AI and Ethics
https://doi.org/10.1007/s43681-024-00504-8

ORIGINAL RESEARCH

Decisional value scores: A new family of metrics for ethical AI-ML

Gabriella Waters[1] · **William Mapp**[1] · **Phillip Honenberger**[1]

Received: 14 October 2023 / Accepted: 10 June 2024

Abstract
Research in ethical AI has made strides in quantitative expression of ethical values such as fairness, transparency, and privacy. Here we contribute to this effort by proposing a new family of metrics called "decisional value scores" (DVS). DVSs are scores assigned to a system based on whether the decisions it makes meet or fail to meet a particular standard (either individually, in total, or as a ratio or average over decisions made). Advantages of DVS include greater discrimination capacity between types of ethically relevant decisions and facilitation of ethical comparisons between decisions and decision-making systems, including across different modalities (for instance: human, machine, or coupled human–machine systems). After clarifying ambiguities in the concept of "decision" itself, including the question of how to individuate the decisions made by a system, we discuss the role and meaning of "decision" in common AI and machine learning approaches such as decision trees, neural networks, SVMs, and unsupervised classifiers. We then show how DVSs may be defined for several ethical values of interest, with an extended discussion of transparency. Finally, we explore how such metrics can be applied to real decision-making systems through two case studies: evaluations of LLMs for transparency; and evaluations of criminal risk assessment tools for utility, rights violations, fairness, and transparency.

Keywords Ethics · Artificial intelligence · Machine learning · Decisions · Transparency · Fairness

1 Introduction

Research in ethical AI has made strides in quantitative expression of ethical values such as fairness, transparency, and privacy [1–3].[1] Here we contribute to this effort by proposing a new family or strategy of metric, which we call "decisional value scores" (DVS). Most basically, DVSs are scores assigned to a system based on whether its decisions meet or fail to meet a given ethical standard, either individually, in total, or as a ratio or average over decisions made. For instance, a DVS for *transparency* might be "ratio of transparent decisions to total decisions made," and, for responsibility, "ratio of responsible decisions to total decisions made."[2]

DVSs come in four basic flavors that we believe may be useful to evaluators: an "individual decisional value score" (IDVS) that gives the score for a single decision; a "total decisional value score" (TDVS) that is the sum of scores for all decisions; a "ratio decisional value score" that is the ratio

✉ Phillip Honenberger
philliphonenberger@gmail.com

Gabriella Waters
gabriella.waters@morgan.edu

William Mapp
william.mapp@morgan.edu

[1] Center for Equitable AI and Machine Learning Systems (CEAMLS), Morgan State University, Baltimore, MD 21251, USA

[1] Quantitative ethics metrics exhibit several well-known benefits and limitations. The potential benefits include (a) reduction of ambiguity in ethical discussions, (b) clear articulation of goals and standards by which systems may be evaluated (either as discrete benchmarks to be met, or ideals to be approximated to), and (c) support in development of ethical evaluation tools that can be integrated with quantitatively expressed ML models and systems. At the same time, the limitations of such metrics include (d) the perhaps intractable *contestability* (in the sense of [44]) of ethical concepts like those the metrics seek to measure, and (e) the inherently approximative or partial status of many if not most of these metrics (i.e. their status as mere "proxies" for the richer and harder-to-measure values of interest). It should also be remembered that (f) insufficient awareness of limitations (c)-(e) can easily lead to additional ethical problems (for instance, [45]); and (g) the ethical value of any particular metric can legitimately derive in part from considerations other than accuracy or completeness as a measure of the value in question, such as *breadth of applicability*, *ease of use*, *ease of measurability* (that is, of collection of data necessary to run calculations), and *public or stakeholder acceptability*.

[2] Given that the metric was first proposed by Gabriella Waters in application to transparency and responsibility, we are tempted to call it the "Waters AI Transparency/Responsibility Score," or WATRS for short. For the purposes of discussion here we use DVS.

Published online: 24 July 2024

of acceptable (or unacceptable) decisions to total decisions (RDVS); and an "average decisional value score" (ADVS) that gives the average decisional value across all decisions. Assignments of ethical value to individual decisions may be made *discretely* (classification of the decision as meeting or failing to meet the standard expressed in the value) or *continuously* (as a variable score assigned to each decision). RDVS relies on discrete ethical value assignments. IDVS, ADVS, and TDVS may rely on discrete or continuous assignments. (See Table 1 for a summary.) Intermediate or combined metrics between these are also possible, but for simplicity we focus on IDVS, TDVS, RDVS, and ADVS here.

Why care about decisional value scores in the context of AI/ML deployment? In particular, why care about them as an option within the currently available menu of metrics for ethical values in AI/ML?

A first advantage of DVSs is that they place the phenomenon of "decision" at the front and center of ethical evaluation of a system, and this phenomenon is arguably of paramount importance to ethical deliberation and evaluation. As will be detailed further below, decisions are selections among possible alternatives at a choice point. As such, they are the points in a system's operation that both "make a difference" to its operation and impact and are subject to revision or correction. They are thus the natural point of focus for ethical evaluation insofar as such evaluation can guide intervention and improvement.

Yet, decisions are of many different types and appear at many different points in a system's operation. The DVS framework adds value in a second way by encouraging and supporting increased clarity and precision in these dimensions.

Third, DVSs provide a common framework within which multiple ethical concerns and values – including, for instance, fairness, transparency, benefits and harms, and respect for rights – can be more informatively related and integrated.

Fourthly and relatedly, DVSs are capable of providing more nuanced and informative *comparisons* between systems of diverse types, including AI systems with very different architectures, human systems, and human-AI hybrid ("Centaur") systems.

To clarify and make tractable the basic idea of a DVS, however, we must first address a few questions, including: How do we propose to define "decision" in the DVS framework? In particular, how can we individuate and count decisions made? What is meant by a "decision" in AI and machine learning contexts in particular? We address these questions in Sects. 2–3. Also, what measures might be appropriate metrics for ethical values and standards such as transparency or responsibility in application to decisions? We address this question in Sect. 4. Finally, how might a DVS framework be applied to real cases of decision-making systems, to evaluate them for their ethicality by the chosen standards? We address this question illustratively through two case studies in Sect. 5.

2 What is a decision?

The paradigmatic case of a "decision-making system" is, of course, an individual human being, but human organizations or institutions, non-human organisms, and automated tools and frameworks such as AI models and computer programs are also often described as "making decisions." However, it is rare to find explicit discussion of what a decision in general *is*,[3] and there is no standard method for individuating and counting the "number of decisions made" by a system. Without such a method, the quantification strategy we're pursuing would be impossible. So we must at least propose a tractable method for getting precise about how decisions can be individuated and counted.[4]

The idea that decisions involve selection of a specific action in circumstances where more than one action is possible, is frequently a component of definitions of decision. Eilon writes that "the definition of decision activity … is associated with making a choice between alternative courses of action" [4] (cf. also [5]). Simon [6] writes "At any moment there are a multitude of alternative (physically) possible actions, any one of which a given individual may undertake; by some process these numerous alternatives are narrowed down to that one which is in fact acted out. The words 'choice' and 'decision' will be used interchangeably in this study to refer to this process" [6, p. 4].[5]

However, the word "decision" as commonly used is ambiguous between two senses. In the first sense, it means "*a situation in which* the agent can only follow one of two or more available paths." In the second sense it means "*the path that an agent follows* in a situation where it can only follow one of two or more paths," i.e. the selection made. Compare "You have an important decision to make" with "She made a good decision." Let us call decisions in the first sense "choice points" and in the second sense simply

[3] This is a common gripe in the decision-theory literature: for instance, [4].

[4] We accept that the answer to this question may be partly stipulative (i.e. is a free choice-point in model building), but there are ways to handle the resulting ambiguities in inter-model comparisons, as discussed further below.

[5] Simon goes on to specify: "Since these terms as ordinarily used carry connotations of self-conscious, deliberate, rational selection, it should be emphasized that as used here they include any process of selection, regardless of whether the above elements are present to any degree" (4).

Table 1 Four "flavors" of decisional value score

Name of metric	Equation	Informal description	Discrete or continuous value assignment	Application contexts
Individual decisional value score (IDVS)	$D_i a_i \begin{cases} D \rightarrow R \in [0..\infty] \\ a \rightarrow R \in [0..1] \end{cases}$	The extent to which a single decision meets an ethical standard (A catalog of major ethical standards and possible metrics for each is given in Sect. 4.)	Discrete or continuous	Ethical evaluation of a single decision; comparison among available decisions at a single choice point; or bottom-up, decision-by-decision construction of TDVS, RDVS, or ADVS for a larger system
Total decisional value score (TDVS)	$\sum_{i=1}^{N} D_i a_i$	The *total number* of decisions made by a system that meet or fail to meet an ethical standard TDVS may be used in at least three ways: **Negative TDVS**: total ethical costs of decisions made by a system; **Positive TDVS:** total ethical benefits of decisions made by a system; **Combined TDVS:** total ethical benefits of decisions made by a system minus the total ethical costs	Discrete or continuous	Comparison between systems that includes not just average ethicality, but also *overall* ethical impact (positive or negative)
Ratio decisional value score (RDVS)	$\frac{1}{N} \sum_{i=1}^{N} D_i a_i$ $a_i \rightarrow [0\|1]$	The ratio of decisions made by a system that meet or fail to meet an ethical standard Note that **positive RDVS** and **negative RDVS** are trivially inter-definable	Discrete	Comparison between systems (or between possible variants of a single system) that focuses on average ethicality of each system or variant
Average decisional value score (ADVS)	$\frac{1}{N} \sum_{i=1}^{N} D_i a_i$ $a_i \rightarrow R \in [0..1]$	The extent to which decisions made by a system meet or fail to meet an ethical standard, on average ADVS can be interpreted as an RDVS that allows a (and, if useful, D and **D**) to take continuous rather than discrete values	Discrete or continuous	Comparison between systems (or between possible variants of a single system) that focuses on average ethicality of each system or variant

D = decision made by a system; a = ethical feature of interest; N = number of decisions made. How D are individuated and N are counted is discussed in Sects. 2 and 3 below. D is normally set equal to 1; however, if magnitude of decision is treated as a variable, D may be set greater or lesser than 1. We assume $D = 1$ here. For some applications a is set between 0 and 1 to maintain normalization; for others it may range across positive and negative values (as in Case Study 2 below)

"decisions" or "choices made."[6] When *individuating* "decisions," we are really interested in identifying the choice points wherein some action was taken, and (sometimes) in *characterizing* those choice points (for instance, in terms of the number and type of possible choices available at them). But when *evaluating* an AI or ML system's decisions, we are primarily interested in evaluating the *specific choices* – that is, not the *choice points* but the *choices made* at those points. These "choices made" are what we call "decisions." (Features of the choice point at which the choice was made are, of course, often relevant to the evaluation of the choice made at that point. We call a specification of these features a *characterization* of the choice point or of the decision-situation.) In what follows, we will primarily reserve the word "decision" for *choices made*, but will revert to more precise terminology when necessary for clarity.

A system's choice points can often be individuated in more than one way. For instance: Are a neural network's choice points defined by the activation at each of its nodes, or also by the weights on connections between nodes? Or, are they defined simply by each output? It seems likely that no one decision-individuation strategy will suffice for every perspective or procedure that we might like to use in answering such questions. We anticipate that decisions will need to be individuated or counted in different ways in different applications. For this reason we sometimes refer to *decision models*, by which we mean representations of the decision-making capacity of the system (including both individuation and characterization of its choice points). Note that systems can often be described as producing *their own* decision models in a sense, on the way to "making" their decisions (for instance: decision-tree or random forest algorithms that parse the dataset via binary decisions the algorithm itself selects); but observers of the system from outside (such as human theorists or evaluators) can also produce decision models that attempt to describe or characterize the decisions of the system, and these can overlap or fail to overlap with the system's own decision model.[7] Comparisons between systems, regarding their DVSs, are unlikely to be meaningful except in cases where at least the systems under comparison are approached through a common set of individuating/counting standard for both.

We may define a system's *aggregate decisional behavior* (ADB) as a set of choices made at its choice points. A system's ADB can be defined for a given time window (e.g. from 12:00 pm-1:00 pm, Thursday, Aug. 17, 2023 Eastern Standard Time) or for an average time window (a system's average ADB per hour) or per operation (a system's ADB when running a specific operation, or average ADB when running operations of a specific type), or across all possible operations.

When individuating and counting a system's decisions (i.e. when making a decision model of the system), it is important to be clear about what parts of the system's operation will be included. At least five "stages" of decision-making with and by AI-ML systems may be delineated [7]. A characterization of these stages, and the types of ethical analyses that are typically conducted in regard to that stage, are given in Table 2.

One might try to formally define and rigorously measure the number of "decisions" of a system in a variety of ways, such as Shannon information or *bits* [8],[8] or Pearl's "do" operator in causal analysis of a system's behavior [9]. Table 3 provides a sketch of a possible information-theoretic measure of the number of decisions for outputs of different kinds.

For machine learning models, one might even posit an identity between the amount of discrete (Shannon) information contained in the model – for, instance, the *bits* of information contained in its training data, as estimated by such features as the number of parameters in that data – and the total number of "decisions" made by the model. (The system's decisions, so defined, may be identical to what has elsewhere been called the "self-information" of the system [10].) But we expect that needs for individuation will sometimes deviate from such formal schemes; hence, we allow a variety of decision-models to be generated for any system.

3 Decisions in AI and machine learning[9]

Decisions in AI and machine learning systems might be individuated in a variety of ways. In Table 4, we provide a summary of how the decision-making behavior of some familiar types of AI models might be characterized. This characterization includes two major components: (1) a description of how *choice-points* may be *individuated* and *characterized* for the AI models; and (2) a description of how the AI-ML

[6] We might further distinguish the "choice made" from actions taken on the basis of that choice (a distinction often relevant in ethical contexts, as emphasized by [46]). Specification of what is meant by the "choice points" in a description of a decision-making system resolves this ambiguity and makes further distinction along these lines unnecessary.

[7] For reflection on the relation between system-produced and observer-of-system-produced models of system behavior, see Luhmann [47] and the tradition of second-order cybernetics more generally.

[8] Bateson intuitively articulates Shannon information as "differences that make a difference" [48, 49].

[9] The features of AI/ML systems discussed in this section are more-or-less common knowledge among AI researchers, as expressed in textbooks such as Russell & Norvig's [50].

Table 2 Major stages of decision-making in and with AI-ML systems

Stages (in time)	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Stage label	Model and data selection	Model construction and/or training	Internal processing	Output	Output application
Decisions at that stage	Selecting the model type and datasets	Building and/or training the model	Decisions made "within" the model when given an input or task	Decisions presented as "output" of the model when given an input or task	Decisions made in application of (and/or in conjunction with) the model's outputs
Decisions made by	Humans	Humans & AI-ML	AI-ML	AI-ML	Humans & AI-ML
Ethical analyses often recommended*	For model selection: Stakeholder feedback For dataset selection: Fairness, Accuracy	Fairness, Bias in datasets	Transparency	Transparency, Fairness, Costs/Benefits, Accuracy	Stakeholder feedback, Accountability, Accuracy, Fairness

*The "ethical analyses often recommended" are based solely on our impressions of current AI-ML ethics literature. We claim no objectivity for these lists as representations of that literature; nor do we claim these are the analyses that *ought* to be performed; nor make any claim about how often they currently *are* performed

Table 3 A classification of common AI and ML system outputs ("decisions") in terms of information content in the "decision" represented by the output

Type of decision	Measure of information content
Binary classification	$1\,bit = \frac{P}{2}\{P = 2$
N > 2 Classification	$\frac{P*D}{2}\begin{cases} N > 2 \\ P = N \text{ classes} \\ D = \text{Diversity of classes} \end{cases}$
Continuous Classification (i.e. Regression)	$\frac{P*D}{2}\begin{cases} N > 2 \\ P = N \text{ possible output values} \\ D = \text{Diversity of possible output values} \end{cases}$
Qualitative ("generative AI") response (for instance: production of novel sentences or images)	$\frac{P*D}{2}\begin{cases} N > 2 \\ P = N \text{ possible output features} \\ D = \text{Diversity of possible output features} \end{cases}$

P Number of possible choices; *D Diversity* of possible choices

model uses inputs and its internal features to *make the decisions* that it does.

In decision trees, each node in the tree represents a choice point, and the paths from that point diverge based on presence or absence of a single feature value [11]. The formation of the choice point, and the decision made at that point, are derived from the training data by determining which splits (decisions) most effectively segregate the classes. In decision tree algorithms, choice-point formation and decision-making occur in tandem, with both based on the attribute that provides the highest information gain. "Information gain" in this context expresses how much uncertainty is reduced with each decision. In rule-based systems (which may be extracted from decision trees), choice points can be explicitly stated as "if–then" rules.

In neural networks, on the other hand, decisions are made through a series of weighted inputs and activation functions. In a trained network's response to input of a single record, the activation or non-activation of each node can be treated as a single decision. (Alternatively, the weights of nodes inside the network might also be treated as intermediate decisions, made in the process of the network's training, though technically these "decisions" are made at the earlier "model construction & training" stage.) The final decision, especially in classification tasks, often involves a normalized exponential (also known as a softmax) function to derive a probability distribution over possible classes. Here the classification, as expressed in the probability distribution, can itself be treated as the system's final decision.

For models that segment the input space into different regions (e.g., support vector machines and unsupervised classifiers), the placement and orientation of these boundaries signify decisions, as well as the resulting classification of records on either side of the boundary or within each cluster.

Table 4 Decision individuation and selection for common AI-ML model-types at internal processing stage (Stage 3) and output stage (Stage 4)

	How are internal choice points individuated?	How does the system make decisions at internal choice points?	How are output decisions individuated?	How does the system make decisions about outputs?
Stage (from Table 2)	Stage 3: Internal processing	Stage 3: Internal processing	Stage 4: Output	Stage 4: Output
Decision Trees	The model recursively splits the dataset based on feature values that result in the best separation of the target variable. The tree is built from the top-down. Choice points are formed based on calculated value of a specific feature as a "decider." Each is formed based on a particular feature and a threshold	Each sample is classified at each choice point by which of the available branches best matches its features	Each leaf node of the decision tree represents an individuated output decision, determined by the path taken through the tree	The output decision is the classification or value at the leaf node, determined by the cumulative decisions made along the path from the root to that leaf
Neural Networks (when supervised classifiers)	Option #1: Each node is a choice point. For each activation event (i.e. each activation at the input layer leading to an output at the output layer), the extent to which a node is activated represents the choice made at that node Option #2: Each potential firing of a node, each activation function, and each weighted connection is a choice point. For each activation event, the choice points are (a) the extent of firing at each node, (b) the activation function at that node, and (c) the weight of the connections between any two nodes Option #3: …	Input data is transformed through layers, each comprising multiple nodes. Each node "makes a decision" about its output based on its input, activation function, and weights of outgoing connections. The final decision is the aggregate outcome of these myriad micro-decisions	The output layer of the neural network represents the final output decision	The model selects the output values or class labels corresponding to the activations in the output layer, which are the result of the collective decisions made throughout the network
Rule-based Systems	Each rule in a rule-based system is a choice point. When a rule is triggered, a choice is made	Decisions are made based on predefined rules. When input data matches a rule's conditions, the corresponding action (i.e. decision) is taken	Any behavior of the system that is presented or interpreted as an "output" is an "output decision"	Output decisions are produced by the combination of triggered rules
Reinforcement Learning	Points at which an action is taken. The agent decides which action to take based on its current state and learned policy Some RL systems also construct and consult *simulations* of action, and make calculations and then decisions on their basis; in these systems there are additional choice points regarding *construction*, *calculation*, and *what parts* of the simulation inform selection of actions	Decisions (actions) are chosen to maximize expected future rewards. The decision-making policy is updated based on feedback (reward) from the environment	With these systems, internal choice points and output choice points are usually identical. (The systems learn *as* they behave; apart from simulations, they have no "internal" decision-making behavior distinct from their observable decision-making behavior.)	Same as for internal processing: Decisions (actions) are chosen to maximize expected future rewards. The decision-making policy is updated based on feedback (reward) from the environment

Table 4 (continued)

	How are internal choice points individuated?	How does the system make decisions at internal choice points?	How are output decisions individuated?	How does the system make decisions about outputs?
Support Vector Machines	There are two levels of choice points in SVMs' internal processing: (1) formation of hyperplanes (or parts of the hyperplane) separating classes; (2) classification of input points on either side of the hyperplane (Unsupervised learning models like k-means can be described by a similar two-level decision process.)	(1) SVMs select the hyperplane that has the maximum margin between two classes (2) SVMs classify input data points based on their relation to this boundary	Output decisions similarly come in two parts: (1) a constructed hyperplane; and (2) a classification of each data point If helpful for analysis, each classification of a datapoint can be treated as a distinct output decision; and portions of the hyperplane could be distinguished as individual decisions (though the hyperplane could of course be partitioned in more than one way, suggesting that decision models of this type are subject to wide-ranging interpretive flexibility)	(1) SVMs select the hyperplane that has the maximum margin between two classes. (2) SVMs classify input data points based on their relation to this boundary
Bayesian Networks	Choice points are individuated during the structuring of the network, where nodes and edges are defined based on dependencies among variables. The structure is determined based on the probabilistic relationships, and parameters are learned from data. Each node in a Bayesian network then represents a probabilistic decision based on the states of its parent nodes	Decisions are made by computing conditional probabilities across the network of interconnected variables	The output decision is the final probabilistic inference made by the network	Decisions are computed using algorithms like variable elimination or message passing, which infer the probabilities of outcomes based on the evidence

In reinforcement learning, a decision is basically equivalent to an action. Actions are chosen based on a policy that is estimated to maximize expected rewards. This policy can be derived from prior exploration of the environment and learning from received rewards. The policy function maps states to actions, representing decisions based on state evaluations.

Decision individuation is likely to be more challenging for more complex or opaque models. A deep learning model with millions of parameters, for instance, is hard to represent or understand in its full interconnected complexity; further, it may be parsed within a decision model in many different ways. The typical decision-tree or rule-based system, on the other hand, is both easier to represent and understand, and leaves fewer degrees of freedom in individuation of its decisions. Providing a reasonable and useful decision model of a complex system may require use of techniques that shed light on the system's inner workings, such as attention mechanisms, feature importance metrics, or saliency maps [12–14]. However, the results of these techniques are often more accurately understood as interpretable reconstructions or "stories" about a system's decision-making process than as accurate descriptions of the actually operative mechanisms in that process.

In general, the greater complexity of a model correlates to a greater number and variety of types of decisions. For instance, deep neural networks can be treated as making a vast number of "micro-decisions" through their many layers and nodes [15, 16]. Simpler models like linear regression make "macro-decisions" based on coefficients of features. Yet the significance of each internal decision of a complex model – either for overall system behavior or for outputs and impacts – may ultimately be much less than a single internal decision of a less complex model. Quantifying decisions in a way that allows intercomparison between varyingly complex models might require aggregative measures or sampling methods with the more complex models, in order to obtain normalized estimates of "how many" choice points one system exhibits in comparison with others [17]. A simple form of normalization along these lines is to focus decision individuation only at the level of *outputs*. For most evaluations of most ethical concerns (*utility* or *fairness*, for instance), output-only decision individuation is likely sufficient. For evaluations of transparency, however, the number and extent of transparency of each internal decision can be especially salient.

4 DVS definitions for seven ethical values of interest

If our work to this point has succeeded, we've established a general framework for comparison of AI-ML models regarding features of their decisions, with some options given for individuating decisions at a few major stages of AI-ML use cycles. It remains now to fill this abstract description of DVS frameworks with content by defining DVSs for ethical values of interest. However, it should be borne in mind that the precise specification or measure of ethical values is often controversial. DVS is consistent with a wide range of choices here, and users of a DVS approach are free to develop their own, alternative operationalization of the ethical values of concern, and/or to develop or use other operationalizations of other values.[10]

4.1 (a) Substantive versus procedural ethical values

There are at least two types of ethical consideration about any decision-making system. One type of consideration is whether the decisions made by the system satisfy *substantive standards* of correctness or preferability; the other is whether the decisions made by a system satisfy *procedural standards* of correctness or preferability.[11] Substantively unethical decisions are wrong because of the content of the decisions themselves, regardless of how they were settled upon. Procedurally unethical decisions are wrong even if the content of the decision is right, because *the procedures by which the decision was arrived at* are ethically problematic or unacceptable. Examples of substantive ethical standards include *classical utility* (benefits and freedom from harm), *respect for rights*, and *attention to justice* (in which *fairness* is usually included). These constitute positive desiderata by which a decision in a decision-making process can be evaluated, regardless of the procedures by which that specific decision was decided upon. The epistemic value of *accuracy*, likewise, is a substantive standard in this sense. The values of *transparency*, *explainability*, and *democracy*, on the other hand, are *procedural* standards. They have to do with the ethicality of the procedure by which decisions are

[10] Ethical terms such as "transparency," "responsibility," and "fairness" are both *ambiguous* and *contestable*. They are *ambiguous* in the sense that they have multiple possible meanings. Responsibility, for instance, may be a synonym for "ethicality" in general, or may mean having an obligation to respond. The obligation to respond that is meant, in turn, may be legal obligation or moral obligation (for more on such distinctions regarding "responsibility," see [51]). And they are *contestable* in the sense that they are characterized by long-standing and perhaps intractable disagreements about what should be taken to be the full or central meaning of the word or concept [44]. Contestability doesn't imply worthlessness of discussion and debate: continued debate about the meaning of contestable concepts such as democracy, truth, art, science, and ethical values like justice, transparency, and responsibility, can reveal things about the world and our values, enable negotiation of commitments by a community, and provide guidance for behavior of individuals or organizations, that is virtually impossible to acquire in any other way.

[11] This distinction is sometimes recognized in the literature (for instance: [19, 20, 52–54]), but seems to be undertheorized.

selected rather than with the ethicality or ethical preferability of the content of the decisions themselves (on which they are neutral). The epistemic value of *rationality* is, plausibly, a procedural standard analogous to these others.

The relation between substantive and procedural ethical metrics is complex and a detailed treatment is beyond the scope of this paper. But two relationships are particularly salient. First, a set of decisions can meet all procedural ethical requirements, including decision-making by democratically acceptable decisions and accommodation of stakeholder feedback, and still be substantively unethical. This could occur, for instance, in circumstances in which a democratic majority favors some decision that is substantively wrong (devastation of the natural environment, for instance, or scapegoating of a few individuals to benefit the many). Conversely, a set of decisions may be substantively ethical but chosen by means that violate procedural ethical standards, such as those of transparency, legal responsibility, or responsiveness to stakeholder feedback. In these cases the system has chosen the right course of action, but the *way* it made the decision was itself unethical. Procedurally ethical decision-making systems may sometimes serve as a bulwark against substantively unethical decisions, but they won't always do so; hence the need for continued ethical argument and advocacy even within the frame of a procedurally just order (e.g. democracy).

In the next two sections, we offer some definitions of DVS metrics for ethical values in these two categories. This definition is conducted at three interrelated levels, with the lower levels determined by features that can be measured more-or-less directly and uncontroversially, and higher levels derived from information at the lower levels. This strategy makes use of the added clarity that ethics metrics can provide to ethical deliberations, while allowing for alternative low-level metrics or mid- and high-level combinations of metrics. While we hope the reader finds our articulation of these metrics plausible, we stress, again, that the specific definitions of the metrics in these tables are merely exemplary: our main purpose in this paper is to explore the capacities of DVSs in general.

4.2 (b) DVSs for substantive ethical values: an overview

Substantive ethical values discussed in the contemporary ethics literature typically include *classical utility* (total benefits minus total harms), *individual rights* (including Kantian approaches), *virtues* (including Aristotelian and perhaps also care ethics approaches), and *justice*. The influential "principlism" approach to medical ethics, for instance, showcases the proposed "four principles" of benefits, harms, autonomy (a type of right), and justice [18]. Sometimes additional values, such as *environmental values*, *aesthetic values*, and *preservation of cultural artifacts*, are added to this set. A set of low-level DVS metrics for these values might be defined as in Table 5.

These four desiderata of ethical decisions involve well-known tensions and tradeoffs with one another, as expressed in thought experiments such as the Trolley Problem [22]. Despite these tensions, promising avenues for linking and integrating the substantive ethics metrics within a single high-level metric have been proposed, including within quantitative frameworks applicable to AI and ML systems. One strategy is to define a highest-level substantive ethics metric (an "overall substantive ethics score," or *OSES*) through a formula such as the following (compare [23]):

$$OSES = F\omega_f\left(R\left(U\omega_u + V\omega_v\right)\right)$$

where U = utility, V = aspirational values, R = a rights filter, F = a fairness function, and $\omega_{f,}$ $\omega_{u,}$ and ω_{v} are weights that define the relative importance of fairness, utility, and aspirational values, respectively. This formula states that the relative substantive ethicality of a decision (in competition with other possible decisions) is a function of (a) the total expected benefits minus the total expected harms (that is, classical utility, or, in modern terms, the recommendation of a cost–benefit analysis), plus (b) the total expected aspirational benefits – with (a) and (b) weighted according to one's judgment of the relative importance of each – and (c) whether the decision passes *filters* for "respect for rights" [23, 24] multiplied by (d) a fairness function (detailed further below). In short, the formula recommends maximizing preference satisfaction, aspirational ideals, and fair and equitable distribution of benefits and harms, while requiring that basic standards of respect for rights and a minimum threshold of fair treatment are met.

Techniques for working with formulae such as these, in application to specific decision situations, are already well-developed. They form the core of cost–benefit analysis and risk analysis, which are widely employed in, for instance, environmental impact assessments (EA), environmental economics, city planning, and public policy research more generally [25, 26]. Good metrics for low-level ethical values, in these contexts, depend greatly on the specifics of the situation modeled and the modelers' purposes in generating a model. As step-wise functions, rights play a cut-off role in some legal and policy decisions. Standards for fairness and aspirational values are often controversial and contestable, but arguments on the basis of such values form an important and influential part of public discourse, and these values have been given some degree of legal codification and enforceability (for instance: the Civil Rights Act of 1964, which raises some fairness standards to the level of legal rights; and animal

Table 5 Substantive ethics metrics

Name of metric (mid-level)	Informal definition	Criteria/Evaluation procedure (low-level)	Commonly used and/or plausible proxies/indicators	Decisional value score per decision
Utility (U)	What is the **total expected benefit** of the action minus the **total expected cost**, in regard to all affected individuals' preference-satisfaction?	What is the expected **benefit** of the action for each individual affected? $B_i + B_{i+1} + \ldots B_n = [0, \infty]$ What is the expected **cost** of the action for each individual affected? $C_i + C_{i+1} + \ldots C_n = [0, \infty]$	Stated preferences (survey responses); market behavior; biological correlates of pleasure and pain; estimates of likelihoods (particularly in risk analysis)	$U = \sum_{i=1}^{n}(B_i - C_i)$
Rights (R)	Does the action meet minimum conditions of **ethically acceptable treatment of individuals** (such as respect for bodily autonomy)?	Does the action exhibit respect for **human rights?** $R = [\varnothing, 1]$ R can be divided into multiple submetrics $R_a, R_b, R_c, \ldots$ for each fundamental right that is recognized (e.g. autonomy, privacy, property, etc.)	Relevant legal codes; rates of compliance with relevant legal codes; evidence of informed consent (e.g. a signed contract)	$R = R_a * R_b * R_c \ldots$ Typically R_x take discrete values such that if $R_x = \varnothing$, the action is classified as unethical due to rights considerations; if $R_x = 1$, the action is not unethical due to rights considerations (but may be unethical for other reasons)
Fairness (F)	Does the action (or, more usually, set of actions) treat all parties **fairly** (for instance, not treating members of some groups worse or better than others without sufficient justification)? (What AI ethicists discuss as "fairness" is basically equivalent to "justice" in the "distributive justice" sense [19, 20].)	Are **false positive rates** equal between groups? $FPP = 1 - \lvert FPR(g') \rvert$ Are **false negative rates** equal between groups? $FNP = 1 - \lvert FNR(g) - FNR(g') \rvert$ Is **predictive accuracy** equal between groups? $PAP = 1 - \lvert PA(g) - PA(g') \rvert$ *Where = 1 indicates equal and* < 1 indicated extent of deviation *fromequality* *FPR=false positive rate* *FNR=false negative rate* *PA=predictive accuracy* g = *group 1* g' = *group 2*	Statistical analysis of performance of decision-making systems (for instance, [2])	$F = \frac{\frac{1}{2}(FPP+FNP)+PAP}{2}$ for all groups definable for the individuals affected by the action Evaluation is usually restricted to comparison between a subset of groups of interest. Consideration of greater number of possible groupings ("intersectionality") generally correlates to greater estimates of system unfairness. [1, 21]
Aspirational (V)	How well does the action align with or facilitate our (or some community's) highest **ideals and values**, including aesthetic ideals?	How much does the action facilitate manifestation of **aspirational ideals** not covered by previous metrics? $V = [0, \infty]$ V can be divided into multiple submetrics $V_a, V_b, V_c, \ldots$ for each aspirational ideal that is recognized (e.g. beauty, diversity, etc.)	Various; Sometimes similar techniques as for *utility* (U) above; Examples include United Nation's Human Development Index (HDI), ecosystem services, and Sen & Nussbaum's capabilities approach	$V = \sum_{i=1}^{n} V_i * w_i$ As with rights (R), the set of $V_a, V_b, V_c, \ldots$ depends on what the analyst adopts as aspirational values. In addition, i is a weight value assigned by the theorist, so that different aspirational values may be assigned appropriately different weights

welfare protection laws and endangered species protection laws regarding aspirational environmental values).

Rights filters can be of two basic kinds: input filters and output filters [23, 24]. The former specify that certain kinds of preference satisfaction or dissatisfaction should be ignored (for instance, the enjoyment of an aggressor at the pain of a victim). The latter specify that certain possible "solutions" to the choice problem (i.e. certain possible choice paths) ought to be striken from consideration – for instance, those that violate individual rights. Options to be stricken from consideration can be rated ϕ. Note that these filters can be made more or less flexible in several useful ways. One might, for instance, set a threshold of benefit maximization (or, more likely, of harm avoidance) that allows choices that violate some subset of individual rights (rights to property, for instance), to "pass through" the filters and be considered as selectable options, so long as the benefits accrued or harms avoided are large enough to meet the threshold (as might, for instance, Ross [27]).[12]

How fairness considerations should be related to other substantive ethical values is controversial. One approach would be to set a threshold of acceptable performance deviation between groups. Differences over that threshold would work like rights violations and set the value of the combined formula to ∅; differences under that threshold would work like negative utility, bringing the overall value of the formula slightly lower than it would otherwise be. The point at which such a threshold value were set would be more or less stipulative but could be decided upon by, say, public or expert deliberation, or context-specific arguments. As a general rule, the "four-fifths" standard for disparate impact, as often employed in American legal contexts, could suffice. Following this rule in application to our formulae above, the fairness threshold would be 0.8 (i.e. any difference greater than 20% would trigger a y = ∅ collapse; any difference less than 20% would be subtracted from overall utility, adjusted by the selected weighting of the fairness value [ω_f]).

4.3 (c) DVSs for procedural ethical values: the case of *transparency*

We now turn to a detailed consideration of an area where DVS scores can add value to existing ethics metrics: namely, the evaluation of a system's *transparency*.

Though distinctions between "transparency," "interpretability" and "explainability" are appropriate on the basis of natural language semantics, machine learning researchers often either (a) do not distinguish these terms, or (b) distinguish them in ways that are incongruent between researchers. Despite this unclarity, a number of distinct types of transparency have been identified, and a variety of methods or techniques have been developed for enhancing the transparency of a system such as SHAP and LIME [12–14]. Less progress has been made on creating quantitative *metrics* for transparency, particularly metrics that might be valid for comparing systems of different models and/or application contexts.[13] This is a dimension to which DVS scoring can contribute.

Doshi-Velez and Kim [28] usefully distinguish three distinct criteria by which the transparency of a model might be evaluated. *Application-grounded* evaluation tests the extent to which the system in its current or some modified form can assist in accomplishment of a specific kind of task or application. Variations in the transparency of the model (for instance, in what kinds of explanations of the system are available) can be studied in regard to whether they improve "success" in uses of the model for that kind of task or application. *Human-grounded* evaluation focuses on human subject preferences for one kind of explanation over another, that is, subjective estimates of explanatoriness.[14] Finally, *functionally-grounded* evaluations rely on some quantitative metric, usually considered as a proxy for explanatoriness in one of the first two senses.

Lipton [29] notes at least three features by which the transparency of a model can be estimated, which he calls *simulatability* (could the model be reconstructed by a human?), *decomposability* (can the relevant parts of the model such as "inputs," "parameters," and "calculation" (p. 14) be delineated, and can the role each plays in the model's decisions be understood?), and *algorithmic transparency* (can the learning algorithm itself be understood?). Lipton notes that some models fare better on these three standards than others (for instance, a few-layered neural network in comparison with a many-layered one), but warns against classifying models as more or less successful on any of these metrics due solely to their model-type alone. (We should not assume, for instance, that every linear model is necessarily more transparent than any neural network).

Despite this warning, many theorists allow that model types can be roughly categorized into (1) those that tend to perform relatively well on simulatability, decomposability, and algorithmic transparency, simply because of their

[12] These and other capacities of such filters are explored in [23].

[13] [3] is an exception, though one without evident application to systems of many types (i.e. not "model-agnostic"). [3] appeared after we had completed and submitted this paper.

[14] Doshi-Velez and Kim [28] define this kind of metric slightly differently, emphasizing small-scaleness of use contexts and layperson test base as defining features of "human-grounded" approaches in contrast to "application-grounded" ones. But these differences seem to us differences of degree rather than kind; the most significant difference is in whether *successful performance of the task*, or *subjective estimate of explanatoriness or interpretability*, is the standard by which the system's transparency is judged.

typical model architecture, and (2) those that tend to perform relatively worse on these, and thus, to be made explainable, require application of *posthoc* methods to improve their transparency [12]. The former are sometimes called *white-box* models or just "transparent models" and usually include linear regression, decision trees, k-nearest neighbors, rule-based learners, and Bayesian models (p. 5). The latter are called *black-box* or *opaque* models and include random forest, support vector machines, and multi-layered neural networks. In working with black-box models, post hoc methods become the central focus of efforts to improve transparency. These methods are sometimes distinguished into *model-specific* and *model-agnostic* methods (depending on whether they are tailored to a particular model type such as neural networks, or rather are applicable to any model type). Model-agnostic approaches usually involve constructing a limited or partial model, or other representation, of the system's behavior around a particular kind of decision. Some well-known model-agnostic approaches include LIME and SHAP [13, 14].

Perhaps the simplest application of DVS, in conjunction with current commonly adopted frameworks, is to estimate the transparency of each output decision by whether or not a LIME- or SHAP-type explanation has been provided (or is readily accessible) for that decision. This might be formalized as the rule that, for each output decision, if such an explanation is available or readily providable, set TRAN = 1; if not, set TRAN = 0. (These scores are IDVSs for each output decision.) The overall transparency of a system's decisions can then be estimated by the RDVS (ratio of transparent to total decisions) or Negative TDVS (total number of non-transparent decisions) of the system. While this procedure is exciting as a means for readily quantifying extent and degrees of transparency, it does have several drawbacks, mostly due to limitations of the LIME and SHAP procedures themselves, such as incomplete capture of intuitive standards of "transparency," "explainability," and/or "interpretability" by the existence or availability of LIME and SHAP interpretations of a system's decisions. For instance, most applications of SHAP assume independence of the features that SHAP estimates the relative significance of, an assumption that is often highly questionable. And the local explanations produced by LIME are plausible but potentially highly misleading if taken as correspondent to a model's internal behavior, and don't necessarily shed light on more global features of the model's structure and operation.

Another, higher-caliber standard of transparency that might be used would be first estimate the total number of internal decisions, and then estimate the proportion of those for which explanations of a certain threshold of informativeness and accuracy are available. (The standard of informativeness could be calibrated to typical SHAP or LIME explanations.) This could lead to a much lower transparency score for opaque and/or complex models such as deep neural networks or closed transformer models if training data, model architecture, or weights are inaccessible to the public and no public or publicly-available posthoc explainability methods comparable to SHAP or LIME are provided.

A possible specification of low-level metrics for "transparency" and related procedural values is provided in Table 6.[15] These include (a) the extent to which features of the model such as data, model architecture, and training process can be observed by outsiders at all (what we've called "accessibility"), (b) the extent to which humans are provided with explanations of the model's decisions that are both understandable and faithful (that is, accurate to the decision-making system's actual decision-making procedures),[16] and (c) the extent to which humans can themselves explore the model by interacting with it. The related values of *responsiveness to stakeholder feedback* and *accountability for errors and harms* are also included because these are necessary conditions of the ethical salience of transparency itself: transparency without these further features does not contribute much to the overall ethicality of a decision-making system.

For instance, *accessibility* measures how observable the system and its operations are to the general public. The overall accessibility (α) of a decision can be estimated by:

$$A = Ad * Am * At$$

$$Ad \rightarrow [0..1]$$

$$Am \rightarrow [0..1]$$

$$At \rightarrow [0..1]$$

where Ad is a value between 0 and 1 indicating the extent to which the data used for training the model is accessible; $Ad = 0$ if training data is inaccessible, closed, or proprietary; and $Ad = 1$ if it is publicly accessible. Likewise, Am is a value between 0 and 1 indicating the extent to which the model is publicly accessible, and At is a value between 0 and 1 indicating the extent to which the training procedures are publicly accessible.[17]

Transparency itself might be defined as in Table 7.

[15] Compare [3], which only appeared after the present article was drafted.

[16] We thank Kofi Nyarko for suggesting "faithfulness" as a value of importance here.

[17] Alternatively, each low-level value might be expressed continuously, as a value between 0 and 1, rather than discretely as 0 or 1. The mid-level metric might be defined as the multiplicative product of the values at the lower-level (as in this table) or as the average of added values of the lower-level metrics (so that 0 scores for lowest level metrics don't force the mid-level metrics to the value of 0).

Table 6 Low- and medium-level ethics metrics for procedural values related to *transparency*

Name of metric (mid-level)	Informal definition	Criteria/Evaluation procedure (low-level)	Commonly used and/or plausible proxies	Decisional value score per decision (average of added values version)
Accessibility (A)	How **observable** are the system's decisions to the general public?	Is the **data** accessible? $Ad \rightarrow [0..1]$ Is the **model** publicly accessible? $Am \rightarrow [0..1]$ Are the **training procedures** publicly accessible? $At \rightarrow [0..1]$	N/A (relatively easy to assess)	$A = Ad * Am * At$
Explainability (E)	How **understandable** are the system's decisions to the public, based on information the system itself provides?	Does the system provide **explicit explanations** for decisions? $Ee \rightarrow [0..1]$ Are the explanations **understandable** to humans? $Eu \rightarrow [0..1]$ Does the system provide **faithful explanations** for decisions? $Ef \rightarrow [0..1]$	Presence or absence of explanations by LIME or SH Performance of human subjects' interpretations of the system (measured on *accuracy*, *timing*, *confidence*, and *subjective sense of interpretability*) [30]	$E = Ee * Eu * Ef$
Explorability (β)	How available is the system for public **exploration** and **experimentation**?	Can the model be **interacted with** and experimented with? $Pp \rightarrow [0..1]$ What **range of scenarios** can be explored? $Pr \rightarrow [0..1]$	N/A (relatively easy to assess)	$P = Pp * Pr$
Stakeholder Feedback (S)	How much does **stakeholder feedback** inform revisions and later versions of the system?	Can stakeholders **provide feedback** about the model and its functioning? $Sf \rightarrow [0..1]$ How **diverse and representative** is the set of stakeholders from which feedback is taken? $Sd \rightarrow [0..1]$ Are there mechanisms for **revising and improving** the model based on feedback? $Sr \rightarrow [0..1]$	Number and frequency of meetings with stakeholders Breadth and representativeness of stakeholder representation Openness of participation to relevant stakeholders	$S = Sp * Sd * Sr$

Table 6 (continued)

Name of metric (mid-level)	Informal definition	Criteria/Evaluation procedure (low-level)	Commonly used and/or plausible proxies	Decisional value score per decision (average of added values version)
Accountability (Y)	To what extent can specific individuals or organizations be held **accountable** for failures (including enforcement mechanisms and restitution)?	Is there an **identifiable party** that will be held accountable if something goes wrong? $Yp \rightarrow [0..1]$ Are there effective mechanisms for **enforcing accountability** (including compensation to victims if something goes wrong?) $Y_e \rightarrow [0..1]$ Are the **compensatory mechanisms** adequate to the compensate for the harms done by errors of the system? $Ya \rightarrow [0..1]$	Legal codes; Statistical analyses of outcomes of prior lawsuits and other accountability-seeking processes	$Y = Yp * Ye * Ya$

5 Case studies

How can DVSs be applied to real decision-making systems? What kinds of novel insights can be expected from their application? In this section we explore these questions through two case studies. Different components of the previously described framework were drawn upon in each case study. The first focuses on DVSs for transparency in large language models (LLMs). This application exhibits the value of DVSs for comparative estimates of model transparency by submetrics. The second application focuses on DVSs for a selection of substantive and procedural ethical concerns, as applied to criminal risk assessment tools such as COMPAS [31, 32]. The second case study shows DVSs' capabilities for comparative evaluation across system types (human, AI, or hybrid) and decision-types (in this case: recidivism prediction, release-or-detention decisions at initial intake ["bail"], and release-or-detention decisions at parole hearings).

5.1 (a) Case study #1: transparency in LLMs

Despite the widely voiced desirability of transparency in AI-ML systems, there are few metrics by which such transparency could be evaluated. The closest thing to a comprehensive quantitative evaluation framework for transparency is the "Foundation Model Transparency Index" (hereafter FMTI) introduced in November 2023 [3]. This 100-parameter checklist enables the "grading" of LLMs on features intuitively relevant to transparency such as the accessibility and explorability of the models [3]. However, the grading system of the FMTI, at least as carried out in the study introducing the instrument [3], relies solely on statements from system producers about which of these attributes apply to their systems. Further, FMTI is designed for transparency evaluation of foundation models, with no discussion of how the transparency of such models might be compared with models of different types. In sum, while FMTI is well-suited to estimating dimensions of transparency in LLMs, DVSs have a broader applicability than FMTI. This is true for the variety of ethical values they can track, the variety of decision-making systems they are designed for evaluating, and the variety of means of estimating sub-metric values that they allow. DVSs are also more specific about what *decision* components of a system's operation they are applied to: individual decisions, aggregates of decisions, or averages of decisions.

To demonstrate the capabilities of DVS on this issue, we selected three LLMs for evaluation: (a) Anthropic AI's Claude 3, (b) Cohere4AI's Command-R, and (c) Meta

Table 7 Definitions of high-level metrics for procedural ethical values

Name of metric (high-level)	Informal definition	Formula for average decisional value score (ADVS)
Transparency (TRAN)	How well can the decisions of a system be **understood by stakeholders and the public**?	$TRAN = A\omega_a * E\omega_e * P\omega_p$ Weighted product of accessibility, explainability, and explorability for all decisions, divided by number of decisions
Democracy (DEM)	How well do the decisions of a system respect basic principles of **democracy**?	$DEM = A\omega_a * E\omega_e * P\omega_p * S\omega_s$ Weighted product of accessibility, explainability, explorability, and stakeholder feedback capacity for all decisions, divided by number of decisions
Legal Responsibility (L-RESP)	To what extent are the decisions of a system ones that some person or organization can be assigned **legal responsibility** for?	Legal Responsibility: L-RESP ~ γ Legal Responsibility is roughly equivalent to accountability (though mechanisms of accountability may sometimes be slightly wider than legal mechanisms – for instance, market mechanisms such as consumer behavior)

The labels "ω_x," in the third column stand for weights representing the relative importance of each metric. These can be set according to user needs and preferences, but should be kept constant through comparative analyses

LlaMa 2. All three are in wide use and widely studied by public and private researchers. We evaluated each model for transparency and stakeholder feedback in comparison with the others using the continuous scoring method (values between 0 and 1). The results are summarized in Table 8.

We now discuss how numerical values were assigned to each system.

Claude 3 is a closed-source, closed-model, closed-data, and closed-weighted model released by Anthropic AI, a company dedicated to research and commercialization of LLMs. Claude 3 is a component of Amazon Web Services' Bedrock LLMs. It is accessible via a RESTful API hosted by Anthropic, Amazon Web Services, or Google Cloud Services. Anthropic doesn't provide model weights, training procedures, or documentation about the model's structure or internals; and doesn't supply a model card in the traditional sense that details the number of parameters, hyper parameters, and other values. (Anthropic's "model card" [33] is rather a marketing document that gives a benchmark comparison against products from OpenAI, Google, and others.) One step of the training process, the Constitutional AI training approach, uses reinforcement learning as a self-improvement mechanism to remove harm. This component is described in a publicly available research paper [34], which slightly increases scores for accessibility and explainability. But this step constitutes perhaps 20% or less of its total training procedure; thus, the increase is very slight. All in all, we rated *Ad*, *Am*, and *At* at 0.1.

Anthropic indicates that Claude 3 is trained on freely available web content and internally generated content. But the company doesn't specify the online web sources nor the scope and tenor of internally used sources. Therefore, no direct explanations of how content sources are used in its training materials are available, and no LIME or SHAP-type explanations are provided automatically. We set $Ep = 0.15$, $Eu = 0.5$, and $Ef = 0.5$.

Other than querying through the API, the model is relatively non-explorable. API-querying alone may be conceived as exploration of just two stages of a model's performance: the input stage and output stage, including how these are correlated. Given that an LLM's internal processing between these two stages is almost certain to constitute 80% or more of its total set of decisions, these limitations leave the explorability of the model rather low. We thus set *Pp* and *Pr* at 0.2.

Command-R is a mixed-source, closed-model, closed-data, and open-weighted model released by Cohere AI, a company dedicated to building enterprise and business-language-focused LLMs. Command-R is designed to accept and execute instructions through a conversational interface. Command-R is specifically tuned for business applications and doesn't perform well generally. Cohere makes Command-R accessible via a RESTful API hosted by Cohere and HuggingFace. The system has a thorough model card outlining parameters, hyperparameters, and training modules accessible via HuggingFace [35]. The card reports on training methodology for Command-R at some level of detail ($At = 0.5$), but its training corpus is unavailable for inspection ($Ad = 0.1$). Command-R model weights and biases can be downloaded in their entirety from HuggingFace as a logged-in, HuggingFace user ($Am = 0.9$).

The relative openness of Command-R's training procedures and model weights means that moderately informative and accurate explanations can be produced for many of its decisions ($Ep = 0.75$, $Ef = 0.75$). Yet explanations are not provided automatically, and producing such explanations may be a technical and interpretive challenge for users ($Eu = 0.5$).

Table 8 DVS scores for transparency and submetrics for three LLMs

	Accessibility of Data	*Accessibility of Model*	*Accessibility of Training procedures*	***Accessibility*** ***=[(Ad+Am+At) / 3]***	*Explanations provided*	*...that are understandable*	*...and are accurate ("faithful")*	***Explanatoriness*** ***=(Ep*Eu*Ef)***	*Model can be explored*	*...across a wide range of its states*	***Explorability*** ***=(Pp * Pr)***	***Transparency*** ***=[(A+E+P) / 3]***
Model	*Ad*	*Am*	*At*	*A*	*Ep*	*Eu*	*Ef*	*E*	*Pp*	*Pr*	*P*	TRANS
(a) Claude 3	0.1	0.1	0.1	**0.1**	0.15	0.5	0.5	**0.0375**	0.2	0.2	**0.04**	**0.06**
(b) Command-R	0.1	0.9	0.5	**0.5**	0.75	0.5	0.75	**0.28**	0.67	0.67	**0.45**	**0.41**
(c) LlaMa 2	0.75	0.9	0.75	**0.8**	0.85	0.6	0.85	**0.43**	0.9	0.9	**0.81**	**0.68**

Users can explore Command-R not just through input-level querying and output-level observation, but also by adjusting weights and biases in the model's processing stages. However, the full architecture of the model is not available for observation or manipulation. This sets the explorability of the model higher than Claude-3, but still below 1. We rated Pp and Pr at two-thirds (0.67).

LlaMa 2 is an mixed-source, open-model, closed-data, and open-weighted model released by Meta. LlaMa 2 is a research specific LLM with a commercial license and is freely available via GitHub. The model's training code can be forked directly from GitHub.com as well. The model weights and biases are available for download after registration with Meta, statement of intended purpose, and Meta's approval. We were able to obtain access to the model very easily: we received customized URLs to provide to download scripts for Llama 2, Llama 3, and Llama Code immediately upon initiating a request. LlaMa 2's model card is presented in an open-access research paper [36] downloadable from ai.meta.com. This paper details Llama 2's construction and training procedures, data sources, and architecture, and explains how Meta incorporates corpus data, annotates the data, and cites 3 major data sources used in the training process. More sources than these were used to train LlaMa 2, but the data reporting in this case is exceptionally transparent by comparison with the others considered ($Ad = 0.75$, $Am = 0.9$, $At = 075$).

The efforts taken to improve accessibility and understandability of model architecture, training procedures, and major data sources, certainly makes LlaMa 2 more explainable than it would be otherwise. However, explanations of particular decisions are not automatic but must be reconstructed through such procedures as investigation of model weights, exploration of the effects of setting weights differently or retraining, or post-hoc explainability methods like SHAP or LIME. Thus, the explainability of the model doesn't reach a maximum value; this is in part due to the complexity and opacity of *any* LLM architecture. Nonetheless, Llama 2's more specific data reporting than Command-R convinced us to rank it slightly higher than Command-R on the explainability sub-metrics (+0.1 for each sub-metric: $Ee = 0.85$, $Eu = 0.6$, $Ef = 0.85$).

LlaMa 2 is exceptionally explorable. The model code is publicly available on GitHub. Potential users can free download, fork, and use the code by installing Python and LlaMa 2 code requirements. Meta makes explorability especially feasible by supplying the requirements.txt file which can be used to automatically download Python dependencies required for training ($Pp = 0.9$, $Pr = 0.9$).

Overall summary: Perhaps unsurprisingly, the "openness" of a system was a strong predictor of its transparency score in the DVS evaluation. Our DVS rating of Llama as more

transparent than Anthropopic also matches the assessment of the FMTI [3].

5.2 (b) Case study #2: Criminal risk assessment systems

Criminal risk assessment (CRA) tools such as COMPAS [31] typically take as inputs some set of information about individual persons – for instance: persons who have been placed under arrest and are awaiting bail hearing; or who have been convicted of a crime, have been serving a prison sentence, and are now up for parole). The information may include prior conviction counts, age, sex, charge type, and answers to questionnaires. On the basis of this information and their internal processing, the assessment tool outputs a risk score (typically from 1 to 10, or in categories such as "low," "medium," and "high").

As Stevenson [37] notes, the increase in attraction and deployment of such tools may be understood as part of a longer-running trend towards "evidence-based" practices in the criminal justice sector, which have themselves been growing since at least the 1970s. However, the term "evidence-based" is ambiguous between at least two meanings: based on large data-sets, or empirically tested for efficacy. Arguably these systems have so far been evidence-based in the former but not in the latter sense. As Stevenson [37: 375] puts it,

> Risk assessment tools wear the clothes of an evidence-based practice – they are developed with the use of large data sets and sophistical techniques and endorsed by social scientists running policy simulations – but risk assessments should not be considered evidence-based until they have shown [*sic*] to be effective.[18]

Other studies of CRA systems echo this sentiment [38, 39]. Too little evidence of the needed kinds has been reported or collected on these systems, making a responsible analysis of their performance close to impossible: "[B]ased on the current published evidence, the highest priority [for researchers in this area] is … to work towards addressing the key methodological limitations identified in previous work" [39: 8].

Notwithstanding the limits of previous studies and (often) the absence or inaccessibility of the data needed for overcoming these limits, CRA tools are a domain in which DVSs can lend ethical insight, primarily through identifying, distinguishing, and providing quantitative assessment tools for ethical standards of different types (e.g. fairness, public utility, transparency, respect for human rights), in application to decisions of different types (particularly algorithmic, human, and hybrid, in regard to such matters as bail-setting, sentencing, and parole). This case study focuses on use of the DVS framework, in combination with results of previous evaluations of CRA tools, to estimate the overall ethicality of CRAs across a range of their current use cases.

The types of DVS that are most relevant to this case are *output IDVS* (an estimate of the ethicality of each output decision of the decision-making system), *output ADVS* (an estimate of the average ethicality of each decision made by the system, answering the questions 'Is it, on average, more *beneficial* or *harmful*, and by how much?' and 'Is it for any reason suspected to be ethically *unacceptable*?'); and *output TDVS* (that is, an estimate of the system's total ethical performance in practice, answering the question 'Has it, overall, done more good or harm, and how much?' and 'Is it, overall, ethically *acceptance* or *unacceptable* for making decisions of a particular type?'). To simplify the problem, we'll focus on estimating the output IDVS and output ADVS here.[19]

CRA tools rarely if ever make any impactful decisions on their own. Most directly, such a tool's output "decisions" are simply the risk scores that the tool assigns to individuals. If we identify selection of a particular risk threshold with a further decision (such as "release" or "do not release"), we can treat such tools as "making decisions" about things like bail and parole. But this construal is a simplification: in actual practice, decisions to release or not release an individual are almost never (if ever) made by these tools alone, but rather by human beings with or without the tools' risk scores in hand. There are thus at least four different types of decision-making system to be distinguished here:

(a) CRA tools deciding about risk.
(b) CRA tools deciding about release (e.g. bail or parole) according to the simplification described above.
(c) Humans without CRA tools deciding about release.
(d) Humans with CRA tools in hand deciding about release.

The ethical dimensions of import for evaluating the decisions of these systems include:

> *Rights*: Are individuals' rights respected through the decision-making process? In particular, are they given 'due process,' not subject to arbitrary treatment, and so on?

[18] And: "A practice should not be considered evidence-based because it references big data sets and sophisticated techniques – it should be considered evidence-based because its impacts have been carefully researched and understood" [37: 311].

[19] However, ADVS and TDVS can each in principle be estimated by the other: TDVS ≈ number of decisions, multiplied by ADVS; ADVS ≈ TDVS, divided by number of decisions.

Fairness: Are individuals treated comparably across demographic groups such as white/non-white, male/female, and abled/disabled? Are some groups disproportionately subject to errors or harms of the system?
Public Utility: Do such systems have overall good or overall bad effects from the standpoint of common measures of public utility (such as social cost measured in dollars, or preference satisfaction measured in surveys)? In particular, what are the average *benefits* of early release and what are the average *costs*? And what are the accruals to public utility of increased precision in predicting and avoiding repeat offenses? How should these be offset by any costs of the avoidance-generating procedures (such as longer detention or costs of applying the CRAs themselves)?
Transparency: To what extent are these systems accessible, understandable, and explainable?
Democracy: To what extent can the application of these systems (in general or in specifics) be contested, critiqued, or reshaped by the public?
Legal Responsibility: Are there individual or corporate entities who can be held responsible if the system makes an error?

As represented in the DVS metrics suggested in previous sections, transparency is itself a component (and thus in a sense a prerequisite) of democracy. It's also worth noting that fairness, transparency, and democracy become especially salient when their violation constitutes a rights violation. Some cases where fairness, transparency, or democracy are violated can count as rights violations, but not all. Yet failures of fairness, transparency, or democracy that don't constitute rights violations can still be ethically salient and warrant criticism.

Unfortunately, the existing literature on these systems only supports evidence-based evaluations for a small subset of these system-types and value-types. One contribution that the DVS framework can make to the evaluation of these systems is a delineation of ethically relevant considerations and the kinds of measurements that would be relevant to evaluating them, thus raising the bar for adequate evaluations going forward. For now, a combination of prior empirical work and reasoned estimation can be used to at least sketch an estimated ADVS for each. This sketch is given in Table 9.

A value of ∅ means that the decision is not just a negative in terms of net utility (which would normally, but not always, make it non-recommendable), but ethically unacceptable (because it violates rights or minimal standards of fairness).

Rights: We take it that mere estimation of a risk of recidivism doesn't violate anyone's rights. Likewise, we assume that parole decisions as such (that is, selection of a decision to release on parole or not-release) don't violate anyone's rights. However, decisions about *bail* as well as *CRA-informed decisions about bail or parole* raise some rights issues.

First, the entire practice of setting bail has come under criticism, particularly as it involves longer detention for individuals who don't have the ability to pay, which may violate their rights to due process. Likewise, paid bail allows early release for individuals who may be high risk but have ability to pay, thereby treating them differently than similarly high risk individuals without ability to pay, violating standards of fairness. Since bail reform is already a common trend, and reforms tend towards $0 bail with decision to release or not-release based solely on risk rather than ability to pay, here we simplify the ethics of bail decisions by construing them simply as the decision to release or not release an individual after initial intake. From this standpoint, bail decisions *as such* do not violate rights, but may do so in some circumstances (e.g. arbitrary and/or undue detention).

CRA-informed decisions about bail or parole may violate rights insofar as it is unclear whether and to what extent automated risk assessments can legitimately inform decision-making processes that affect individuals (for some arguments that they usually *cannot*, see [40]). If some of the data on which such systems are trained is illegitimate grounds for making such decisions (for instance: race, religion, or sexual orientation; or even more apparently benign considerations such as "how much criminal activity occurs in the vicinity of the individual's primary residence"), then use of these systems to inform decisions may violate individuals' rights on that account alone. However, an ethical problem doesn't necessarily arise for *all* CRA-assisted decision-making about bail or parole; this rather depends on precisely how the decisions are made (i.e. on the basis of what data points and what processing procedures). Analogous ethical problems can arise with human-only decision-making about bail or parole (for instance: use of race or class to inform decisions).

Fairness: The performance of CRAs on *fairness* is debatable. Some researchers are comfortable with *predictive accuracy parity* (PAP) alone as a measure of fairness (see [37] for supporting arguments), whereas others argue that metrics such as *false positive rate parity* and *false negative rate parity* should also be satisfied [31, 41]. We tried to take an intermediate approach in our evaluation by averaging between (TPP + FNP)/2, on the one hand, and PAP on the other.

Given the limited evidence in support of fairness of these systems by standards other than PAP [38], and some evidence of violation of these other standards [31, 32], we rate such systems a 0.75 for fairness, with a ± 0.25 window of potential variation, hopefully to be closed by future evaluations (or, closed in practice by more precise studies of particular CRA tools or systems). However, it is difficult on the basis of current data to say confidently whether

Table 9 A table of IDVS for the output decisions of criminal risk assessment tools and related decision-making systems

	(Decision-Maker) →Decision Type	Rights (R)	Fairness (F) If $F<0.8$, then $F=\varnothing$	Public Utility (U) If decision is accurate, value is + If decision is inaccurate, value is -	Overall Substantive Ethical Score (OSES) If F < 0.8 OR $R=\varnothing$, then OSES = ∅ Else OSES = $\pm U \pm F$	Transparency (T)	Democracy (D)	L-Responsibility (L)	Overall Procedural Ethical Score (OPES) = [(T + D + L) / 3]
(a)	(Human) → Risk	1	0.75 ± 0.25	± 1	If F < 0.8, then OSES = ∅ Else OSES = $\pm 1 \pm 0.5$	0.5	0.5	0.5	0.5
(b)	(CRA) → Risk	1	0.75 ± 0.25	± 1	If F < 0.8, then OSES = ∅ Else OSES = $\pm 1 \pm 0.5$	0.1	0.1	0.1	0.1
(c)	(Human + CRA) → Risk	1	0.75 ± 0.25	± 1	If F < 0.8, then OSES = ∅ Else OSES = $\pm 1 \pm 0.5$	0.2	0.2	0.2	0.2
(d)	(Human) → Bail	1	0.75 ± 0.25	± 5	If F < 0.8, then OSES = ∅ Else OSES = $\pm 5 \pm 2.5$	0.5	0.5	0.5	0.5
(e)	(CRA) → Bail	?{∅,1}	0.75 ± 0.25	± 5	If F < 0.8 OR $R=\varnothing$, then OSES = ∅ Else OSES = $\pm 5 \pm 2.5$	0.1	0.1	0.1	0.1
(f)	(Human + CRA) → Bail	?{∅,1}	0.75 ± 0.25	± 5	If F < 0.8 OR $R=\varnothing$, then OSES = ∅ Else OSES = $\pm 5 \pm 2.5$	0.2	0.2	0.2	0.2
(g)	(Human) → Parole	1	0.75 ± 0.25	± 10	If F < 0.8 OR $R=\varnothing$, then OSES = ∅ Else OSES = $\pm 10 \pm 5$	0.5	0.5	0.5	0.5
(h)	(CRA) → Parole	?{∅,1}	0.75 ± 0.25	± 10	If F < 0.8 OR $R=\varnothing$, then OSES = ∅ Else OSES = $\pm 10 \pm 5$	0.1	0.1	0.1	0.1
(i)	(Human + CRA) → Parole	?{∅,1}	0.75 ± 0.25	± 10	If F < 0.8 OR $R=\varnothing$, then OSES = ∅ Else OSES = $\pm 10 \pm 5$	0.2	0.2	0.2	0.2

non-CRA-based decision-making systems are more or less fair than CRA-based ones (for comments on this, see [37, 39]). We thus opted to rate *all* systems in our list (human, non-human, and hybrid) of equivalent performance for fairness, at least in this exploratory sketch.

Public Utility: Public utility is a measure of total *benefits* minus total *harms*, where "benefit" and "harm" have classically been understood in terms of pleasures and pains, but today are usually gauged by preferences satisfied, for which one uses survey instruments or market behavior to make estimates [26]. Public utility is more-or-less identical to the core quantitative part of what is sometimes called a "cost–benefit" analysis or a "risk analysis." Since rights and fairness are included elsewhere in our framework, these values shouldn't be treated as having any utility intrinsically, though their "knock-on" effects could be so treated.

The overall utility of *risk scoring alone* is relatively small, dependent as it is on the further use of such scores to guide behavior. Nonetheless it is not 0. Whether it is a net positive or negative remains to be determined. We set the magnitude of its value at ± 1 while remaining agnostic about directionality ($U = \pm 1$). The magnitude of utility for decisions about *bail* and *parole* are likely to be somewhat larger; we set these at $U = \pm 5$ and $U = \pm 10$, respectively. The main components of the utility of these decisions are probably (a) the total *cost* of early release (primarily via costs due to (i) increased rates of criminal activity and (ii) increased rates of rearrest), and (b) the total *benefit* of early release (primarily via benefits due to (i) increased preference-satisfaction for suspects, their families, and their communal partnerships and (ii) lowered costs on the criminal justice systems). (Conversely, one can estimate the same values in terms of overall costs and benefits of longer detention.)

Though the evidence is disappointingly scarce on this question, it does appear that CRA-assisted decision-making about bail and parole has modest effects in a few directions of relevance here. In particular, it appears that such systems lead to slightly higher rates of release overall; slightly greater precision deciding who to release or not release (thus generating slight decreases in the cost of further crime due to early release); and (in the case of bail decisions) slightly higher rates of "failures to appear," which should be counted a *cost* of the changed decision-making regime [37, 38]. Thus, overall, CRA-assisted systems show evidence of a slight increase in public utility by comparison with human-only systems. Somewhat surprisingly, it appears that the net benefit would be even higher if human decision-makers always followed the recommendations of the CRAs (i.e. rows (e) and (h) outperform rows (f) and (i), respectively). However, human decisions not to follow the CRAs appear in some cases to be effective parts of strategies aimed at not violating individuals' rights or other ethical desiderata [38] (though in others they appear to be simply reversions to older "habits" of decision-making practices [37]). Any suggestion that hybrid human-and-CRA systems should bring their decisions more into line with CRA recommendations should keep these dimensions in mind.

Transparency, democracy, and legal responsibility: The traditional, human-led systems by which bail and parole decisions were made prior to the introduction of CRAs were certainly imperfect as far as transparency, democracy, and legal responsibility for errors are concerned. Nonetheless, they undeniably had mechanisms for ensuring *some* degree of each of these ethical values – public records and the appeal process, for instance. Given that many CRAs are themselves relatively "black boxed" systems, trained on unshared data and employing unshared training procedures, the introduction of almost any current CRA into bail and parole decision-making processes will lower the overall transparency of those processes. Without transparency, democracy likewise is compromised. And, given the lack of official channels by which the scores of CRAs, or decisions made partly on their basis, could be challenged, the legal responsibility of CRA-assisted decision-making can in general be expected to be lower than human-only decision-making. Though precise numerical estimates are hard to justify without more extensive research on these questions, we set estimates for these three values to 0.5 for the human-only systems, to 0.1 for the CRA-only systems, and to 0.2 for the CRA-human hybrid systems.

It should be noted that these values are ones on which different CRA systems (for instance: COMPAS; the Arnold Foundation's Public Safety Assessment (PSA); or various checklist style risk assessment tools that preceded ML-trained CRAs) appear to vary significantly. Some relevant questions here include: To what extent do the developers of such systems make their training data, developed models, or training processes available to public scrutiny? To what extent do the systems provide explicit SHAP- or LIME-like explanations of individual decisions, and to what extent can we expect explanations so-provided to be accurate descriptions of these systems? To what extent are producers of the systems answerable to the public, either directly or through representatives? To what extent can be they be held legally responsible when something goes wrong?

Likewise, different hybrid human + CRA systems exhibit different levels of transparency, democratic accountability, and legal responsibility by the metrics described above. A comparative study of how such hybrid systems in different states or jurisdictions fare on these measures would be a welcome contribution to the testing and evaluation of such systems.

6 Conclusion

Decisions and decision-making are of central importance to the ethical performance of human and AI systems. Decisional Value Scores (DVSs) provide a promising new tool to study, compare, and improve the ethicality of such systems on a per-decision, aggregate-of-decisions, or average-decisions basis.

A key advantage of the DVS framework is that it places the fundamental unit of decision-making at the center of ethical evaluation. Prior approaches have tended to focus on high-level ethical principles without a clear way to connect these principles to the decisions made by an AI/ML system. In contrast, DVS's decision-centric framework starts by clarifying what will be counted as a decision in a particular evaluation, and focuses evaluations on comparisons across decisions of the same or different types.

The ability to evaluate decisions individually, rather than relying on aggregate system-level metrics, is another key strength of the DVS approach. It may eventually allow for much more granular ethical assessment and improvement of AI-ML systems insofar as designers could identify specific decisions or decisions of a certain type or stage that fail to meet ethical standards and work to improve them instead of just optimizing for an overall fairness or accuracy score. Even estimates of fairness, which have traditionally been made through statistical analyses (that is, via "statistical fairness" concepts [2]), could eventually be improved by a DVS approach: as methods for estimating fairness of individual decisions improve (for instance, counterfactual methods [42]), individual fairness scores for each system decision could be aggregated, giving a finer and more reliable measure of a system's fairness as a whole.

A third advantage of the DVS approach is that it supports tractable and clear comparative metrics for transparency. For example, researchers often aim to improve transparency through application of post-hoc interpretability techniques like LIME or SHAP. But these methods only provide local explanations for individual predictions, not a cross-model measure for transparency. The DVS framework allows for the transparency of each decision to be evaluated according to multiple relevant factors (e.g. accessibility of the underlying data and model, the clarity and faithfulness of explanation provided, and the ability of stakeholders to explore and interact with the system). The overall transparency of the system can be estimated by averaging across decisions.

Fourth, DVS provides a common language and set of metrics for comparing the ethical performance of AI/ML systems, even those with very different architectures. This allows for more meaningful benchmarking and accountability, as diverse systems can be evaluated against a standardized set of ethical criteria. By focusing on the fundamental unit of the decision, the DVS framework provides a way to assess and compare ethical performance across diverse AI/ML models.

Fifth and finally, the DVS framework has the potential to integrate ethical analyses across ethical principles and values, including both substantive and procedural ethical considerations. ("Decisions" are at least one meeting point between substantive and procedural ethicality!) This paper demonstrates how DVSs can be defined for specific values like transparency or responsibility, but the framework is flexible enough to accommodate an even wider range of ethical considerations. Integration is valuable because real-world AI systems often need to balance and trade off between competing ethical priorities. A DVS-based approach allows these tradeoffs to be made explicit and quantified instead of relying on more subjective or implicit assessments.

The granular, decision-centric nature of DVSs makes them well-suited to support iterative refinement and optimization of AI systems from an ethical standpoint. Designers can use DVSs to pinpoint specific areas for improvement, test the impact of design changes, and gradually work towards systems that better uphold the desired ethical standards. This iterative capability is necessary as ethical AI is an ongoing challenge that requires continuous evaluation and adjustment as systems are developed and deployed in the real world.

The DVS framework is designed to be adaptable and extensible. The specific metrics defined in this paper are meant to be illustrative rather than exhaustive. Researchers and practitioners can define new metrics to capture emerging ethical values or context-specific concerns. This flexibility is crucial as the ethical landscape for AI/ML systems is rapidly evolving. The DVS framework provides a stable foundation for ethical evaluation that can evolve alongside the technology and societal expectations.

Acknowledgements We thank Kofi Nyarko, Olusola Olabanjo, and two anonymous reviewers for feedback on earlier drafts of the paper.

Funding Funding for this research was provided by the Center for Equitable AI and Machine Learning Systems (CEAMLS) at Morgan State University.

Data availability Data used in Case Study #1 is available from references [33–36]. Data used in Case Study #2 is available from [31, 32, 37–39, 43]. The numerical values assigned to DVS scores in these case studies were informed by data in these sources but also by author interpretations of the significance of that data for the metrics of interest; we make no claim that scores were directly and objectively read from the data alone.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Kearns, D., Roth, A.: The Ethical Algorithm. Oxford University Press, Oxford UK (2020)
2. Carey, A., Wu, X.: The Statistical Fairness Field Guide. AI and Ethics **3**, 1–23 (2023)

3. Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., Liang, P.: The Foundation Model Transparency Index. ArXiv. **19**, 2023 (2023)
4. Eilon, S.: What is a Decision? Manage. Sci. **16**(4), B172–B189 (1969)
5. Pomerol, J.-C.: Decision-Making and Action. Wiley, New Jersey (2012)
6. Simon, H.A.: Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization, 3rd edn. Free Press, Mumbai (1976)
7. Danks, D. and A.Y. London: Algorithmic Bias in Autonomous Systems. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence AI and autonomy track. Pp. 4691–4697 (2017)
8. Shannon, C.: A Mathematical Theory of Communication. Bell Systems Journal, New York (1948)
9. Pearl, J.: Causality. Cambridge University Press, Cambridge UK (2009)
10. A Kumar 2023 Information Theory, Machine Learning, & Cross-Entropy Loss. https://vitalflux.com/information-theory-machine-learning-concepts-examples-applications/, accessed Sept 13 2023
11. J Fürnkranz 2011 Decision Tree In: C Sammut GI Webb (eds) Encyclopedia of Machine Learning Springer Boston https://doi.org/10.1007/978-0-387-30164-8_204
12. Belle, V., Papantonis, I.: Principles and Practice of Explainable Machine Learning. Frontiers in Big Data. **4**, 1 (2021)
13. M. Ribeiro, S. Singh, C. Guestrin 2016 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. KDD 2016 San Francisco
14. Lundberg, S., Lee, S.: A Unified Approach to Interpreting Model Predictions. Tamil Nadu, NIPS (2017)
15. Hu, X, L. Chu, J. Pei, W. Liu, & J. Bian: Model Complexity of Deep Learning: A Survey. ArXiv. https://arxiv.org/pdf/2103.05127.pdf (2021)
16. Zhang, Z., Cheng, H., Yang, T.: A recurrent neural network framework for flexible and adaptive decision making based on sequence learning. PLoS Comput. Biol. Comput. Biol. **16**(11), e1008342 (2020). https://doi.org/10.1371/journal.pcbi.1008342
17. P Pochelu 2022 Ensembles of Deep Neural Networks distributed and parallel for the energy industry. Computer Vision and Pattern Recognition [cs.CV]. Université de Lille. English. NNT: 2022ULILB009 . tel-03936792
18. Beauchamp, T.: Principles of Biomedical Ethics. Oxford University Press, Oxford (1979)
19. Rawls, J.: A Theory of Justice. Harvard University Press, Cambridge (1974)
20. D Miller 2021 Justice. Stanford Encyclopedia of Philosophy https://plato.stanford.edu/entries/justice/ Accessed May 29, 2024
21. A Bell L Bynum N Drushchak L Rosenblatt T Zakharchenko J Stoyanovic 2013 The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice FAccT '23 Chicago IL
22. Foot, P.: The Problem of Abortion and the Doctrine of Double Effect. Oxford Review **5**, 5–15 (1967)
23. Lowry, R., Peterson, M.: Cost-benefit analysis and non-utilitarian ethics. Politics, Philosophy & Economics **11**(3), 258–279 (2011)
24. RE Goodin 1995 Laundering preferences In: H Elster A Hylland (eds) Foundations of Social Choice Theory Cambridge University Press Cambridge UK
25. Harris, C., Pritchard, M., Rabins, M., James, R., Englehardt, E.: Engineering Ethics: Concepts and Cases, 6th edn. Cengage, Boston (2018)
26. Smith, S.: Environmental Economics: A Very Short Introduction. Oxford University Press, Oxford (2011)
27. Ross, D.: The Right and the Good. Oxford University Press, Oxford (1930)
28. F Doshi-Velez B Kim 2017 Towards a Rigorous Science of Interpretable Machine Learning. ArXiv accessed September 16, 2023
29. Lipton, Z.: The Mythos of Model Interpretability. ACM Queue, New York (2018)
30. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support. Syst.. Support. Syst. **51**(1), 141–154 (2011)
31. JJ Angwin J Larson S Mattu L Kirchner 2016 Machine Bias ProPublica
32. J Larson S Mattu L Kirchner J Angwin 2016 How We Analyzed the COMPAS Algorithm. ProPublica. May 23, 2016.
33. Anthropic: The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf Accessed June 2, 2024 (2023)
34. Bai, Y., Kadavath, S., Kundu, S., et al.: Constitutional AI: Harmlessness from AI Feedback. ArXiv (2022)
35. HuggingFace.co. Cohere4AI. C4AI-Command-R-V01 Model Card. https://huggingface.co/CohereForAI/c4ai-command-r-v01 Accessed June 2, 2024 (2024)
36. Touvron, H., Martin, L., Stone, K., et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv (2023)
37. Stevenson, M.: Assessing risk assessment in action. Minnesota Law Review **103**, 302–384 (2018)
38. Berk, R.: An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. J. Exp. Criminol. Criminol. **13**, 193–216 (2017)
39. Fazel, S., Burghart, M., Fanshawe, T., Gil, S.D., Monahan, J., Yu, R.: The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. J. Crim. Just. **81**, 101902 (2022)
40. Wang, A., Kapoor, S., Barocas, S., Narayanan, A.: Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. ACM Journal on Responsible Computing **1**(1), 1–45 (2023)
41. Hardt, M., E. Price, N. Srebo: Equality of Opportunity in Supervised Learning. ArXiv. (2016)
42. Carey, A.N., Wu, X.: The causal fairness field guide: perspectives from social and formal sciences. Front Big Data. **5**, 892837 (2022). https://doi.org/10.3389/fdata.2022.892837
43. Vaccaro, M., Waldo, J.: The Effects of Mixing Machine Learning & Human Judgment. ACM Queue July-August 2019, New York (2019)
44. WB Gallie 1995 Essentially contested concepts Proceedings of the Aristotelian Society
45. Nguyen, T.: Value Capture. Journal of Ethics and Social Philosophy. Forthcoming.
46. Beigang, F.: On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. Mind. Mach. **32**, 655–682 (2022)
47. Luhmann, N.: Introduction to Systems Theory. Polity Press, Cambridge (2012)
48. Bateson, G.: Form, Substance, and Difference. In: Steps to an Ecology of Mind. Chicago University Press, Chicago (1972)
49. Waters, C.K.: Causes that Make a Difference. Journal of Philosophy **104**(11), 551–579 (2007)
50. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Londaon, Pearson (2021)
51. Hart, H.L.A.: Punishment and Responsibility. Oxford University Press, Oxford (2008)
52. Shrader-Frechette, K.: Environmental Justice. Oxford University Press, Oxford (2002)
53. M Heikkilä 2023 It's High Time for More AI Transparency. MIT Technology Review. July 25, 2023. (https://www.technology

review.com/2023/07/25/1076698/its-high-time-for-more-ai-transparency/, accessed 8–29–2023) (2023)
54. Peters, U.: Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. AI and Ethics **3**, 963–974 (2023)