A FROM VALUES TO SPECIFIC CRITERIA MANIFESTATIONS

	Value	Criteria	Manifestations
Conservation	Privacy	(1) **Consent for data usage** [3, 56, 60] (2) **Data protection** [3, 60, 61] (3) **Control over data / ability to restrict processing** [56, 60] (4) **Right to rectification** [3, 56, 60] (5) **Right to erase the data** [3, 56, 60] (6) **Right of access by data subject, data agency** [56, 167]	• Written declaration of consent [56] • 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]

	Security	(1) **Resilience to attacks**: protection of privacy [86, 127, 178], vulnerabilities, fallback plans [3, 60, 75, 133]
		(2) **Predictability** [3, 57, 60]
		(3) **Robustness / reliability**: prevent manipulation [3]
			AGAINST INTEGRITY THREATS [183]:
			• Training time [183] Ex.:
			  • Data sanitization [12] [23, 40]
			  • Robust learning [13] [23, 73]
			• Prediction time [183]
			  • Model enhancement [23, 74, 122, 141] Ex.:
			    • Adversarial Learning [14]
			    • Gradient masking [15]
			    • Defensive Distillation [16]
			
			AGAINST PRIVACY THREATS [183]:
			• Mitigation techniques [136]:
			  • Restrict prediction vector to top k classes [17] [161]
			  • Coarsen the precision of the prediction vector [18] [161]
			  • Increase entropy of the prediction vector [19] [161]
			  • Use regularization [20] [101, 161]
			• Differential privacy mechanisms [136]:
			  • Differential privacy [21] [53, 187]. Ex.:
			    • Adversarial regularization [22] [136]
			    • MemGuard [23] [97]

[12] It ensures data soundness by identifying abnormal input samples and by removing them [183].
[13] It ensures that algorithms are trained on statistically robust datasets, with little sensitivity to outliers [183].
[14] Adversarial samples are introduced to the training set [183].
[15] Input gradients are modified to enhance model robustness [183].
[16] The dimensionality of the network is reduced [183].
[17] 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 [161].
[18] It consists in rounding the classification probabilities down [161].
[19] Modification of the softmax layer (in neural networks) to increase its normalizing temperature [161].
[20] Technique to avoid overfitting in ML that penalizes large parameters by adding a regularization factor λ to the loss function [161].
[21] 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]
[22] 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 [136].
[23] 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, 57, 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) [130, 180]
			• False Positive and False Negative rates [130, 180]
			• False Discovery and Omission Rate [130]
			• Mean and median error [180]
			• R2 score [25]
			• Precision and recall rates [180]
			• Area under ROC curve (AUC) [25]
			• Estimation of energy consumption through [68]:
			  • performance counters
			  • simulation
			  • instruction- or architecture-level estimations
			  • real-time estimation
			• Estimation of GPU memory consumption [67, 123]
			• Wall-clock training time [14, 41]

Universalism	Respect for public interest	(1) **Desirability of technology** [1, 34, 104] (2) **Benefit to society** [60–62, 133] (3) **Environmental impact** [3, 21]	• Diverse and inclusive forum for discussion [60, 129] • Measure of social and environmental impact [21, 133, 147]
	Fairness	(1) **Individual fairness** [24] [18, 52, 110, 126] (2) **Demographic parity** [25] [18, 52, 80, 86, 102, 110, 126, 163, 177] (3) **Conditional Statistical parity** [26] [126, 177] (4) **Equality of opportunity** [27] [79, 126, 175] (5) **Equalized odds** [28] [126] (6) **Treatment equality** [29] [22, 126] (7) **Test fairness** [30] [37, 126, 177] (8) **Procedural fairness** [31] [77, 110, 126]	• Accuracy across groups (for classification, sum of true positive and true negative rates) [37, 80, 105, 133] • False positive and negative rates across groups [37, 105, 126, 151, 179] • False discovery and omission rates across groups [130, 151] • Pinned AUC [48, 130] • Debiasing algorithms [19] • Election of protected classes based on user considerations [77]
	Non-discrimination	(1) **Quality and integrity of data** [60, 70, 86, 133, 144] (2) **Inclusiveness in design** [57, 60, 133] (3) **Accessibility** [3, 26, 60, 133]	• Inclusive data generation process [3, 34, 70, 133] • Analysis of data for potential biases, data quality assessment [3, 60, 69, 86, 126] • Diversity of participant in development process [3, 60, 114, 189] • Access to code and technology to all [3, 26, 60, 133]

[24] 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].

[25] The probability of getting a positive outcome should be the same whether the individual belongs to a protected group or not [126].

[26] 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].

[27] 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].

[28] 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].

[29] The ratio of false positives and negatives has to be the same for both groups [126].

[30] 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].

[31] It deals with the fairness of the decision-making process that leads to the outcome in question [77].

	Value	Criteria	Manifestations
Openness	Transparency	(1) **Interpretability of data and models** [26, 168] (2) **Enabling human oversight of operations** [60, 133] (3) **Accessibility of data and algorithm** [3, 60, 168] (4) **Traceability** [133] (5) **Reproducibility** [26]	• Description of data generation process [3, 20, 34, 69, 70, 133] • Disclosure of origin and properties of models and data [3, 130, 168] • Open access to data and algorithm [3, 26, 60, 168] • Notification of usage/interaction [60] • Regular reporting [60]
	Explainability	(1) **Ability to understand AI systems and the decision reached** [26, 57, 61, 62, 139, 168] (2) **Traceability** [133] (3) **Enable evaluation** [60, 133]	• Interpretability by design [18] • Post-hoc explanations [18]
Individual empowerment	Contestability	(1) **Enable argumentation / negotiation against a decision** [6, 16, 57, 60, 100, 113, 121, 168] (2) **Citizen empowerment** [16, 57, 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 [87, 131]
	Human Control	(1) **User/collective influence** [26, 113] (2) **Human review of automated decision** [60] (3) **Choice of how and whether to delegate** [60]	• Continuous monitoring of system to intervene [57, 60, 166] • Establishment levels of human discretion during the use of the system [57, 127] • Ability to override the decision made by a system [57]
	Human agency	(1) **Respect for human autonomy** [57, 60, 133] (2) **Power to decide. Ability to make informed autonomous decision** [26, 57] (3) **Ability to opt out of an automated decision** [57, 60]	• Give knowledge and tools to comprehend and interact with AI system [57] • Opportunity to self-assess the system [57]

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** [164]	**Nature of knowledge**	**Purpose of insight**
Development team	ML, Formal + Instrumental + Personal	• "Knowledge of the math behind the architecture" [164] • "Stakeholder involved in an ex-ante impact assessment of the automatic decision system"[84]	• Ensure/improve product efficiency and debug [18] • Research new functionalities [18]
Auditing team	Milieu, Formal + Instrumental	• "Familiarity with broader ML-enabled systems" [164] • "Experts who intervene wither upstream or downstream" [84]	• Verify model compliance with legislation [18]
Data domain experts	Data domain, Formal + Instrumental	• "Theories relevant to the data domain" [164] • "Professional involved in the operational phase of the automatic decision system" [84]	• Gain scientific or domain-specific knowledge [18, 164] • Trust the model [18, 164] • Act based on the output [164]
Decision subjects	Data domain + Milieu, Personal	• "Lived experience and cultural knowledge" [164] • "Layperson affected by the outcomes of the automatic decision system" [84]	• Understand their situation [18] • Verify fair decision [18] • Contest decision [164] • Understand how one's data is being used [164]

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. [164]. The nature of their knowledge and the purpose of gaining insight for each of them have also been defined.

A 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 14. 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 15, where each means is documented.

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[A]	Iconsets for data privacy declarations [56, 89, 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 [54, 58, 59, 85, 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								
		Transparency	• Disclosure of origin and properties of data								
[C]	Risk matrix [3, 107]	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 [86]	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	

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[E]	Social impact assessment [147]	Respect for public interest	• Measure of social impact	✓	✓	✓		Agnostic	Anticipate scenarios		
[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 [130]	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 DT	Stakeholder AT	Stakeholder DE	Stakeholder DS	Application (model)	Approach	Visual elements	Additional details
[H]	What-if tool [32] [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								
[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								

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

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
		Explainability	• Post-hoc explanations								
[K]	Datasheets for datasets [69]	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								
[L]	Data centric explanations [12]	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								

	Means	Value	Manifestation(s)	Stakeholder				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[M]	Example-based explanations [18, 24, 32, 49, 98, 117, 118]	Transparency	• Disclosure of properties of data	✓	✓	✓	✓	Agnostic	• Similar example • Typical example • Counter-factual example	• Example images from dataset if in the visual domain	Normative vs comparative explanations [32]
		Explainability	• Post-hoc explanations								
[N]	Explanation by simplification [18, 98]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Decision rule • Decision tree		
[O]	Feature relevance explanation [7, 18, 24, 49, 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
[P]	Contrastive explanations [47, 118, 134]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• Example of minimum change that leads to different outcomes		
[Q]	Text-based explanation [18, 175]	Explainability	• Post-hoc explanations	✓	✓	✓	✓	Agnostic	• With or without outcome comparison		

	Means	Value	Manifestation(s)	Stakeholder				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[R]	Interactive demonstrations [120]	Explainability	• Post-hoc explanations				✓	Agnostic			
[S]	Experiential AI [82]	Explainability	• Post-hoc explanations				✓	Agnostic	• Art mediated between computer code and human comprehension		
[T]	Interactive contestations [84, 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								
[U]	Challenge justifications provided by operator using the same means [84]	Contestability	• Mechanisms for users to ask questions and record disagreements with system behavior	✓	✓	✓		Agnostic	• Further testing • Verification		

	Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[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 [33] [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								
[Y]	AI Fairness 360 [34] [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								

[33] https://github.com/dssg/aequitas
[34] https://github.com/Trusted-AI/AIF360

Means	Value	Manifestation(s)	Stakeholder DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[Z] Fairlearn [35] [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 [36] [90]	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 [37]	Security	• Defence against integrity threats • Defence against privacy threats	✓				Agnostic			
[AC] InterpretML [38] [39] [134, 137]	Explainability	• Interpretability by design • Post-hoc explanations	✓	✓			Both white-box and blackbox models		• Bar charts • Line charts • Decision trees	

[35] https://github.com/fairlearn/fairlearn
[36] https://github.com/microsoft/HAXPlaybook
[37] https://github.com/Azure/counterfit
[38] https://github.com/interpretml/interpret/
[39] https://github.com/interpretml/DiCE

	Means	Value	Manifestation(s)	Stakeholder				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[AD]	Error analysis dashboard [40]	Non-discrimination	• Analysis of data for potential biases, data quality assessment	✓	✓			Agnostic		• Decision tree • Error heatmap	
		Explainability	• Post-hoc explanations								
		Fairness	• Accuracy across groups								
[AE]	Breakend Impact tracker [41] [83]	Performance	• 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 [174]	Contestability	• Mechanisms for users to ask questions and record disagreement with system behaviour				✓	Agnostic			

Table 15. 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 14. 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.

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

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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 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, 1 (2018), 95–122. 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, 2 (7 2020), 2053951720949566. https://doi.org/10.1177/2053951720949566

[5] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 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. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA. 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, 1 (2021), 15. 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. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. 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. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 289–295. https://doi.org/10.1145/3306618.3314243

[11] 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

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

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

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

[15] 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).

[16] 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/

[17] Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2021. Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 116–128. https://doi.org/10.1145/3442188.3445875

[18] 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

[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, 4/5 (2019), 1–4. 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?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 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.* Technical Report MSR-TR-2020-32. Microsoft. 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. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 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, 3 (9 2017), 273–291. 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. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A Friedler and Christo Wilson (Eds.). PMLR, 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. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA. 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, 4 (2019), 478–506. https://doi.org/10.1080/1463922X.2019.1697390

[34] Kyla Chasalow and Karen Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 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. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

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

[37] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. (10 2016).

[38] 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. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi-org.tudelft.idm.oclc.org/10.1145/3313831.3376660

[39] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The Politics of Training Sets for Machine Learning.

[40] 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. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 81–95. https://doi.org/10.1109/SP.2008.11

[41] Steven Dalton, Iuri Frosio, and Michael Garland. 2019. Accelerating Reinforcement Learning through GPU Atari Emulation. (7 2019).

[42] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 525–534. https://doi.org/10.1145/3351095.3372878

[43] Dasha Simons. 2019. *Design for fairness in AI: Cooking a fair AI Dish.* Technical Report. Delft University of Technology. Graduation project. MSc in Strategic Product Design. http://resolver.tudelft.nl/uuid:5a116c17-ce0a-4236-b283-da6b8545628c

[44] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Stroudsburg, PA, USA. https://doi.org/10.18653/v1/W19-3504

[45] Janet Davis and Lisa P. Nathan. 2015. Value Sensitive Design: Applications, Adaptations, and Critiques. In *Handbook of Ethics, Values, and Technological Design*. Springer Netherlands, Dordrecht, 11–40. https://doi.org/10.1007/978-94-007-6970-0_3

[46] 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

[47] 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).

[48] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729

[49] 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

[50] Ravit Dotan and Smitha Milli. 2019. Value-laden Disciplinary Shifts in Machine Learning. (12 2019).

[51] 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, 1 (10 2021), 48–59. https://ojs.aaai.org/index.php/HCOMP/article/view/18939

[52] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. (4 2011).

[53] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. 265–284. https://doi.org/10.1007/11681878_14

[54] Julia Earp and Jessica Staddon. 2016. "I had no idea this was a thing". In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*. ACM, New York, NY, USA, 79–86. https://doi.org/10.1145/3046055.3046062

[55] 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, 2 (2020), 197–213. https://doi.org/10.1007/s41060-019-00181-5

[56] 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

[57] European Commission. 2019. Ethics guidelines for trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[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. 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, 2 (6 2019). 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, 4 (12 2018). 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] W. Fred van Raaij and Theo M.M. Verhallen. 1994. Domain-specific Market Segmentation. *European Journal of Marketing* 28, 10 (10 1994), 49–66. https://doi.org/10.1108/03090569410075786

[65] Batya Friedman, David G. Hendry, and Alan Borning. 2017. A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction* 11, 2 (2017), 63–125. https://doi.org/10.1561/1100000015

[66] Georges Gaillard. 2016. La conflictualité : une modalité de lien où s'arrime la destructivité humaine. *Connexions* 106, 2 (2016), 71. https://doi.org/10.3917/cnx.106.0071

[67] Yanjie Gao, Yu Liu, Hongyu Zhang, Zhengxian Li, Yonghao Zhu, Haoxiang Lin, and Mao Yang. 2020. Estimating GPU memory consumption of deep learning models. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, New York, NY, USA, 1342–1352. https://doi.org/10.1145/3368089.3417050

[68] 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

[69] Timnit Gebru, Google Jamie Morgenstern, Briana Vecchione, and Jennifer Wortman Vaughan. 2020. Datasheets for Datasets.

[70] 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?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 325–336. https://doi.org/10.1145/3351095.3372862

[71] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, and Klaus Mueller. 2020. Measuring Social Biases of Crowd Workers using Counterfactual Queries. (4 2020).

[72] 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 (11 2021), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[73] Amir Globerson and Sam Roweis. 2006. Nightmare at test time. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, New York, New York, USA, 353–360. https://doi.org/10.1145/1143844.1143889

[74] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. (12 2014).

[75] Google. 2018. AI at Google: Our Principles. https://www.blog.google/technology/ai/ai-principles/

[76] Ben Green and Lily Hu. 2018. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*. Stockholm, Sweden.

[77] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32.

[78] Christopher Groves. 2015. Logic of Choice or Logic of Care? Uncertainty, Technological Mediation and Responsible Innovation. *NanoEthics* 9, 3 (12 2015), 321–333. https://doi.org/10.1007/s11569-015-0238-x

[79] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[80] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 392–402. https://doi.org/10.1145/3351095.3372831

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

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

[83] 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).

[84] 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

[85] 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. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 169–174. https://doi.org/10.1145/3386392.3399557

[86] César 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

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

[88] 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).

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

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

[91] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi-org.tudelft.idm.oclc.org/10.1145/3290605.3300637

[92] 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. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

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

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

[95] Stefania Ionescu, Anikó Hannák, and Kenneth Joseph. 2021. An Agent-Based Model to Evaluate Interventions on Online Dating Platforms to Decrease Racial Homogamy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 412–423. https://doi.org/10.1145/3442188.3445904

[96] Technology Japanese Cabinet Office, Council for Science and Innovation. 2019. Social Principles of Human-Centric Artificial Intelligence. https://www8.cao.go.jp/cstp/english/humancentricai.pdf

[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

[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, 6 (7 2021). 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. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3411764.3445096

[129] 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

[130] 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

[131] Tanushree Mitra. 2021. Provocation: Contestability in Large-Scale Interactive {NLP} Systems. In *Proceedings of the First Workshop on Bridging Human{–}Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, 96–100.

[132] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[133] 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

[134] 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

[135] Pradeep K Murukannaiah and Munindar P Singh. 2014. Xipho: Extending Tropos to Engineer Context-Aware Personal Agents. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 309–316.

[136] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine Learning with Membership Privacy using Adversarial Regularization. (7 2018).

[137] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. (9 2019).

[138] 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).

[139] OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0406

[140] 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

[141] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 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, 1 (2020), 66–89. https://doi.org/10.1080/10714421.2019.1704111

[143] Reema Patel. 2021. Reboot AI with human values. *Nature* 598, 7879 (10 2021). 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, 4 (12 2012), 285–303. 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[148] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A F Almeida, and Wagner Meira. 2020. Auditing Radicalization Pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 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. In *2017 Conference for E-Democracy and Open Government (CeDEM)*. IEEE, 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, 4 (12 2015). 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.. In *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. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 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, 1 (12 2012). https://doi.org/10.9707/2307-0919.1116

[157] 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).

[158] 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

[159] 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

[160] 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, 4 (10 2020). https://doi.org/10.1145/3419764

[161] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership Inference Attacks against Machine Learning Models. (10 2016).

[162] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA. https://doi.org/10.1145/3351095.3372870

[163] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[164] 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. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445088

[165] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483305

[166] 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

[167] 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.

[168] The Royal Society. 2019. Explainable AI: the basics . https://royalsociety-org.tudelft.idm.oclc.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

[169] Sarah Thew and Alistair Sutcliffe. 2018. Value-based requirements engineering: method and experience. *Requirements Engineering* 23, 4 (11 2018). 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. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19)*. Association for Computing Machinery, New York, NY, USA, 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] National Science United States Executive Office of the President and 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

[173] 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, 2 (7 2019), 205395171987946. https://doi.org/10.1177/2053951719879468

[174] 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, CSCW2 (10 2020), 1–22. https://doi.org/10.1145/3415238

[175] 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. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445365

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

[177] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, New York, NY, USA, 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] Zezhong Wang, Jacob Ritchie, Jingtao Zhou, Fanny Chevalier, and Benjamin Bach. 2021. Data Comics for Reporting Controlled User Studies in Human-Computer Interaction. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2 2021), 967–977. https://doi.org/10.1109/TVCG.2020.3030433

[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. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[182] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109, 1 (1980), 121–136. 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. In *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, 01 (4 2020), 1079–1087. 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. In *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. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 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. In *2014 Ninth International Conference on Availability, Reliability and Security*. IEEE, 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, 6 (2019), 34–42. https://doi.org/10.1109/MIS.2019.2958393