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				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[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				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[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	AT	DE	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				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
		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 DT	AT	DE	DS	Application (model)	Approach	Visual elements	Additional details
[M]	Example-based explanations [18, 24, 32, 49, 98, 117, 118]	Transparency	• Disclosure of properties of data	✓	✓	✓	✓	Agnostic	• Similar example • Typical example • Counterfactual 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				Application (model)	Approach	Visual elements	Additional details
				DT	AT	DE	DS				
[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