AI Pluralism and the Worlds It Misses

Rashid Mushkani [1 2]

Abstract

AI pluralism is often framed as a problem of representing diverse values, preferences, users, or outputs. This paper argues that this framing is incomplete because AI systems also impose ontologies: they define what counts as an entity, relation, feature, harm, benefit, and valid form of evidence. We define *ontological flattening* as the conversion of situated, contested, and historically specific meanings into a restricted technical category, proxy, aggregation rule, or benchmark target that is treated as neutral and difficult to contest. The paper develops a bounded conceptual and qualitative synthesis across value pluralism, pluralistic alignment, participatory and democratic AI, procedural justice, science and technology studies, accountability research, aggregate themes from 11 expert interviews, and three urban AI companion cases. The cases illustrate how pluralistic methods can improve or structure model behavior while still compressing categories, proxies, aggregation rules, and revision rights before affected actors have procedural standing. We introduce Pluralistic Lifecycle Governance (PLG) as a preliminary qualitative audit scaffold for documenting ontological openness, epistemic inclusion, procedural authority, evaluation pluralism, and lifecycle accountability. PLG is not presented as a validated scoring instrument; it is a framework for making the evidence and governance conditions of pluralistic AI explicit.

1. Introduction

AI systems do not only predict, rank, classify, or generate. They define the terms under which the world becomes computable. A street becomes a vector of visual features. A public-space rendering becomes a preference comparison. A community judgment becomes a benchmark label. A contested concept such as safety, accessibility, comfort, or inclusion becomes a category that a model can optimize. These translations are often necessary for computation, but they are not neutral. They decide which distinctions matter, which forms of knowledge count, and which conflicts remain visible.

Pluralistic alignment and participatory AI have become important responses to this problem. Recent work argues that AI systems should represent a wider range of values, expose multiple reasonable answers, support steering across perspectives, or match population distributions (Sorensen et al., 2024). Normative accounts further show that pluralistic value alignment requires choices about criteria, origins, measurement, aggregation, and legitimacy (Kasirzadeh, 2024). These developments reject the assumption that alignment means optimizing one global preference function. Yet they leave a prior question underdeveloped: when do pluralistic AI methods preserve contestable ways of world-making, and when do they compress them into fixed technical ontologies?

This paper argues that AI pluralism must address ontological pluralism as well as value pluralism. By *ontology*, we mean the practical representational commitments embedded in an AI system about what exists, what can be measured, how entities relate, and what counts as evidence. By *lifeworld*, we mean the situated background of meanings, identities, histories, and practical judgments through which people inhabit social life. By *procedural standing*, we mean the ability of affected actors to influence, contest, revise, or appeal decisions about a system rather than merely supply data to it. A public street can be enacted as infrastructure, memory, risk, care, commerce, surveillance, refuge, exclusion, or belonging. A benchmark that reduces this plurality to a stable label such as *inclusive* or *safe* may be useful, but it also performs a world-making act.

Habermas's account of lifeworld colonization provides one diagnostic vocabulary for this failure. He distinguishes communicatively reproduced lifeworlds from systems governed by money, power, and instrumental rationality (Habermas, 1984; 1987). In AI, colonization appears when contested judgments are turned into prediction tasks, local knowledge becomes a feature vector, public reason becomes a benchmark score, and communities may contest outputs but not the representational scheme that produced them. This is

[1]Université de Montréal, Montréal, Québec, Canada [2]Mila – Québec AI Institute, Montréal, Québec, Canada. Correspondence to: Rashid Mushkani <rashidmushkani@gmail.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

not only bias, misclassification, or measurement error. It is *ontological flattening*: the conversion of situated, contested, and historically specific worlds into restricted technical categories or proxies treated as neutral.

The empirical scope is deliberately bounded. We focus on urban, public-space, and vision-based systems because accessibility, safety, inclusion, comfort, and belonging are not stable visual properties. The evidence base combines a concept map of relevant literatures, 11 semi-structured expert interviews, and secondary analysis of three companion cases: Case A, a pluralistic public-space generation dataset and Direct Preference Optimization experiment; Case B, a participatory streetscape inclusivity model; and Case C, a reliability-aware urban vision-language benchmark (Mushkani et al., 2025; Mushkani & Koseki, 2026; Mushkani, 2025). The cases support analytic generalization about lifecycle mechanisms, not statistical generalization across AI domains.

The paper makes three contributions. First, it defines ontological flattening as a failure mode that can persist even when a system supports pluralistic outputs or collects pluralistic feedback. Second, it distinguishes *outcome pluralism*, where systems produce, steer, or approximate diverse outputs, from *procedural pluralism*, where affected actors have standing over categories, evidence, aggregation, evaluation, and revision. Third, it proposes Pluralistic Lifecycle Governance (PLG) as a qualitative framework for documenting and contesting ontological openness, epistemic inclusion, procedural authority, evaluation pluralism, and lifecycle accountability across the AI lifecycle.

2. Related Work

Classical value pluralism rejects the reduction of legitimate values to one master scale. Berlin argues that human goods may be real and legitimate while remaining incompatible (Berlin, 1969). Anderson and Chang show that values can be incomparable or incommensurable, making it misleading to assume that every conflict can be converted into one utility calculus (Anderson, 1993; Chang, 2015). For AI, the practical problem is that learning and evaluation procedures often require commensuration through labels, losses, rankings, utilities, scores, or preferences.

Science and technology studies and feminist epistemology shift the issue from value conflict to world-making. Classification systems organize work and social memory (Bowker & Star, 1999); objects are enacted differently across practices rather than merely viewed differently from one neutral standpoint (Mol, 2002); and objectivity requires accountable positioning rather than a view from nowhere (Haraway, 1988). Measurement and construct validity scholarship further shows that operational choices can fail when a construct is underspecified or when a social target is treated as directly measurable (Jacobs & Wallach, 2021). Fairness research likewise warns that abstraction can hide institutional context, background conditions, and social meaning (Selbst et al., 2019). These literatures imply that AI pluralism cannot be limited to different preferences over a fixed ontology. It must also ask who defines entities, relations, proxies, and evidentiary standards.

Alignment research asks how AI systems can act in accordance with human values and intentions (Gabriel, 2020; Kasirzadeh & Gabriel, 2023). Reinforcement learning from human feedback, instruction tuning, and constitutional approaches have improved model behavior (Christiano et al., 2017; Ouyang et al., 2022; Askell et al., 2021; Bai et al., 2022), but they often aggregate heterogeneous judgments into one training signal. Pluralistic alignment reframes disagreement as signal. Roadmaps distinguish Overton, steerable, and distributional pluralism; social choice approaches treat feedback as collective decision-making; and projects such as STELA, PRISM, and Collective Constitutional AI show concrete ways to elicit or represent plural input (Sorensen et al., 2024; Conitzer et al., 2024; Bergman et al., 2024; Kirk et al., 2024; Huang et al., 2024). Our argument is not that these methods fail. The distinction is that outcome-level pluralism can still operate over fixed categories, proxies, and aggregation rules. PLG asks whether those choices are themselves procedurally open.

Democratic AI and procedural justice provide the institutional counterpart. Democratic AI distinguishes advisory input from processes that bind, initiate, or govern metagovernance (Ovadya et al., 2025). Participatory AI warns that participation becomes tokenistic when communities are consulted after core decisions are fixed (Arnstein, 1969; Birhane et al., 2022; Delgado et al., 2023; Sloane et al., 2022). Design justice identifies community-led design and accountable redistribution of design power as central to equitable systems (Costanza-Chock, 2020). Procedural justice and due process add that algorithmic legitimacy requires voice, reasons, contestation, and audit trails (Kinchin, 2024; Citron, 2008; Pasquale, 2015). Accountability research explains why this cannot be solved at the model layer alone: AI responsibility is fragmented across model providers, data producers, deployers, and downstream institutions, and harms can arise throughout the machine-learning lifecycle (Widder & Nafus, 2023; Suchman, 2007; Suresh & Guttag, 2021). Documentation frameworks make disclosure more systematic, but disclosure does not by itself transfer authority (Gebru et al., 2021; Mitchell et al., 2019).

The contribution of ontological flattening is therefore not to restate that categories are political. That claim is established by prior work. The contribution is to specify a lifecycle failure mode for pluralistic AI: pluralistic inputs or outputs

Table 1. Terminology used throughout the paper.

Term	Meaning in this paper
Ontology	Practical representational commitments embedded in an AI system about entities, relations, measurable attributes, evidence, and valid outputs.
Lifeworld	Situated background of meanings, identities, histories, and practical judgments through which people experience and organize social life.
Ontological flattening	Conversion of situated and contested meanings into a restricted technical category, proxy, aggregation rule, or benchmark target that is treated as neutral and difficult to contest.
Outcome pluralism	Pluralism at the level of outputs, answers, steering options, or represented preference distributions.
Procedural pluralism	Pluralism at the level of authority over problem framing, categories, evidence, aggregation, evaluation, revision, and appeal.
Procedural standing	Capacity of affected actors to influence, contest, revise, veto, or appeal decisions about a system, including decisions about the ontology through which the system operates.
Lifecycle accountability	Assignment of responsibilities, evidence, recourse, audit cadence, revision triggers, and decommissioning conditions after design and deployment.

can coexist with fixed ontologies when category formation, proxy choice, aggregation, evaluation, and revision remain outside affected actors' authority.

3. Ontological Flattening

We define **ontological flattening** as the transformation of situated, contested, embodied, or historically specific meanings into singular or restricted categories and proxies treated as neutral descriptions of the world. Flattening occurs when disagreement is represented only as noise, non-response is treated as missingness, local categories are replaced by global labels, and a system's output is contestable but its ontology is not.

A diagnosis of flattening requires four conditions. The source concept is situated, contested, context-dependent, or historically specific. The AI system fixes that concept into a restricted category, proxy, aggregation rule, metric, or benchmark target. Disagreement, uncertainty, abstention, or context is erased, converted into noise, or made unavailable to downstream interpretation. Affected actors lack meaningful standing to contest or revise the representational commitment before or after use. These conditions distinguish flattening from ordinary abstraction. Compression is acceptable when the system's scope is limited, its rationale is documented, known alternatives and dissent are retained, and affected actors have revision or appeal rights.

Flattening is related to bias, misclassification, measurement error, proxy failure, representational harm, and tokenistic participation, but it is not identical to them. Bias and misclassification are often diagnosed relative to an existing label space. Flattening asks whether the label space itself has unjustified authority. Measurement error concerns whether an instrument measures a specified target consistently or validly. Flattening can occur even when measurement is consistent if the target suppresses alternative meanings. Proxy failure is one mechanism of flattening, but flattening also concerns who can contest a proxy and how dissent is recorded. Representational harm concerns damaging depiction or recognition; flattening can produce representational harm, but it can also occur through apparently neutral benchmarks or maps that suppress contestable categories.

Technical abstraction is not automatically flattening. AI systems require compression, categorization, and operationalization. Abstraction becomes flattening when it is imposed without procedural standing for affected actors, erases known disagreement, substitutes visible proxies for situated meanings, or remains unavailable to revision after deployment. Diagnostic indicators include fixed label spaces without schema rationale, erased neutral or abstention categories, aggregation without dissent records, unavailable category appeal, and absent revision triggers.

4. Method

This paper is a conceptual and qualitative synthesis. It asks where pluralism fails across the AI lifecycle, how participatory AI projects preserve or compress plural values and ontologies, and what institutional mechanisms make pluralism durable beyond model outputs. The unit of analysis is the AI lifecycle: problem framing, data generation, labeling, training, evaluation, deployment, maintenance, and retirement. The paper is not a systematic review, a new controlled experiment, or a validation study for a scoring instrument. Quantitative findings are reported as source-reported results from companion case manuscripts and are used to identify mechanisms, not to estimate population-level effects.

The concept map used purposive theoretical sampling. Seed literatures were selected because they provide concepts needed to analyze pluralism in AI lifecycles: value pluralism, pluralistic alignment, democratic AI, participatory AI, procedural justice, science and technology studies, design justice, algorithmic pluralism, supply-chain accountability,

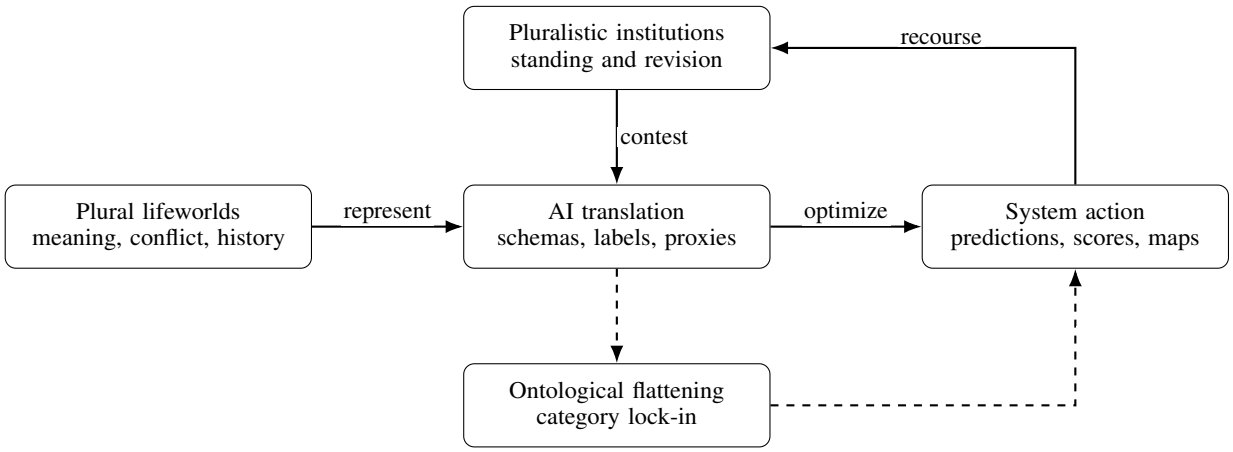

Figure 1. AI pluralism as resistance to ontological flattening. The issue is not only inaccurate representation, but the conversion of plural lifeworlds into singular technical ontologies.

Table 2. Boundary between ontological flattening and adjacent concepts.

Concept	**Primary diagnostic question**	**Relation to ontological flattening**
Bias or misclassification	Are outputs wrong or uneven relative to a label space?	Flattening asks whether the label space itself wrongfully fixes a contested world.
Measurement or construct validity failure	Does the measure capture the intended construct?	Flattening asks whether the construct should have been singularized and who had authority to define it.
Proxy failure	Does a measurable proxy diverge from the target?	Proxy failure can be a mechanism of flattening when the proxy replaces situated meaning and cannot be contested.
Representational harm	Does the system depict or recognize a group in damaging ways?	Flattening can cause such harm, but it also applies to neutral-looking schemas, maps, and benchmark targets.
Tokenistic participation	Are participants consulted without power?	Tokenism is a procedural mechanism through which flattening persists despite participation.
Legitimate abstraction	Is compression bounded, documented, revisable, and proportionate to the task?	Abstraction is not flattening when disagreement and revision rights are preserved and the output is not presented as a complete social description.

and documentation. Works were included when they addressed at least one construct relevant to the framework: object of pluralism, locus of authority, evidence standard, aggregation method, lifecycle stage, or failure mode. The map was used to identify whether a source treated pluralism primarily as a value, preference, output, population, ontology, institution, or procedure.

The synthesis also incorporates 11 confidential expert interviews. Seven initial semi-structured interviews with experts in law, ethics, and AI refined construct boundaries. Four additional interviews focused on PLG, auditability, institutional use, and external evaluation. The interviews are used as construct-refinement evidence, not as representational inference. The paper does not report individual quotations, frequencies, or transcript-level claims. Aggregate interview themes shaped three analytic decisions: procedural authority was separated from epistemic inclusion, lifecycle accountability was treated as distinct from documentation, and auditability was defined as evidence traceability rather than numerical scoring.

The three cases were selected because they cover different lifecycle stages in a bounded empirical domain. Case A centers data formation and preference optimization for public-space generation. Case B centers participatory labels, predictive modeling, and spatial inference for streetscape inclusivity. Case C centers reliability-aware evaluation for urban vision-language models. The cases are companion manuscripts. This paper therefore treats them as illustrative evidence for mechanisms of flattening rather than independently verified empirical validation. Where the case materials do not provide enough information to resolve an interpretation, the analysis records the claim as partial, source-reported, or unknown rather than treating it as established.

The case analysis used a five-code matrix: ontological entry point, disagreement handling, authority and recourse, proxy

Table 3. Evidence base and evidentiary status.

Source	Role	Material available to this synthesis	Evidence status
Concept map	Theoretical grounding	Pluralistic alignment, value pluralism, democratic AI, participatory AI, procedural justice, STS, accountability, documentation, and design justice.	Purposive map, not exhaustive review.
Expert interviews	Construct refinement	11 interviews: seven initial interviews plus four PLG-focused interviews.	Aggregate themes only; no frequency, prevalence, or transcript-level claims.
Case A	Generative alignment case	Two-year participatory process; 30 community organizations; 634 initial concepts; six consolidated criteria; 37,710 pairwise comparisons across 13,462 images.	Source-reported results; used for mechanism illustration.
Case B	Predictive modeling case	28 interviews; 12 focus-group raters; 20 streets; three data points per street; 60 rated locations; around 45,000 street images; 28 group-criterion outputs.	Source-reported performance; non-independence and split design remain interpretation limits.
Case C	Evaluation case	100 urban scenes; 12 participants from seven organizations; 230 forms; 30 dimensions; seven vision-language models.	Source-reported benchmark; small local sample; chance baselines vary by item.

limitation, and lifecycle continuation. Each case was examined across data and criteria, model and evaluation, and governance and deployment. The analysis separated four claim types: claims established by prior literature, aggregate interview-derived construct refinements, source-reported case results, and authorial inferences from the cross-case comparison. Construct validity is addressed through traceability across these claim types rather than through inter-rater statistics. PLG is presented as a qualitative framework and audit scaffold, not as a validated instrument with calibrated scores.

5. Empirical Synthesis

Case A: pluralistic generation. Case A asks whether community-defined, multi-criteria feedback can align text-to-image models for inclusive public spaces. Participants generated 634 initial concepts, consolidated through workshops into six locally defined criteria: Accessibility, Safety, Comfort, Invitingness, Inclusivity, and Diversity. The final dataset contains 37,710 pairwise comparisons across 13,462 images (Mushkani et al., 2025). The model was fine-tuned with Direct Preference Optimization (DPO), a method that trains directly from paired preferred and dispreferred outputs (Rafailov et al., 2023). In this case, the DPO objective encouraged transformation of multi-criteria judgments into pairwise preference targets through majority voting. The source study reports that the aligned model was favored more often than the baseline in heldout evaluation, while neutral comparisons remained common. This is not evidence that preference optimization fails. It illustrates that outcome gains can coexist with procedural compression. Neutral labels may indicate ambiguity, context-dependence, balanced trade-offs, insufficient visual evidence, or disagreement that should not be forced into a binary preference.

Case B: participatory prediction. Case B examines whether participatory urban perception labels can support prediction of streetscape inclusivity. The project recruited 28 interview participants and 12 focus-group raters. Participants evaluated 20 streets, with three data points per street, producing 60 rated locations. The training set expanded each location through street-level image frames, and the trained model was applied to around 45,000 street images (Mushkani & Koseki, 2026). The source study reports internal validation and test performance, but this paper does not use the magnitude of those results as evidence of generalization because frames from the same street, nearby vantage points, or participant rating contexts could introduce non-independence if splits are not grouped by street or location. The governance issue remains even when prediction performs well internally: the model can shift inclusivity toward visible aesthetics, while participants linked inclusivity to accessibility, activity, acceptance, safety, and belonging.

Case C: reliability-aware evaluation. Case C asks whether vision-language models classify urban scenes in ways that align with local human annotations. It curates 100 scenes and collects annotations from 12 participants across 30 dimensions. The evaluation compares seven models under a deterministic zero-shot prompt contract (Mushkani, 2025). The task includes single-choice and multi-label items. Accuracy is used for single-choice dimensions, Jaccard sim-

ilarity for multi-label dimensions, and macro averages for dimension-level summaries. Because option counts and label structures vary, the reported macro scores from 0.16 to 0.31 cannot be interpreted against one uniform chance baseline. The source study reports that dimensions closer to visible attributes were easier than appraisal dimensions and that model performance co-varied with human reliability. Negative Krippendorff's α in some dimensions indicates systematic disagreement beyond chance expectations, which means those dimensions should not be treated as stable ground truth for ordinary accuracy claims. Different uses of `Not applicable` by humans and models also made abstention semantics part of the measurement result.

The cases motivate, rather than validate, a lifecycle account of pluralism. Pluralism is most fragile where a lifeworld becomes a technical object. If a system begins with fixed labels, later participation can only fill in values for those labels. Disagreement and neutrality should therefore be treated as information rather than missingness. Model gains are conditioned by aggregation: binary preferences, averaged participant scores, majority labels, thresholds, and prompt contracts are governance decisions. Outcome pluralism and procedural pluralism must therefore be evaluated separately.

6. Pluralistic Lifecycle Governance

We propose Pluralistic Lifecycle Governance as a preliminary qualitative framework for implementing AI pluralism as a lifecycle property. *Ontological openness* means that affected actors can contest categories, proxies, labels, and metrics. *Epistemic inclusion* means that multiple forms of knowledge, including lived experience, are admissible evidence. *Procedural authority* means that participation includes decision rights, not only advisory input. *Evaluation pluralism* means that disagreement, abstention, neutrality, and reliability are reported rather than erased. *Lifecycle accountability* means that pluralistic commitments persist through deployment, monitoring, revision, and retirement.

PLG does not assign numerical scores in this paper. A minimal qualitative audit records each dimension as documented, partially documented, absent, or unknown. *Documented* means that artifacts directly demonstrate both process and authority. *Partially documented* means that participation or reporting exists but authority, revision, or evidence requirements remain incomplete. *Absent* means that expected artifacts are not present. *Unknown* means that the evidence available to the auditor is insufficient to classify the dimension. Disagreement among auditors should be recorded rather than averaged away. In institutional use, unresolved disagreement is itself evidence about governance ambiguity.

A PLG audit begins by defining the artifact and decision context: dataset, model, benchmark, procurement, or deployment. It then collects evidence on problem framing, data formation, label schemas, annotation instructions, aggregation rules, model objectives, evaluation reports, deployment plans, recourse procedures, and revision logs. Each dimension receives an evidence state with a supporting artifact and a limitation note. The audit report should distinguish documentation from authority. A model card may disclose a label schema, but it does not show that affected actors could revise it. A workshop may show epistemic inclusion, but it does not show procedural authority unless the workshop had binding consequences or response obligations.

Procedural authority is the dimension most likely to be overstated. Consultation gives participants an opportunity to speak. Co-design gives them influence over categories or workflows. Delegated authority gives them a defined decision right. Veto or pause rights allow them to stop a release, map publication, or use case under specified conditions. Appeal rights require a named decision-maker to respond to objections. Binding review requires that a decision cannot proceed until a defined review procedure is completed. PLG treats these as different evidence states because voice without response obligations can leave ontological flattening intact.

The cases illustrate diagnostic use. Case A documents community-defined concepts and criteria, but binary preference optimization leaves residual compression risk. A less compressive lifecycle design could retain criterion-specific preference heads, report neutral outcomes as a target distribution, or use multi-objective methods that preserve tradeoffs. Case B documents interviews and group-specific ratings, but heatmaps create risks of proxy substitution, neighborhood stigmatization, and decontextualized planning use. A pluralistic deployment would limit map resolution, display uncertainty, prohibit punitive uses, and give community partners authority over publication and revision. Case C reports reliability, abstention, and distributional mismatch, but consensus labels can still conceal unstable semantics. A pluralistic benchmark should disclose prompt contracts, aggregation rules, label definitions, abstention semantics, and dimensions that should not be ranked.

7. Discussion

The synthesis explains why pluralistic outputs are insufficient. A model can produce multiple answers, steer to different users, or match a target distribution while still reproducing a fixed ontology. It may answer planning questions from several perspectives while assuming that accessibility is a stable visual property. It may generate diverse public-space images while using a training objective that collapses community disagreement. It may score well on a benchmark whose label space affected communities never

Table 4. Cross-case mechanisms and evidentiary status.

Mechanism	Case basis	Flattening risk	Procedural remedy
Category formation	Cases A and B began with participant concepts before consolidating criteria; Case C fixed a benchmark schema before model evaluation.	Later participation can only fill a pre-existing ontology.	Require public schema rationales, alternative label spaces, and category revision rights.
Disagreement and neutrality	Case A preserved neutral comparisons; Case B used group-specific ratings; Case C reported reliability and abstention.	Majority labels and averaged scores can hide ambiguity, low reliability, and minority meanings.	Report distributions, neutral rates, abstentions, reliability, aggregation sensitivity, and minority reports where appropriate.
Visual proxy limits	Images support spatial cues but poorly capture lived accessibility, acceptance, cultural meaning, maintenance, and belonging.	Visible form substitutes for situated meaning.	Combine visual outputs with field audits, testimony, maintenance records, and community review.
Lifecycle continuation	Cases document participatory inputs, but long-term authority over publication, revision, and use remains partial or unknown in the synthesis evidence.	Pluralism disappears after data collection or benchmark construction.	Specify audit cadence, recourse logs, review triggers, prohibited uses, revision rights, and decommissioning criteria.

authorized. Such systems are pluralistic at the output layer but monistic at the ontological layer.

Pluralism also does not require institutional paralysis. Public decisions still need to be made. Infrastructure must be built, models must be evaluated, and institutions must allocate resources. The alternative to monism is provisional consensus: decisions whose legitimacy rests on contestable procedures, public reasons, recorded dissent, and revision paths. A pluralistic AI lifecycle should preserve disagreement in the record, explain aggregation choices, and create revision triggers when deployed outputs conflict with situated experience.

Pluralism has boundaries. Some asserted ontologies deny equal standing to others, erase legally protected rights, enable targeted harm, or convert participation into harassment. A pluralistic lifecycle therefore needs boundary-setting procedures that are public, rights-respecting, and contestable. Boundary decisions should identify the excluded claim, state the rights or safety rationale, document the decision authority, preserve dissent when safe, and provide an appeal or revision path. In the present domain, a public-space model should not encode notions of safety that equate marginalized presence with risk, a generative dataset should not reward stereotypes of inclusion, and a benchmark should not treat contested social inference from pixels as a stable property of a scene.

For alignment research, the synthesis suggests three design implications. Alignment data should be treated as institutional data because labels record who had authority to define what mattered. Dataset documentation should therefore report category origin, disagreement handling, abstention semantics, and revision rights, not only demographics and collection protocols. Alignment objectives should support multi-dimensional and neutral feedback when the domain contains known trade-offs rather than treating all pluralism as a ranking problem. Benchmark scores should be reliability-aware because low human reliability can make ordinary accuracy interpretation misleading.

For governance, the urban vision cases suggest that pluralism must be tied to rights and institutions. Affected actors need standing to contest categories and not only to appeal outputs. Procurement for public-sector AI should require evidence of community involvement in problem framing and evaluation when systems operate on contested social concepts. Deployments should include revision procedures, audit cadences, recourse pathways, prohibited uses, and decommissioning triggers. These requirements must be proportionate. Participation can become burdensome when communities are asked to supply unpaid labor, when accessibility supports are absent, or when consultation fatigue accumulates. PLG therefore treats compensation, accessibility, multilingual materials, bounded decision scope, and response obligations as governance requirements rather than optional engagement practices.

8. Limitations and Future Work

The study has several limits. The empirical cases are concentrated in urban and public-space domains and rely heavily on vision-based systems. This focus makes ontological conflict observable, but it limits empirical breadth. Healthcare, education, labor, social services, and criminal justice may reveal different mechanisms. The cases were selected for thematic relevance and artifact availability rather than by an independent sampling frame. This creates a risk of confirmatory case selection, which is why the paper treats them

Table 5. Operationalizing Pluralistic Lifecycle Governance as a qualitative audit scaffold.

Dimension	Diagnostic question	Sufficient evidence	Insufficiency indicator
Ontological openness	Can affected actors contest categories, labels, proxies, metrics, and problem definitions?	Schema rationale, contested-category log, alternatives considered, revision path, and response record.	Fixed label space with no rationale, appeal path, or record of rejected alternatives.
Epistemic inclusion	Which forms of knowledge count as evidence?	Recruitment rationale, compensated participation, accessibility supports, multilingual materials where needed, and inclusion of lived, technical, and institutional knowledge.	Expert-only framing, inaccessible participation, or reliance on visible proxies as complete evidence.
Procedural authority	Do participants have power or only voice?	Documented authority to co-define, pause, veto, escalate, appeal, or require response from a named decision-maker.	Consultation after core decisions are fixed or no obligation to respond to participant objections.
Evaluation pluralism	Are disagreement, neutrality, abstention, and low reliability preserved?	Distributional reporting, neutral and abstention rates, reliability, aggregation sensitivity, subgroup analysis, and minority reports where safe.	Single aggregate score or majority label that hides instability and dissent.
Lifecycle accountability	Who maintains pluralism after deployment or release?	Audit cadence, recourse logs, incident reporting, revision triggers, prohibited uses, versioning, and decommissioning criteria.	Responsibility ends at model release, publication, or procurement.

as analytic anchors rather than validation evidence. Expert interviews are reported only in aggregate and are used for construct refinement, not prevalence claims.

The synthesis reuses results from companion projects rather than running a new controlled experiment across all lifecycle stages. The present manuscript does not provide transcript-level interview evidence, raw annotation data, full model training logs, or independent reanalysis of the companion case results. These materials would be needed for a stronger empirical paper. The current paper therefore separates source-reported case results from authorial inferences and avoids treating the cases as independent proof that PLG improves outcomes.

PLG is not yet validated as an evaluation instrument. It has no demonstrated inter-rater reliability, calibrated numerical scoring rule, or evidence that its use improves governance outcomes. Future work should apply PLG prospectively to external AI systems with independent raters, recorded disagreements, adjudication procedures, and outcome tracking. Additional work should test whether PLG changes procurement decisions, documentation quality, recourse use, or deployment revision in practice. Pluralistic procedures also require resources for workshops, accessibility supports, compensation, translation, revision, and maintenance.

9. Conclusion

AI pluralism is not achieved by producing many outputs, supporting many users, or collecting many preferences. Those steps matter, but they can leave intact a deeper problem: AI systems often impose one way of world-making over many others. The cases illustrate that pluralism is stronger when affected communities help define criteria, when disagreement and neutrality remain visible, when evaluation reports reliability and abstention, and when institutions allow revision. It is weakened when situated meanings are flattened into binary labels, visible proxies substitute for lived experience, or benchmark scores conceal unstable targets. This paper introduced ontological flattening as a failure mode for AI pluralism and proposed Pluralistic Lifecycle Governance as a qualitative framework for making ontological contestability an explicit object of AI governance.

References

Anderson, E. *Value in Ethics and Economics*. Harvard University Press, 1993.

Arnstein, S. R. A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4):216–224, 1969. doi: 10.1080/01944366908977225.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., Das-

Sarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. A general language assistant as a laboratory for alignment. 2021. doi: 10.48550/arXiv.2112.00861. URL https://arxiv.org/abs/2112.00861.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI feedback. 2022. doi: 10.48550/arXiv.2212.08073. URL https://arxiv.org/abs/2212.08073.

Bergman, S., Marchal, N., Mellor, J., Mohamed, S., Gabriel, I., and Isaac, W. Stela: A community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616, 2024. doi: 10.1038/s41598-024-56648-4.

Berlin, I. Two concepts of liberty. In *Four Essays on Liberty*, pp. 118–172. Oxford University Press, Oxford, 1969.

Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., and Mohamed, S. Power to the people? opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pp. 1–8, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3551624.3555290.

Bowker, G. C. and Star, S. L. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999. doi: 10.7551/mitpress/6352.001.0001.

Chang, R. Value incomparability and incommensurability. In Hirose, I. and Olson, J. (eds.), *The Oxford Handbook of Value Theory*, pp. 205–224. Oxford University Press, 2015.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Citron, D. K. Technological due process. *Washington University Law Review*, 85(6):1249–1313, 2008.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Position: Social choice should guide ai alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360. PMLR, 2024.

Costanza-Chock, S. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, Cambridge, MA, 2020. ISBN 9780262043458.

Delgado, F., Yang, S., Madaio, M., and 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*, 2023. doi: 10.1145/3617694.3623261.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020. doi: 10.1007/s11023-020-09539-2.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Habermas, J. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Beacon Press, Boston, 1984.

Habermas, J. *The Theory of Communicative Action, Volume 2: Lifeworld and System: A Critique of Functionalist Reason*. Beacon Press, Boston, 1987.

Haraway, D. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599, 1988. doi: 10.2307/3178066.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1395–1417, 2024. doi: 10.1145/3630106.3658979.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Kasirzadeh, A. Plurality of value pluralism and ai value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024.

Kasirzadeh, A. and Gabriel, I. In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, 36(2):27, 2023. doi: 10.1007/s13347-023-00606-x.

Kinchin, N. Voiceless: The procedural gap in algorithmic justice. *International Journal of Law and Information Technology*, 32:eaae024, 2024. doi: 10.1093/ijlit/eaae024.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344, 2024. doi: 10.52202/079017-3342. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be2e1b68b44f2419e19f6c35a1b8cf35-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596.

Mol, A. *The Body Multiple: Ontology in Medical Practice*. Duke University Press, 2002. doi: 10.1515/9780822384151.

Mushkani, R. Do vision-language models see urban scenes as people do? an urban perception benchmark, 2025. URL https://arxiv.org/abs/2509.14574.

Mushkani, R. and Koseki, S. Street review: A participatory ai-based framework for assessing streetscape inclusivity. *Cities*, 170:106602, 2026. doi: 10.1016/j.cities.2025.106602. URL https://doi.org/10.1016/j.cities.2025.106602.

Mushkani, R., Nayak, S., Berard, H., Cohen, A., Koseki, S., and Bertrand, H. LIVS: A pluralistic alignment dataset for inclusive public spaces. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45311–45341. PMLR, July 2025. URL https://proceedings.mlr.press/v267/mushkani25a.html.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. doi: 10.48550/arXiv.2203.02155.

Ovadya, A., Redman, K., Thorburn, L., Chen, Q. Z., Smith, O., Devine, F., Konya, A., Milli, S., Revel, M., Feng, K. J. K., Zhang, A. X., Chandra, B., Bakker, M. A., and Kasirzadeh, A. Position: Democratic ai is possible. the democracy levels framework shows how it might work. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81930–81961. PMLR, 2025.

Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, 2019. doi: 10.1145/3287560.3287598.

Sloane, M., Moss, E., Awomolo, O., and Forlano, L. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–6, 2022. doi: 10.1145/3551624.3555285.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024.

Suchman, L. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press, 2007.

Suresh, H. and Guttag, J. V. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9, 2021. doi: 10.1145/3465416.3483305.

Widder, D. G. and Nafus, D. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1):1–12, 2023. doi: 10.1177/20539517231177620.