stakeholders based on the purpose of their insight and the nature of their knowledge. These means have been classified into three main categories: descriptive documents specifying whether/how a value manifestation is fulfilled (red), strategies for fulfilling value manifestations (blue), and complete tools for enabling the fulfillment of value manifestations (green). This table aims at facilitating the navigation of table 6, where each means is documented.

	Means	Value	Manifestation(s)	Stakeholder				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[A]	Iconsets for data privacy declarations [38, 88, 125, 149]	Privacy	• Description of what data is collected • Description of how data is handled • Purpose statement of data collection • Statement of how long the data is kept				✓	Agnostic		Iconsets	
[B]	Privacy dashboards [56, 58, 59, 84, 191]	Privacy	• Description of what data is collected • Description of how data is handled • Purpose statement of data collection				✓	Agnostic		• Timelines • Bar charts • Maps • Network graphs	
		Human agency	• Opportunity to self-assess the system								
		Trans-parency	• Disclosure of origin and properties of data								
[C]	Risk matrix [3, 108]	Respect for public interest	• Measure of social impact				✓	Agnostic		Two dimensional space (vulnerability vs dependence of the decision)	
		Human Control	• Ability to override the decision made by a system								
[D]	Moral space [85]	Respect for public interest	• Measure of social impact				✓	Agnostic	Based on human judgement	Three dimensional moral space. Wrongness as a function of intention and harm	
[E]	Social impact assessment [147]	Respect for public interest	• Measure of social impact	✓	✓	✓		Agnostic	Anticipate scenarios		

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[F]	Model Tracker interactive visualization [9]	Performance	• Accuracy • False Positive and Negative rates	✓				Classification tasks		• Summary statistics • Confusion matrices • Labels chart • Precision-recall curves • Connector lines to identify similar examples in feature space • Highlighted boxes for correlations between features and target classes	
[G]	Model cards for models [129]	Performance	• Accuracy • False Positive and Negative rates • False Discovery and omission rates	✓	✓			Agnostic		• Confidence bars • Bar charts	
		Fairness	• Accuracy across groups • False Positive and Negative rates across groups • False Discovery and omission rates across groups								

	Means	Value	Manifestation(s)	Stakeholder				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[H]	What-if tool [33] [180]	Performance	• Accuracy • False Positive and Negative Rates • False Discovery and omission rates	✓	✓			Classification tasks, Regression tasks		• Confusion matrices • (Two-dimensional) Histograms • Scatterplots • Summary statistics of datasets • Partial dependence plots	Interactive modules include: list of feature values, inference values, and counterfactual controls
		Fairness	• Accuracy across groups • False Positive and Negative Rates across groups • False Discovery and omission rates across groups								
		Transparency	• Disclosure of origin and properties of data								
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[I]	Interactive transfer learning tools [128]	Performance	• Accuracy • False Positive and Negative Rates			✓		Convolutional Neural Networks		• Confusion matrices • Z-scored of each filter • Bar charts • Activation heatmaps • t-SNE clusters	
		Fairness	• Accuracy across groups • False Positive and Negative Rates across groups								
		Transparency	• Disclosure of properties of data								

[33] https://github.com/pair-code/what-if-tool

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[J]	Question-Driven XAI Design [118]	Performance	• Accuracy			✓	✓	Agnostic		• Summary statistics (percentage scores) for data explanations and performance metrics • Feature importance • Contrastive explanations	End users were more interested in the limitation of the model: uncertainty
		Fairness	• Accuracy across groups								
		Transparency	• Disclosure of origin and properties of data								
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
		Explainability	• Post-hoc explanations								
[K]	Datasheets for datasets [68]	Transparency	• Description of data generation process • Disclosure of origin properties of models and data	✓	✓			Agnostic		• Summary statistics • Visual examples of datasets (if images, for instance)	
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
		Privacy	• Written declaration of consent • Description of what data is collected • Description of how data is handled • Purpose statement of data collection • Statement of how long the data is kept								
		Fairness	• Election of protected classes								
		Security	• Membership inference								

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[L]	Data centric explanations [11]	Transparency	• Description of data generation process • Disclosure of origin and properties of the models and data			✓	✓	Agnostic		• Interactive list • Q&A format • Pie charts • Bar charts • Process diagrams • Timelines • Icons	
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[M]	Example-based explanations [13, 24, 32, 51, 98, 117, 118]	Transparency	• Disclosure of properties of data	✓	✓	✓	✓	Agnostic	• Similar example • Typical example • Counterfactual ex	• Example images from dataset if in the visual domain	Normative vs comparative explanations [32]
		Explainability	• Post-hoc explanations								
[N]	Explanation by simplification [13, 98]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Decision rule • Decision tree		
[O]	Feature relevance explanation [7, 13, 24, 51, 98, 118]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Feature attribute • Feature shape • Feature interaction • Sensitivity / perturbation -based • Saliency maps (visual domain)	• Bar charts • Visualization of element importance, saliency (visual domain)	Usability of saliency maps for non-experts [7]. They should be accompanied by global descriptors

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[P]	Contrastive explanations [49, 118, 133]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Example of minimum change that leads to different outcomes		
[Q]	Text-based explanation [13, 174]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• With or without outcome comparison		
[R]	Interactive demonstrations [120]	Explainability	• Post-hoc explanations				✓	Agnostic			
[S]	Experiential AI [81]	Explainability	• Post-hoc explanations				✓	Agnostic	• Art mediated between computer code and human comprehension		
[T]	Interactive contestations [83, 106]	Contestability	• Mechanisms for users to ask questions and record disagreements with system behavior			✓	✓	Agnostic	• Statements restricted to natural language		
		Human Control	• Ability to override the decision made by the system								
		Human agency	• Opportunity to self-assess the system								

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[U]	Challenge justifications provided by operator using the same means [83]	Contestability	• Mechanisms for users to ask questions and record disagreements with system behavior	✓	✓	✓		Agnostic	Further testing Verification		
[V]	Mapping of actors and tasks depending on automation level [33]	Human Control	• Establishment of levels of human discretion during the use of the system	✓	✓	✓	✓	Agnostic		• Relationship diagrams	
[W]	Failure Modes and Effects Analysis [147]	Security	• Threats against integrity (adversarial learning) and mitigation techniques	✓	✓			Agnostic			
		Fairness	• Accuracy across groups • False positives and negatives across groups								
[X]	Aequitas [34] [151]	Fairness	• Accuracy across groups • False Positive and Negative rates across groups • False Discovery and Omission rates across groups • Counterfactual examples	✓	✓			Agnostic			
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								

[34]https://github.com/dssg/aequitas

	Means	Value	Manifestation(s)	Stakeholder				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[Y]	AI Fairness 360 [35] [19]	Performance	• False Positive and Negative rates	✓	✓			Classifiers: logistic regression, random forest classifier and neural networks		• Bar charts • Confidence bars	
		Fairness	• False positive and negative rates across groups • Debiasing algorithms								
		Non-discrimination	• Analysis of data for potential biases, data quality assessment								
[Z]	Fairlearn [36] [25]	Performance	• Accuracy • False Positive and False Negative rates • Precision and recall rates	✓	✓			Agnostic		• Bar charts • Pie charts	
		Fairness	• Accuracy across groups • False negative and false positive rates across groups • Debiasing algorithms								
[AA]	Playbook AI [37] [89]	Human agency	• Give knowledge and tools to comprehend and interact with AI systems • Opportunity to self-assess the system				✓	NLP	Early AI prototyping	• Interactive survey	
[AB]	Counterfit [38]	Security	• Defence against integrity threats • Defence against privacy threats	✓				Agnostic			

[35] https://github.com/Trusted-AI/AIF360
[36] https://github.com/fairlearn/fairlearn
[37] https://github.com/microsoft/HAXPlaybook
[38] https://github.com/Azure/counterfit

Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[AC] InterpretML [39] [40] [133, 136]	Explain-ability	• Interpretability by design • Post-hoc explanations	✓	✓			Both whitebox and blackbox models		• Bar charts • Line charts • Decision trees	
[AD] Error analysis dashboard [41]	Non-discrimination	• Analysis of data for potential biases, data quality assessment	✓	✓			Agnostic		• Decision tree • Error heatmap	
	Explain-ability	• Post-hoc explanations								
	Fairness	• Accuracy across groups								
[AE] Breakend Impact tracker [42] [82]	Perfor-mance	• Estimation of energy consumption • Estimation of GPU memory consumption	✓	✓			Agnostic		• Dot plots • Bar charts	
	Respect for public interest	• Measure of environmental impact								
[AF] Representative contestations [173]	Contest-ability	• Mechanisms for users to ask questions and record disagreement with system behaviour				✓	Agnostic			

Table 6: Mapping of available means for transmitting value-specific manifestations to different stakeholders based on the purpose of their insight and the nature of their knowledge (DT = Development Team; AT = Auditing Team; DE = Data Domain Experts; DS = Decision Subjects). The identification and color code correspond to those on table 5. Each means is linked to the value and criteria manifestations that they communicate, the stakeholders that the original papers address, model specificity, deployed approach, visual elements and any additional details.

[39] https://github.com/interpretml/interpret/
[40] https://github.com/interpretml/DiCE
[41] https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/erroranalysis-dashboard-README.md
[42] https://github.com/Breakend/experiment-impact-tracker