Proceedings of ICGPSH 2026 Symposium: EdTech & AI in Learning: Large Language Models in Business, Politics, and Humanities
DOI: 10.54254/2753-7048/2026.LD33969

The Normative Limits of Empirical AI Alignment: Preferences, Consensus, Data, and Normative Authority

Zekai You

School of Philosophy, Zhongnan University of Economics and Law, Wuhan, China
gearssocial@gmail.com

Abstract. Many influential approaches to AI value alignment rely, at least in part, on empirical inputs drawn from human experience, including individual preferences, collective judgments, and large-scale datasets. These inputs are indispensable for training, evaluating, and coordinating AI systems. However, they are sometimes treated, implicitly or explicitly, as if they could also provide the normative authority of alignment standards themselves. This article challenges that assumption through a meta-ethical analysis. It argues that preference-based, consensus-based, and data-based approaches all encounter the same structural limit: they can describe how people judge, choose, or behave, but they cannot by themselves explain why such judgments, choices, or behaviours should be treated as normatively authoritative. Taking universalizability and justifiability as two minimal conditions for normative grounds, the article examines the strongest versions of the three empirical approaches and shows that each requires an additional normative premise that cannot be generated from empirical inputs alone. The argument does not reject empirical alignment as a practical coordination mechanism. Rather, it repositions empirical alignment as normatively insufficient but epistemically and operationally useful. The normative authority of AI alignment standards must therefore be secured by independent ethical, legal, or institutional justification.

Keywords: AI ethics, value alignment, normative justification, meta-ethics, human preferences

1. Introduction

Artificial intelligence value alignment has become a central topic in debates on AI ethics, AI safety, and responsible AI. As AI systems are increasingly deployed in communication, recommendation, decision support, and institutional coordination, alignment can no longer be understood merely as technical performance or task completion. It also concerns whether AI systems can be made responsive to human values, social norms, and legitimate expectations. In this broad sense, alignment is commonly taken to require more than the optimisation of predefined objectives; it involves specifying what it would mean for an AI system to act in ways that are acceptable, justifiable, or appropriate in human social contexts [1].

Many influential approaches to AI value alignment rely, at least in part, on empirical inputs drawn from human experience. Human feedback is used to train models to produce outputs that

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.LD33969

evaluators prefer. Preference learning attempts to infer what users choose or would endorse under specified conditions. Participatory and deliberative approaches seek to include the judgments of diverse groups in order to make alignment standards more representative. Large-scale datasets, including textual corpora and other traces of human practice, are used to train systems that can reproduce, predict, or respond to patterns of human language, evaluation, and conduct [2-6]. These approaches differ in their technical methods and philosophical motivations, but they share a common orientation: they treat human experience as an indispensable source of information for aligning AI systems.

There is a strong practical reason for this empirical orientation. AI systems do not operate in a moral or social vacuum. If they are to interact with human beings in meaningful contexts, they must be informed by how people evaluate situations, express preferences, identify harms, negotiate disagreement, and respond to practical constraints. Empirical inputs can therefore play an important epistemic and operational role. They help reveal what people care about, where disagreements arise, which outputs are perceived as harmful or acceptable, and how models may be adjusted in response to human evaluation.

The difficulty begins when empirical inputs are treated not merely as resources for training, evaluation, or coordination, but as grounds of normative authority. By normative authority, this article means the capacity of an alignment standard to count as a justified reason for guiding AI design, deployment, or governance, rather than merely as a description of what people prefer, accept, or do. A model output may match human preferences, reflect a collective judgment, or reproduce patterns found in data. These are claims about descriptive fit. The stronger claim that such preferences, judgments, or patterns should guide AI behaviour is different. It concerns normative justification rather than empirical accuracy.

This distinction matters because empirical alignment can succeed operationally while remaining normatively incomplete. Human feedback may improve model behaviour, yet the feedback itself may reflect adaptive preferences, local conventions, or unjust expectations. A deliberative process may generate broad agreement, yet its scope of participation or supporting reasons may remain contestable. A dataset may capture large-scale patterns of language or behaviour, yet those patterns may encode prejudice, exclusion, or historically contingent norms. Even when empirical inputs are accurate, representative, and carefully processed, they still describe how people judge, choose, speak, or act. They do not, on their own, explain why those judgments, choices, linguistic patterns, or practices should have normative authority over AI systems.

This article argues that preference-based, consensus-based, and data-based approaches to AI alignment all encounter this structural limit. The problem is not merely that empirical inputs may be biased, incomplete, or noisy, although these are serious concerns. The deeper problem is that no increase in the quantity, representativeness, or technical precision of empirical inputs can by itself transform a description of human judgment or conduct into a justification of how AI systems ought to be governed. The issue, then, is not whether empirical inputs are useful for AI alignment. They clearly are. The issue is whether they can independently ground the normative authority of alignment standards. This article argues that they cannot do so without additional normative premises.

The argument should not be understood as a rejection of empirical alignment. Human feedback, preference modelling, participatory evaluation, and data-driven learning can improve the responsiveness, usability, and social sensitivity of AI systems. They can also reveal conflicts among values, expose patterns of harm, and make AI governance more attentive to the plurality of human expectations. The claim is narrower: empirical inputs are necessary for many alignment practices,

DOI: 10.54254/2753-7048/2026.LD33969

but they are not normatively self-sufficient. If an alignment standard is to guide AI behaviour in ways that affect multiple agents, communities, or institutions, it must be defensible not only as a reflection of human experience, but also as a standard that can be justified to those subject to it.

This concern is connected to a broader meta-ethical difficulty in AI ethics. In their discussion of benchmarking AI ethics, LaCroix and Luccioni argue that attempts to measure the "ethicality" of AI systems depend on contested assumptions about the nature and objectivity of ethics; they further suggest that it is often more conceptually useful to speak in terms of values and value alignment rather than ethics as such [7]. The present article takes up a related question within value alignment itself. Even if the discussion is framed in terms of values rather than ethics, one must still ask how alignment standards acquire normative authority. The central issue is therefore not only whose values are represented, but whether the empirical representation of those values can justify the standards by which AI systems are guided.

To develop this argument, the article examines three empirical routes to AI value alignment. The first is preference-based alignment, which treats human preferences, especially informed or reflectively endorsed preferences, as privileged indicators of value. The second is consensus-based alignment, which seeks normative weight in collective agreement, deliberative convergence, or cross-cultural endorsement. The third is data-based alignment, which treats large-scale datasets, textual corpora, or revealed patterns of human practice as resources from which AI systems may learn socially embedded values. The article reconstructs the strongest philosophical appeal of each route before identifying its normative limit.

The analysis is organized around two minimal conditions for normative grounds: universalizability and justifiability. Universalizability requires that a proposed normative ground be capable, at least in principle, of applying to all relevantly situated agents, rather than deriving its authority merely from the standpoint of a particular individual, group, or dataset. Justifiability requires that a standard be defensible to those who are subject to it, especially when they do not already share the preferences, consensus, or practices from which the standard is derived. These conditions do not amount to a complete moral theory. They function as minimal criteria for distinguishing normative authority from empirical convergence.

The contribution of this article is twofold. First, it clarifies a hidden assumption in empirical approaches to AI alignment: the assumption that human preferences, collective agreement, or large-scale data can generate normative authority from within their own empirical structure. Second, it offers a more precise account of the role that empirical inputs should play in AI ethics. The article distinguishes empirical input, training signal, evaluative metric, and normative standard, arguing that success at the first three levels should not be treated as sufficient for the fourth. Empirical inputs should therefore be understood as epistemic and operational resources within a broader justificatory structure. They can inform normative reasoning by revealing human concerns, conflicts, expectations, and vulnerabilities, but the authority of alignment standards must ultimately be secured by ethical argument, legal principles, institutional procedures, or other forms of public justification that cannot be reduced to empirical description. Whereas much of the alignment literature asks how AI systems can be made responsive to human values, this article asks a prior justificatory question: when, and under what conditions, empirically derived alignment standards can claim normative authority.

The remainder of the article proceeds as follows. Section 2 identifies the role of empirical inputs in AI value alignment and distinguishes preference-based, consensus-based, and data-based routes. Section 3 develops the analytical framework by explaining why normative grounds require universalizability and justifiability. Sections 4, 5, and 6 examine the three empirical routes in turn,

DOI: 10.54254/2753-7048/2026.LD33969

reconstructing their strongest philosophical appeal and showing why each remains normatively insufficient on its own. Section 7 repositions empirical alignment as a valuable but non-self-sufficient coordination mechanism. The conclusion argues that AI alignment should not be understood as a self-standing source of normativity, but as a set of technical and institutional practices that require independent normative justification.

2. Empirical inputs in AI value alignment

The previous section identified a justificatory question that arises once empirical inputs are treated as more than practical resources for alignment. This section clarifies the three empirical routes examined in the article. The labels "preference-based," "consensus-based," and "data-based" alignment are analytical rather than classificatory. They do not describe mutually exclusive schools of alignment research, nor do they exhaust the ways in which empirical information enters AI design and governance. In practice, the three often overlap. Human feedback can express individual preferences, reflect local norms, and become training data. Participatory procedures can generate collective judgments that are later translated into evaluation criteria. Datasets may contain traces of both individual choices and socially embedded expectations. The distinction is nevertheless useful because each route gives a different answer to the same underlying question: what kind of human experience should guide alignment standards?

The first route is preference-based. Here human preferences, choices, or evaluative judgments are treated as central signals for shaping AI behaviour. Reinforcement learning from human feedback, for example, uses human comparisons or evaluations to train reward models and adjust system behaviour [2]. In language model alignment, human feedback has been used to make models more responsive to user instructions and more likely to produce outputs preferred by human evaluators [3]. These methods should not be read as claiming to solve the whole philosophical problem of value alignment. Their immediate aim is narrower: to make AI systems better track patterns of human approval, disapproval, or preference in specified tasks.

The attraction of this route is not merely technical. If AI systems are to be responsive to human values, it seems natural to begin with what human agents prefer, endorse, or reject. The stronger version of the view does not equate value with whatever users happen to want at a given moment. It appeals instead to preferences that are informed, considered, or reflectively endorsed. Gabriel's distinction between revealed preferences, ideal preferences, interests, and values is important here, because it prevents preference-based alignment from collapsing into the simple reproduction of immediate user choice [1]. The preference-based route is therefore most plausible when preferences are treated not as raw desires, but as structured forms of human evaluation.

The second route is consensus-based. Here the relevant empirical input is not the preference of a single individual, but some form of collective judgment. In AI ethics and governance, this route appears in appeals to participatory design, stakeholder engagement, cross-cultural evaluation, democratic input, and deliberative procedures. Its motivation is straightforward: AI systems often affect many people, including those who do not directly provide feedback or interact with the system as users. Alignment standards derived only from developers, corporate actors, expert communities, or narrow pools of annotators risk reflecting the assumptions of a limited source group. Consensus-based approaches respond by widening the range of voices that count in the formation and assessment of alignment standards.

This route has a different appeal from preference-based alignment. It seeks legitimacy not primarily through individual endorsement, but through collective convergence under conditions that are, at least ideally, inclusive and deliberative. Recent philosophical work on language models and

DOI: 10.54254/2753-7048/2026.LD33969

human values emphasizes that alignment cannot be reduced to satisfying individual user requests, since conversational systems operate within social contexts shaped by norms, roles, and expectations [4]. This supports the need to consider the collective and institutional dimensions of alignment, even though it does not yet show that consensus itself can serve as a sufficient normative ground.

The third route is data-based. Here alignment is informed by large-scale datasets, textual corpora, behavioural traces, and other records of human practice. Modern AI systems, especially large language models, are trained on bodies of text and other data that record how people describe the world, express evaluations, negotiate norms, and reproduce social assumptions. These data are not merely technical raw material. They shape what models can generate, which associations they learn, and which social patterns they reproduce. For this reason, discussions of large language models often emphasize risks arising from training data, including bias, stereotyping, toxicity, exclusion, and the amplification of harmful patterns [5, 6].

The data-based route is appealing because it appears to capture human practice at scale. Unlike self-reported preferences or small deliberative exercises, large-scale data can reveal patterns that are distributed across many contexts and would otherwise remain difficult to observe. In its strongest form, this route does not treat data as mere frequency counts. It treats them as evidence of socially embedded practices from which AI systems can learn expectations, norms, and evaluative regularities. That makes data epistemically powerful. It also makes the normative question harder to avoid: if data reveal how people actually speak, classify, and behave, which of those patterns should AI systems reproduce, correct, or refuse?

The three routes therefore identify different aspects of human experience that any serious account of alignment must take into account. Preferences matter because AI systems need to respond to human evaluation. Consensus matters because alignment standards operate across communities and institutions. Data matter because models learn from the traces of social life that datasets make available. The target of the following analysis is not the use of these empirical inputs as such. It is the stronger claim that preferences, consensus, or data can generate the authority of alignment standards from within their own empirical structure. The next section develops the criteria by which that claim will be assessed.

3. Normative authority, universalizability, and justifiability

The preceding section distinguished three empirical routes through which human experience enters AI alignment: preferences and feedback, collective judgment, and large-scale data. Each can inform alignment practice. The question now is more demanding: what would be required for such inputs to function not merely as evidence, signals, or resources, but as normative grounds?

In this article, normative authority refers to the capacity of an alignment standard to provide justified reasons for guiding AI design, deployment, or governance. A standard does not have such authority simply because it is widely accepted, frequently observed, or technically useful. It has normative authority only if it can be defended as a reason-giving standard: something that can explain why a system ought to be designed, constrained, or evaluated in one way rather than another. This definition is deliberately modest. It does not presuppose a comprehensive moral theory, nor does it require actual agreement among all reasonable agents. It requires only that an alignment standard be more than a record of what people prefer, accept, or do.

This distinction can be framed as a difference between empirical convergence and normative justification. Empirical convergence occurs when individuals, groups, or datasets display stable patterns of judgment or behaviour: users may prefer one model response over another; evaluators

DOI: 10.54254/2753-7048/2026.LD33969

may agree that a certain output is harmful; a dataset may reveal recurring associations between concepts, identities, or social roles. Such convergence can be practically significant. It may help train models, detect risks, or guide institutional design. Yet it remains a fact about how people judge, choose, speak, or act. It does not, by itself, establish why those judgments, choices, linguistic patterns, or practices should guide AI systems.

The underlying issue is familiar in meta-ethics. Hume's discussion of the transition from "is" to "ought" identifies a gap between descriptive claims and normative conclusions [8]. Moore's criticism of attempts to define the good in purely natural or descriptive terms points to a related difficulty: even if a property is widely desired, observed, or approved, it remains a further question whether it is good or normatively authoritative [9]. This article does not rely on a full reconstruction of either Hume's or Moore's moral philosophy. It uses their shared insight in a limited way: facts about human attitudes, practices, or social patterns do not automatically generate reasons that should govern action.

The distinctive problem in AI alignment is that this transition can be obscured by operational success. A feedback signal may make a model more responsive to evaluators. A participatory process may produce a stable standard for system design. A dataset may allow a model to reproduce social practice with high accuracy. These achievements are not trivial. They can improve performance, reveal harms, and make systems more responsive to context. But they do not settle the justificatory question. A model can be better aligned with a feedback signal without that signal being normatively justified. A consensus can stabilize decision-making without showing that the resulting standard is fair. A dataset can represent social practice accurately without showing that the practice should be reproduced.

For an empirical input to ground an alignment standard, it must therefore satisfy conditions that are not reducible to empirical description. Two such conditions are central here: universalizability and justifiability. They are not offered as a complete moral theory, but as minimal tests for whether a proposed ground can plausibly claim normative authority rather than merely empirical relevance.

The first condition is universalizability. A proposed normative ground must be capable, at least in principle, of applying to all relevantly situated agents rather than deriving its force merely from the standpoint of a particular individual, group, platform, institution, or dataset. Kant's formulation of the moral law in terms of universal law provides the classical expression of this requirement [10]. The present argument does not assume Kant's full moral philosophy. It takes a thinner point from the universalizability tradition: a standard cannot claim normative authority if its validity depends solely on the special position of those who happen to express, endorse, or generate it.

This matters for AI alignment because the sources of alignment input are always partial. Users who provide feedback are not identical with all affected parties. Annotators and evaluators may not represent the full diversity of those subject to the system's outputs. Training data may be drawn disproportionately from certain languages, platforms, regions, or social groups. If an alignment standard is derived from such inputs, one must still ask why it should apply beyond the source from which it was drawn. Better sampling may improve representativeness, but representativeness alone does not answer the normative question. The issue is whether the principle expressed by the input can be justified beyond the contingent boundaries of its origin.

The second condition is justifiability. A normative standard must be defensible to those who are subject to it, especially when they do not already share the preferences, consensus, or practices from which it is derived. Scanlon's contractualist account captures this point by connecting moral justification to principles that others could not reasonably reject under appropriate conditions [11]. Again, the present article does not adopt Scanlon's contractualism as a complete moral theory. It

DOI: 10.54254/2753-7048/2026.LD33969

draws on a minimal requirement: a standard must be supportable by reasons that can be offered to others, not merely by pointing to the fact that some people prefer it, some group accepts it, or some dataset contains it.

Justifiability becomes especially important where alignment standards concern contested matters: harmful speech, refusal rules, paternalistic intervention, fairness trade-offs, cultural variation, or the weighting of safety against user autonomy. In such cases, an appeal to preference, consensus, or data does not by itself settle the matter. Those who reject the standard may reasonably ask why they should accept it. A satisfactory answer cannot simply repeat the empirical fact that others preferred, accepted, or practised it. It must give reasons capable of addressing those affected by the standard as participants in a shared normative space.

Universalizability and justifiability therefore mark the threshold between empirical relevance and normative authority. Universalizability asks whether a proposed ground can extend beyond the contingent standpoint of its source. Justifiability asks whether the resulting standard can be defended to those governed or affected by it. A preference, consensus, or data pattern may be valuable as evidence, as a training signal, as a heuristic, or as a coordination device. But unless it can meet these conditions, it cannot by itself function as a normative ground for AI alignment.

This framework also clarifies the scope of the argument that follows. The claim is not that empirical inputs are irrelevant to normative reasoning. They may reveal needs, vulnerabilities, harms, conflicts, and expectations that ethical theory and institutional design must take seriously. The claim is that empirical inputs require interpretation and justification before they can guide AI systems as normative standards. Preferences must be assessed in light of the conditions under which they are formed. Consensus must be examined in relation to inclusion, deliberation, and dissent. Data must be evaluated in relation to the social practices and power relations it records. In each case, empirical information can inform normative judgment, but it cannot replace it.

The next three sections apply this framework to the strongest versions of the three empirical routes. Section 4 considers whether informed or reflectively endorsed preferences can ground alignment standards. Section 5 asks whether deliberative or inclusive consensus can provide normative authority. Section 6 examines whether large-scale data and traces of human practice can function as normative grounds. The analysis will show that each route contributes something important to AI alignment, but each ultimately requires a normative premise that cannot be generated from empirical input alone.

4. Preference-based alignment: its appeal and limit

Preference-based alignment begins from an appealing thought: if AI systems are to be aligned with human beings, then human choices, approvals, rejections, and evaluative judgments must matter. In technical practice, this thought appears most clearly in methods that use human feedback to shape model behaviour. Reinforcement learning from human feedback uses human comparisons or evaluations as signals for reward modelling and policy improvement [2]. In language model alignment, human feedback has also been used to make models more responsive to user instructions and more likely to produce outputs preferred by human evaluators [3]. In these settings, preferences are not treated as abstract philosophical entities. They are operationalised through rankings, comparisons, corrections, approvals, and other evaluative signals.

The appeal of this approach is not merely one of technical convenience. Designers of AI systems often cannot specify in advance all the values, constraints, and context-sensitive expectations that should govern model behaviour. Human feedback offers a way of incorporating evaluative information that would be difficult to encode directly. It can reveal whether system outputs are

DOI: 10.54254/2753-7048/2026.LD33969

intelligible, useful, intrusive, offensive, unsafe, or socially inappropriate. A system trained with human feedback may therefore be more responsive to users and evaluators than one optimised only for a predefined objective. This is a genuine achievement of preference-based methods.

The strongest version of the preference-based route does not identify value with whatever users happen to want at a given moment. That view would be too crude. Human beings can have impulsive, misinformed, manipulated, or self-defeating preferences. A more defensible version appeals instead to preferences that are informed, considered, or reflectively endorsed. On this view, what matters is not an agent's immediate desire, but what the agent would endorse under conditions of adequate information, reflection, and freedom from obvious distortion. Gabriel's distinction between revealed preferences, ideal preferences, interests, and values is important here because it prevents preference-based alignment from collapsing into the simple reproduction of immediate user choice [1].

This idealised version has a serious philosophical attraction. Several strands of liberal moral and political philosophy connect respect for persons with attention to individuals' own judgments about what is good for them. Rawls, for example, links a person's good to a rational plan of life under appropriate conditions of deliberation [12]. Parfit's discussion of desire-fulfilment theories similarly distinguishes crude desire satisfaction from more informed or idealised forms of preference satisfaction [13]. These traditions suggest that preferences can matter normatively when they are connected to agency, reflection, and self-authorship. To ignore a person's considered preferences may appear paternalistic; to take them seriously may appear to respect that person as a rational agent.

The difficulty is that even idealised preference remains vulnerable at two levels. The first concerns the conditions under which preferences are formed. Preferences often reflect social background, unequal options, internalised expectations, or adaptive responses to constraint. Sen's work on development and freedom is relevant here: people living under deprivation or social subordination may adapt their expectations to unjust circumstances, coming to prefer or accept arrangements that limit their own freedom or well-being [14]. In AI alignment, this means that human feedback may reproduce not only considered judgments, but also social pressure, unequal expectation, and normalised constraint.

A defender of preference-based alignment might reply that this objection targets only defective preferences. The relevant input, they might say, should be purified: only well-informed, reflective, non-adaptive, and procedurally reliable preferences should count. This reply strengthens the view, but it does not remove the deeper problem. Even an idealised preference tells us what an agent would endorse under specified conditions. It does not, by itself, explain why that endorsement should govern the behaviour of an AI system in relation to other agents, institutions, or affected communities. The move from "this is what a person would reflectively prefer" to "this is what an AI system ought to do" still requires an additional normative premise.

Williams's distinction between internal and external reasons helps clarify the point [15]. A preference may provide an internal reason for the agent who has it, especially when it is embedded in that agent's motivational structure. If I reflectively prefer X, then X may give me a reason to act. But it does not automatically give others a reason to organise institutions, technologies, or social systems around X. For that further conclusion, one needs an additional principle: perhaps that autonomy should be respected, that each person's preferences should count equally, that welfare consists in preference satisfaction, or that systems should maximise the satisfaction of aggregated preferences. These are normative principles. They are not generated by the mere fact that preferences exist.

DOI: 10.54254/2753-7048/2026.LD33969

The two criteria developed in the previous section make this limit more precise. Preference-based alignment struggles with universalizability when the authority of a preference depends on the standpoint of the person or group from which it is drawn. A user's preference may be important evidence of what matters to that user, but it is not automatically a standard for all similarly situated agents. The problem becomes sharper when AI systems affect people who did not provide feedback, whose preferences conflict with those of evaluators, or whose interests are not represented in the preference data. Expanding the pool of feedback providers may improve representativeness, but it does not by itself show that the resulting preference structure has universal normative force.

The same difficulty appears with justifiability. Those subject to an alignment standard can reasonably ask why they should accept it. A satisfactory answer cannot simply be that some users preferred it, some annotators endorsed it, or some reward model learned it. If a model refuses a request, prioritises one user's interest over another's, filters certain content, or ranks certain outputs as preferable, the affected parties may demand reasons. Preference data may help explain how the system was trained, but explanation is not justification. To justify the standard, one must appeal to considerations such as harm prevention, fairness, autonomy, rights, welfare, or institutional legitimacy. These considerations go beyond preference as such.

Preferences therefore matter, but their significance is conditional. A defensible alignment standard should normally take account of what users and affected parties care about. Ignoring human preferences altogether would make AI systems unresponsive and potentially authoritarian. Yet preferences can inform alignment standards without independently justifying them. Their normative role depends on further principles specifying which preferences count, under what conditions they count, how conflicts among preferences should be handled, and why the resulting standard should be acceptable to those affected by it.

For AI alignment, the conclusion is limited but important. Human feedback is a powerful training signal, not a complete source of normative authority. Preference learning can help systems approximate human evaluative patterns, but it cannot settle which patterns should be preserved, revised, aggregated, or rejected. The design of feedback procedures, the selection of evaluators, the treatment of disagreement, and the interpretation of preference data all require normative judgment. Preference-based alignment contributes to the epistemic and operational side of AI ethics, but it cannot by itself ground the legitimacy of alignment standards.

5. Consensus-based alignment: its appeal and limit

Consensus-based alignment shifts the focus from individual endorsement to collective judgment. The motivation is straightforward. AI systems often affect people who are not direct users, who never provide feedback, and who may be harmed by standards formed within narrow technical or institutional settings. If alignment standards are derived only from developers, corporate actors, expert communities, or limited pools of annotators, they risk reflecting a restricted social standpoint. For this reason, AI ethics and governance increasingly emphasize stakeholder participation, inclusive design, cross-cultural evaluation, deliberative procedures, and mechanisms for incorporating diverse perspectives into AI development and assessment [16, 17].

The attraction of this route is that it seems to answer a weakness in preference-based alignment. Individual preferences are plural, unstable, and often conflicting. Collective judgment appears to offer a more public basis for alignment, especially when AI systems mediate communication, rank information, filter content, support institutional decisions, or shape access to opportunities. A standard formed through broader participation seems less likely to reflect the preferences of a single user group and more likely to register the concerns of those affected by the system. Consensus-based

DOI: 10.54254/2753-7048/2026.LD33969

alignment therefore seeks legitimacy not through isolated choice, but through forms of collective convergence.

The strongest version of this route is not simple majoritarianism. It does not hold that whatever most people believe should become the alignment standard. Majorities can be misinformed, oppressive, or inattentive to minority claims. A more plausible version appeals to inclusive and deliberative agreement: collective judgment matters when it emerges from procedures in which affected parties can offer reasons, challenge assumptions, and contest proposed standards. The relevant ideal is not agreement as such, but agreement shaped by participation, reflection, and freedom from domination.

This stronger version has clear philosophical roots. Habermas's discourse ethics connects normative validity with the possibility of rational acceptance in practical discourse [18]. A norm is not legitimate merely because it is accepted as a social fact; its authority depends on whether it could be accepted by those affected under conditions of free and equal deliberation. Scanlon's account of justifiability supports a related idea from a different perspective: principles must be defensible to others in terms they could not reasonably reject [11]. Both views link normative authority to public reason-giving rather than to private preference or mere aggregation.

In AI alignment, this orientation gives participatory and deliberative approaches real force. Affected communities should not be treated merely as data sources, annotator pools, or passive recipients of technical decisions. Their judgments can reveal harms, contextual meanings, and practical constraints that developers or external experts may miss. Participatory processes can also expose value conflicts hidden by apparently neutral technical objectives. In this respect, consensus-based alignment improves upon narrow preference-based approaches: it shifts attention from individual satisfaction to social legitimacy, inclusion, and contestability.

The difficulty is that actual consensus is rarely formed under ideal conditions. Participation may be limited, symbolic, or unevenly distributed. Some groups may lack the resources, time, language access, technical knowledge, or institutional power needed to shape the process meaningfully. Others may be consulted without having any real influence over design decisions. Recent work on participatory AI design emphasizes both the appeal of participation and the difficulty of granting stakeholders substantive agency [16]. Critics of participation in machine learning similarly warn that participatory procedures can become a design fix that leaves deeper structures of power, extraction, and inequality intact [17].

This matters because the existence of a participatory process does not by itself make the resulting agreement authoritative. A group may converge because alternatives were excluded, dissenting voices were marginalised, participants adapted to institutional expectations, or the problem was framed in a way that made some options appear unavailable. In such cases, consensus may tell us more about the structure of participation than about the legitimacy of the norm. The fact that a judgment is collectively produced does not yet show that it is justified.

A defender of consensus-based alignment might reply that this objection targets defective procedures, not consensus as such. The relevant standard, they might argue, is not actual agreement, but idealised agreement under conditions of inclusion, equality, information, and freedom from coercion. This reply is important. It explains why consensus-based alignment is stronger than simple aggregation. Yet it also reveals the deeper dependence of consensus on prior normative standards. To say that a procedure is fair presupposes an account of fairness. To say that affected parties had a meaningful voice presupposes an account of who counts as affected and what meaningful participation requires. To say that no one could reasonably reject the outcome presupposes standards

DOI: 10.54254/2753-7048/2026.LD33969

of reasonable rejection. These criteria are not generated by agreement itself; they are the conditions under which agreement becomes normatively significant.

Universalizability exposes the same point from another angle. Every actual consensus has boundaries. It includes some participants and excludes others; it is formed within particular languages, institutions, jurisdictions, communities, and historical circumstances. Even broad stakeholder processes cannot literally include all present and future persons affected by an AI system. A consensus may therefore be highly informative without being universally authoritative. The question is not only whether agreement was reached, but why that agreement should extend to those outside the process, to those whose dissent was not incorporated, or to those who will be affected later. Expanding participation can reduce this problem, but it cannot remove the need for a principle explaining how far the authority of the consensus extends.

Justifiability raises a related challenge. Those subject to an alignment standard may reasonably ask why they should accept it, especially when they disagree with the consensus or when the consensus reflects values they do not share. In conditions of value pluralism, disagreement is not always a sign of ignorance, irrationality, or bad faith. Berlin's account of value pluralism emphasizes that human values may be multiple, genuine, and sometimes in conflict [19]. If so, consensus cannot be assumed to dissolve normative conflict. It may organise disagreement, reduce uncertainty, or provide a workable institutional settlement, but it does not eliminate the need to justify why one settlement should guide the system rather than another.

The implication for AI alignment is that participatory evaluation, stakeholder engagement, and cross-cultural consultation are necessary but incomplete. They can reveal whose values are ignored, which harms are invisible to designers, and where apparently universal standards are in fact local. But they do not make normative reasoning unnecessary. Designers and institutions still have to justify why a procedure is fair, why certain participants were included, how dissent was treated, and why the resulting standard should guide system behaviour. Without such justification, consensus remains an empirical or procedural fact rather than a complete normative ground.

Consensus-based alignment therefore has genuine value, but its authority depends on principles that govern participation, deliberation, equality, dissent, and the scope of application. Those principles cannot be derived from consensus alone, because they are what allow us to distinguish legitimate consensus from distorted, exclusionary, or merely strategic agreement. Consensus can inform alignment standards, and in many cases it should. What it cannot do is independently ground their normative authority.

6. Data-based alignment: its appeal and limit

Data-based alignment turns from explicit endorsement to the traces of human practice. Instead of asking primarily what individuals prefer or what groups collectively accept, it looks to large-scale datasets, textual corpora, behavioural traces, and other records of social life. In contemporary AI systems, especially large language models, such data are not peripheral. They shape the patterns of language, association, classification, and evaluation that models learn. If alignment concerns how AI systems respond to human social contexts, then the data through which those contexts are represented become ethically significant.

The appeal of this route is that data seem to provide access to human practice at scale. People may state their values inconsistently, respond differently across contexts, or describe themselves in socially desirable terms. Textual corpora, by contrast, contain traces of how people describe the world, express approval and disapproval, negotiate norms, justify institutions, and reproduce social categories. Behavioural data may reveal what people actually choose, not merely what they say they

DOI: 10.54254/2753-7048/2026.LD33969

value. Data-based alignment can therefore appear more realistic than approaches that rely only on abstract principles, individual testimony, or small-scale deliberation.

This thought has a recognizable epistemic background. In economics, Samuelson's revealed preference theory shifted attention from unobservable mental states to observable choice behaviour [20]. The point was not that observed choices are morally authoritative, but that they can provide evidence about preference structures without relying on introspection. A related, though broader, pragmatist impulse can be found in Dewey's theory of valuation, where value judgments are connected to human practices, consequences, and processes of inquiry rather than detached from experience [21]. These traditions help explain why data-based approaches can appear powerful: they promise access to value-relevant patterns through observable practice rather than through moral speculation alone.

In AI alignment, the appeal is intensified by scale. Large datasets include ordinary language use, institutional documents, cultural products, technical materials, and everyday interactions. A model trained on such data may learn not only grammar and factual associations, but also patterns of politeness, refusal, explanation, stereotyping, authority, social roles, and moral evaluation. In its strongest form, the data-based route does not treat data as mere frequency counts. It treats them as evidence of socially embedded practices from which AI systems can learn expectations, norms, and evaluative regularities.

This stronger version should not be dismissed. Data can reveal patterns that individual reflection or formal deliberation may miss. They can expose recurrent harms, discriminatory associations, toxic language, representational gaps, and differences across social contexts. Work on language model risks has shown that large-scale models can reproduce and amplify harmful associations found in training data, including stereotypes, exclusionary representations, toxic content, and misleading or unsafe outputs [5, 6]. Data analysis can therefore play an important diagnostic role: it can make visible the social patterns that alignment efforts must address.

The first difficulty is that data are not neutral mirrors of human practice. Datasets are produced through selection, collection, filtering, labelling, platform design, moderation policy, language distribution, and institutional incentives. What appears in data is not "humanity" as such, but the result of historically and technically mediated processes. Some groups, languages, and social positions are overrepresented; others are underrepresented or misrepresented. Some communities leave fewer digital traces, while others are disproportionately exposed to surveillance or extraction. Data therefore carry not only information about social life, but also the conditions under which that information was produced.

This problem is not merely theoretical. Buolamwini and Gebru's study of commercial gender classification systems showed substantial intersectional accuracy disparities, with darker-skinned women being misclassified at much higher rates than lighter-skinned men [22]. The significance of this case is not limited to facial analysis. Although the mechanisms differ across facial analysis systems and language models, the case illustrates a broader point about data-driven systems: when datasets underrepresent or misrepresent particular groups, systems trained on those datasets can reproduce the structure of that exclusion. In language models, similar concerns arise when training data encode stereotypes, toxic associations, or dominant cultural assumptions.

A defender of data-based alignment might reply that these are correctable defects. Better datasets, better documentation, better filtering, better sampling, and better evaluation could reduce bias and improve representation. This response is important. Dataset documentation and accountability mechanisms can make data practices more transparent and contestable [23]. More representative data may also reduce certain forms of harm. But this response addresses only the first layer of the

DOI: 10.54254/2753-7048/2026.LD33969

problem. It assumes that if data were sufficiently accurate, inclusive, and well curated, they could provide a reliable foundation for alignment. The deeper question is whether even ideal data could generate normative authority on their own.

They could not. Even a dataset that was accurate, comprehensive, and representative would still show how people speak, act, classify, value, and interact under particular historical and social conditions. It would not show which of those patterns should guide AI systems. A dataset may accurately record that a stereotype is common; that does not mean the stereotype should be reproduced. It may accurately record that certain groups are associated with lower status in existing discourse; that does not mean a model should preserve that association. It may accurately record that exclusionary or hierarchical norms have been widespread; that does not make those norms legitimate. Accuracy and authority are different properties.

This is the structural limit of data-based alignment. Data can reveal regularities, but they cannot decide which regularities should be preserved, modified, or rejected. That decision requires normative judgment. A model designer may suppress toxic patterns, counteract stereotypes, privilege safety over literal imitation, or refuse to reproduce certain forms of social bias. These decisions cannot be made by data alone, because they involve standards about harm, fairness, dignity, equality, autonomy, or public legitimacy. The more ethically consequential the system is, the less adequate it becomes to say that the system should simply reflect the data distribution.

Universalizability helps identify the problem. Data are generated from particular sources: platforms, languages, regions, institutions, communities, and historical periods. Even when a dataset is large, its source conditions remain particular. An alignment standard derived from such data therefore requires justification beyond the fact that the data exist. Why should patterns from one set of digital environments guide systems used in other social, linguistic, or cultural contexts? Why should the behaviour of those who are most visible in the data become a standard for those who are less visible or absent? Scale can make a pattern more statistically robust, but it does not make it universally valid.

Justifiability raises an equally serious challenge. People affected by an AI system may reasonably reject the idea that past data should determine how the system treats them. A person may object to being represented through historically biased categories. A community may reject dominant narratives embedded in textual corpora. A marginalized group may reasonably resist a system that treats existing patterns of speech or behaviour as evidence of legitimate social norms. In such cases, pointing to the dataset explains where the pattern came from, but it does not justify why the pattern should govern the system's output or decision.

Data-based alignment is therefore best understood as epistemically powerful but normatively insufficient. Data can help identify the social world in which AI systems operate. They can reveal expectations, conflicts, harms, and inherited biases. They can also support evaluation by showing how models reproduce or transform patterns found in their training environments. But they cannot determine, on their own, what should count as an acceptable alignment standard. The ethical task is not simply to learn from data, but to decide which aspects of the data should be treated as informative, which should be corrected, and which should be rejected.

The implication for AI alignment is direct. Large-scale data and textual corpora are indispensable for building modern AI systems. They can supply linguistic competence, social knowledge, and empirical evidence of human practice. They can also expose risk. Yet the transition from "this pattern exists in the data" to "this pattern should guide AI behaviour" requires independent justification. Data-based alignment therefore shares the same structural limit as preference-based

DOI: 10.54254/2753-7048/2026.LD33969

and consensus-based alignment: it depends on empirical inputs that are indispensable for practice, but insufficient for normative authority.

7. Repositioning empirical alignment: coordination without normative self-sufficiency

The preceding analysis does not show that empirical alignment should be abandoned. It shows that its role must be more carefully specified. Preferences, consensus, and data each provide important forms of access to human evaluation, social disagreement, and patterns of practice. What they do not provide, on their own, is the authority of the standards into which they are translated. Empirical alignment should therefore be understood not as a source of normativity in itself, but as a set of epistemic and operational practices that must be situated within a broader justificatory structure.

A useful way to make this point is to distinguish four levels that are often conflated in discussions of alignment. The first is empirical input: the preferences, judgments, data, and practices from which information is drawn. The second is training signal: the way such input is converted into reward modelling, fine-tuning, filtering, model optimisation, or evaluation. The third is evaluative metric: the criterion by which system performance is assessed, such as helpfulness, harmlessness, user satisfaction, toxicity reduction, fairness, or refusal accuracy. The fourth is normative standard: the reason-giving principle that explains why a particular objective, constraint, or metric should guide the system in the first place. Confusion arises when success at the first three levels is treated as sufficient for the fourth.

This distinction matters because technical success can mask justificatory incompleteness. A system may optimise a reward model trained from human feedback, but the procedure still leaves open why the feedback should be authoritative. A model may perform well on an evaluation benchmark, but the benchmark still requires justification as a measure of what matters. A moderation system may reflect a community standard, but the legitimacy of that standard depends on how the community was defined, whose speech was protected or restricted, and what reasons support the rule. A language model may reduce toxic outputs, but the definition of toxicity itself depends on normative choices about harm, context, power, and contestability. In each case, empirical success makes the standard operationally effective; it does not make it normatively justified.

This point also clarifies the role of AI ethics principles. High-level principles such as fairness, accountability, transparency, privacy, beneficence, and non-maleficence are often invoked to guide AI governance, and influential frameworks have sought to organise such principles into more coherent ethical architectures [24]. Comparative work on AI ethics guidelines has shown both convergence around recurring principles and significant divergence in their interpretation and implementation [25]. This supports the present argument. Agreement at the level of general principles does not remove the need for normative reasoning about meaning, priority, scope, and institutional application. Empirical inputs can inform such reasoning, but they cannot replace it.

Nor can abstract principles solve the problem on their own. Mittelstadt argues that high-level AI ethics principles do not guarantee ethical AI unless they are connected to institutional mechanisms, professional responsibilities, and accountability structures [26]. This observation prevents a simple reversal. The answer to empirical insufficiency is not to abandon empirical inputs in favour of moral theory alone. A more plausible view is that empirical inputs, ethical principles, legal norms, and institutional procedures must be connected. Empirical alignment needs normative guidance; normative guidance needs institutional implementation; institutional implementation needs empirical feedback about effects, failures, and unintended consequences.

DOI: 10.54254/2753-7048/2026.LD33969

The result is a layered account of alignment. At the empirical level, designers and institutions gather information from users, affected parties, evaluators, datasets, and social contexts. At the interpretive level, they determine what that information means, which conflicts it reveals, and which limitations it contains. At the normative level, they justify the principles that should guide system behaviour, including principles concerning harm, fairness, autonomy, rights, dignity, or public legitimacy. At the institutional level, they establish procedures for accountability, contestation, revision, and oversight. These levels are interconnected, but they are not reducible to one another.

Such a layered account changes how empirical alignment should be evaluated. The relevant question is not simply whether a system has incorporated human feedback, consulted stakeholders, or been trained on large-scale data. It is whether those empirical inputs are embedded in processes that make their normative role explicit and contestable. Who provided the feedback, and under what conditions? Which affected parties were excluded? How was disagreement interpreted? Which data patterns were treated as informative, and which were treated as harmful? What principles guided these decisions? What mechanisms allow affected persons to challenge or revise the resulting standard? These questions show that empirical alignment becomes ethically meaningful only when linked to justification and accountability.

This has direct implications for AI governance. Governance should not treat empirical alignment as a substitute for public justification, nor should it treat technical metrics as neutral proxies for ethical legitimacy. If preferences are used, there should be an account of whose preferences count and why. If consensus is invoked, there should be an account of participation, exclusion, disagreement, and procedural fairness. If data are used, there should be an account of provenance, representation, bias, and the normative criteria for correction or refusal. In each case, legitimacy depends not only on empirical responsiveness, but on the reasons and institutions that organise that responsiveness.

The positive claim, then, is not that empirical alignment should be replaced by external moral theory. It is that empirical alignment should be placed within a broader structure of justification. Ethical theory can clarify the principles at stake. Legal norms can identify rights, duties, and institutional constraints. Participatory procedures can reveal affected perspectives and enable contestation. Technical methods can operationalise and test proposed standards. Empirical feedback can show whether those standards work as intended or reproduce new harms. None of these elements is sufficient on its own. Together, they provide a more credible model of alignment than any purely preference-based, consensus-based, or data-based route.

This repositioning avoids an unrealistic picture of normative justification. AI alignment need not wait until moral philosophy has resolved all deep disagreements. That would make governance impossible. The point is rather that practical alignment decisions should not obscure their normative character. When designers choose a reward model, select annotators, define harmful content, weight competing values, or filter training data, they are not merely making technical choices. They are making decisions with normative significance. Responsible alignment requires that these decisions be made explicit, reasoned, revisable, and accountable.

Empirical alignment can coordinate AI behaviour with human judgments, expectations, and practices. It can improve model responsiveness, reveal ethically relevant information, and support institutional learning about system failures and harms. What it cannot do is explain, by itself, why a particular alignment standard is legitimate. The normative authority of alignment standards must come from the justificatory structure in which empirical inputs are interpreted, constrained, and made accountable.

DOI: 10.54254/2753-7048/2026.LD33969

8. Conclusion

This article has examined a justificatory assumption that often remains implicit in empirical approaches to AI value alignment: the assumption that human preferences, collective judgments, or large-scale data can provide not only useful inputs for alignment, but also the normative authority of alignment standards themselves. The analysis has argued that this assumption cannot be sustained. Empirical inputs are indispensable for designing, training, evaluating, and governing AI systems, but they do not independently explain why a given alignment standard should be treated as normatively justified.

The argument has distinguished empirical convergence from normative justification. Preferences can show what individuals endorse, reject, or find acceptable under specified conditions. Consensus can show where collective judgment converges, especially when participation and deliberation are taken seriously. Data can show large-scale patterns of language, behaviour, classification, and social practice. These are all valuable sources of information. They make alignment more responsive to actual human contexts and help reveal harms, disagreements, vulnerabilities, and expectations that abstract principles alone may miss. Yet none of them can, by itself, explain why the resulting standard should guide AI behaviour.

The same structural limit appears in each route. Preference-based alignment requires a further principle explaining why particular preferences should count, under what conditions, and with what authority over others. Consensus-based alignment requires principles governing inclusion, deliberation, dissent, and the scope of collective agreement. Data-based alignment requires normative judgment about which patterns should be preserved, corrected, or refused. In each case, empirical inputs can inform normative reasoning, but they cannot replace it.

Universalizability and justifiability clarify why this limit matters. An alignment standard must be capable of extending beyond the contingent standpoint of the particular users, participants, platforms, institutions, or datasets from which it is derived. It must also be defensible to those governed or affected by it, especially when they do not already share the preferences, consensus, or practices on which it relies. Without these conditions, empirical inputs may still function as evidence, training signals, evaluative metrics, or coordination devices, but they cannot claim normative authority.

The conclusion is not that empirical alignment should be abandoned. Human feedback, participatory evaluation, dataset documentation, risk analysis, model auditing, and institutional monitoring remain essential to responsible AI development. They make AI systems more responsive to human concerns and more accountable to the contexts in which they operate. The point is rather that these practices should be located within a broader justificatory structure. Empirical alignment can help identify what people value, contest, fear, or experience as harmful; it cannot by itself determine which standards are legitimate. For governance purposes, this means that alignment procedures should be assessed not only by the quality of their empirical inputs or technical metrics, but also by the justificatory processes through which their standards are selected, contested, and revised.

AI alignment should therefore not be understood as a self-sufficient source of normativity. It is better understood as a set of technical and institutional practices that require independent normative justification. Recognising this boundary provides a more defensible starting point for AI ethics: one that preserves the practical value of empirical alignment while making explicit the ethical, legal, and institutional work required to justify the standards by which AI systems are guided.

DOI: 10.54254/2753-7048/2026.LD33969

References

[1] Gabriel, I.: Artificial intelligence, values, and alignment. Minds and Machines 30, 411–437 (2020). https: //doi.org/10.1007/s11023-020-09539-2

[2] Christiano, P.F., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems 30, pp. 4299–4307 (2017).

[3] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems 35, pp. 27730–27744 (2022).

[4] Kasirzadeh, A., Gabriel, I.: In conversation with artificial intelligence: aligning language models with human values. Philosophy & Technology 36, 27 (2023). https: //doi.org/10.1007/s13347-023-00606-x

[5] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. Association for Computing Machinery, New York (2021). https: //doi.org/10.1145/3442188.3445922

[6] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, M., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., Gabriel, I.: Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 214–229. Association for Computing Machinery, New York (2022). https: //doi.org/10.1145/3531146.3533088

[7] LaCroix, T., Luccioni, A.S.: Metaethical perspectives on 'benchmarking' AI ethics. AI and Ethics 5, 4029–4047 (2025). https: //doi.org/10.1007/s43681-025-00703-x

[8] Hume, D.: A Treatise of Human Nature. Edited by Norton, D.F., Norton, M.J. Oxford University Press, Oxford (2000).

[9] Moore, G.E.: Principia Ethica. Revised edn. Cambridge University Press, Cambridge (1993).

[10] Kant, I.: Groundwork of the Metaphysics of Morals. Translated and edited by Gregor, M., Timmermann, J. Cambridge University Press, Cambridge (2012).

[11] Scanlon, T.M.: What We Owe to Each Other. Harvard University Press, Cambridge, MA (1998).

[12] Rawls, J.: A Theory of Justice. Revised edn. Belknap Press of Harvard University Press, Cambridge, MA (1999).

[13] Parfit, D.: Reasons and Persons. Oxford University Press, Oxford (1984).

[14] Sen, A.: Development as Freedom. Oxford University Press, Oxford (1999).

[15] Williams, B.: Internal and external reasons. In: Moral Luck: Philosophical Papers 1973–1980, pp. 101–113. Cambridge University Press, Cambridge (1981).

[16] Delgado, F., Yang, S., Madaio, M., Yang, Q.: The participatory turn in AI design: theoretical foundations and the current state of practice. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–23. Association for Computing Machinery, New York (2023). https: //doi.org/10.1145/3617694.3623261

[17] Sloane, M., Moss, E., Awomolo, O., Forlano, L.: Participation is not a design fix for machine learning. In: Proceedings of the 2022 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–6. Association for Computing Machinery, New York (2022). https: //doi.org/10.1145/3551624.3555285

[18] Habermas, J.: Moral Consciousness and Communicative Action. Translated by Lenhardt, C., Nicholsen, S.W. MIT Press, Cambridge, MA (1990).

[19] Berlin, I.: Liberty: Incorporating Four Essays on Liberty. Edited by Hardy, H. Oxford University Press, Oxford (2002).

[20] Samuelson, P.A.: A note on the pure theory of consumer's behaviour. Economica 5(17), 61–71 (1938). https: //doi.org/10.2307/2548836

[21] Dewey, J.: Theory of Valuation. University of Chicago Press, Chicago (1939).

[22] Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (2018).

[23] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. Communications of the ACM 64(12), 86–92 (2021). https: //doi.org/10.1145/3458723

[24] Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harvard Data Science Review 1(1) (2019). https: //doi.org/10.1162/99608f92.8cd550d1

[25] Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 389–399 (2019). https: //doi.org/10.1038/s42256-019-0088-2

DOI: 10.54254/2753-7048/2026.LD33969

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