Co-Producing AI: Toward an Augmented, Participatory Lifecycle

Rashid Mushkani[1,2*], **Hugo Berard**[1], **Toumadher Ammar**[1], **Cassandre Chatonnier**[1], **Shin Koseki**[1,2]

[1]Université de Montréal
[2]Mila–Quebec AI Institute
{rashid.ahmad.mushkani, hugo.berard, toumadher.ammar, cassandre.chatonnier, shin.koseki}@umontreal.ca

Abstract

Despite efforts to mitigate the inherent risks and biases of artificial intelligence (AI) algorithms, these algorithms can disproportionately impact culturally marginalized groups. A range of approaches has been proposed to address or reduce these risks, including the development of ethical guidelines and principles for responsible AI, as well as technical solutions that promote algorithmic fairness. Drawing on design justice, expansive learning theory, and recent empirical work on participatory AI, we argue that mitigating these harms requires a fundamental re-architecture of the AI production pipeline. This re-design should center co-production, diversity, equity, inclusion (DEI), and multidisciplinary collaboration. We introduce an augmented AI lifecycle consisting of five interconnected phases: co-framing, co-design, co-implementation, co-deployment, and co-maintenance. The lifecycle is informed by four multidisciplinary workshops and grounded in themes of distributed authority and iterative knowledge exchange. Finally, we relate the proposed lifecycle to several leading ethical frameworks and outline key research questions that remain for scaling participatory governance.

Introduction

AI systems are increasingly embedded in decision-critical infrastructures, including law-enforcement identification (Dwivedi et al. 2021; Zhao et al. 2019), medical image interpretation, patient-outcome prediction, drug discovery (Yu, Beam, and Kohane 2018), and personalized recommendation. The breadth and speed of these deployments create substantial societal leverage.

Empirical evidence indicates that this leverage can intensify existing inequities. Documented harms include predictive-policing bias (Angwin et al. 2016), disparate facial-recognition error rates for people of color (Buolamwini and & Gebru 2018), and discriminatory screening in automated hiring systems (Dastin 2022). These risks are compounded by limited model transparency (Burrell 2016), privacy violations arising from data aggregation (Véliz 2021; Nissenbaum 2010), and a broad range of governance challenges cataloged by prior work (Arslan 2017; Brayne 2017; Galaz et al. 2021; Gerdes 2018; Golpayegani, Pandit, and & Lewis 2022; Koseki et al. 2022; Leslie 2019; Mohamed, Png, and Isaac 2020; Alim and Adebayo 2022; Mushkani et al. 2025a; Sorensen et al. 2024).

Policymakers and professional bodies have responded with high-level guidance. The Montreal Declaration for Responsible AI (Université de Montréal 2018), the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019), the European Commission's AI Act (European Commission 2018), and the NIST AI Risk Management Framework collectively articulate principles such as autonomy, equity, and accountability (NIST 2023). Yet these instruments remain largely aspirational and center responsibility within expert or organizational hierarchies rather than within affected publics.

Concurrently, technical research has introduced concepts of bias, fairness, and explainability (Arrieta et al. 2020; Miller 2019). Numerous fairness criteria, such as statistical parity, equal opportunity, and individual fairness, coexist and sometimes conflict (Bellamy et al. 2019; Chouldechova 2017; Dwork et al. 2012; Hardt, Price, and Srebro 2016; Binns 2020; Hoffmann et al. 2022; Lepri et al. 2018). The multiplicity of metrics, combined with the socio-technical nature of bias (Korobenko, Nikiforova, and Sharma 2024), complicates operationalization in practice.

Recent scholarship underscores the cultural contingency of ethical criteria. Studies on Indian art traditions (Divakaran, Sridhar, and Srinivasan 2023) and non-Western honor systems (Wu, Demetriou, and Husain 2023) demonstrate that prevailing frameworks often mirror Western, educated, industrialized, rich, and democratic settings. Cross-cultural surveys reveal divergent alignment preferences (Kirk et al. 2024; Mushkani et al. 2025b), and pluralistic-alignment proposals advocate participatory resolution of normative conflict (Sorensen et al. 2024; Anthropic 2023).

Sector-specific investigations likewise show that ethical challenges vary by context. Domain studies expose distinctive issues in agriculture (Burema et al. 2023) and human–computer interaction (Li and Lu 2022), while practitioner reflections identify organizational power dynamics that limit the utility of existing ethics toolkits (Wong, Madaio, and Merrill 2023). These findings suggest that a checklist-based ethics approach is insufficient without con-

*Corresponding author

tinuous stakeholder engagement.

Participatory and co-design traditions offer a partial remedy. Case studies demonstrate improvements in logistics optimization (Berditchevskaia, Malliaraki, and & Peach 2021), clinical decision support (Zicari et al. 2021), and public-sector recommender systems (Donia and Shaw 2021). Nevertheless, most participatory efforts involve brief or single-phase engagements (Gerdes 2022; Birhane et al. 2022), and large technology firms still tend to design *for* rather than *with* users (Aizenberg and & van den Hoven 2020). This gap motivates our research question: *How can citizen participation be integrated throughout the AI lifecycle—balancing process-oriented and outcome-oriented considerations—to produce systems that are both effective and just?*

To address this question, we draw on expansive-learning theory (Engeström and Sannino 2010) and design-justice principles (Costanza-Chock 2020). Building on multidisciplinary workshops (Lember, Brandsen, and Tõnurist 2019; McBride et al. 2023), we propose an augmented AI lifecycle comprising five interconnected phases—co-framing, co-design, co-implementation, co-deployment, and co-maintenance. Each phase positions citizens, domain experts, and technologists as co-producers, thereby operationalizing DEI commitments while fostering iterative knowledge exchange.

Contributions

This study makes three key contributions:

1. **Lifecycle proposal:** Synthesizes design-justice and expansive-learning perspectives into a five-phase AI lifecycle that meaningfully includes citizens as co-producers.
2. **Ethical mapping:** Maps this lifecycle to prevailing ethical frameworks, highlighting where they converge and diverge from participatory practice.
3. **Workshop synthesis:** Distills a shared understanding from four multidisciplinary workshops, of how participatory checkpoints can shape technical and governance choices across the lifecycle.

The remainder of the paper details this lifecycle, relates it to leading ethical frameworks, and outlines future research directions for scalable participatory governance.

Related literature

Lifecycle models across domains

Product development research identifies three generic phases—beginning-of-life, middle-of-life, and end-of-life—through which artefacts progress (Kiritsis, Bufardi, and Xirouchakis 2003). Early craft models placed all lifecycle tasks in a single agent, as illustrated by the cobbler metaphor, whereas contemporary industrial workflows distribute responsibilities across specialized actors while seeking continuity of data and knowledge (Ameri and Dutta 2005; Ibrahim and Paulson 2008; Terzi et al. 2010). Software engineering codified similar concerns in the software development lifecycle (SDLC), a sequence of planning, analysis, design, implementation, testing, deployment, and maintenance that evolved from the PDCA cycle and agile methods (Deming 1986; Beck et al. 2001; Mohammed et al. 2017; Assal and Chiasson 2018). Machine-learning pipelines inherit these stages but foreground data collection, preprocessing, and algorithm selection. Canonical models present a linear path from problem formulation to deployment, yet in practice, training outcomes routinely prompt returns to prior stages, and pre-trained models introduce additional loops devoted to transfer learning and fine-tuning (Hummer et al. 2019; Haakman et al. 2021; Sculley et al. 2015; Qian, Lee, and Schwanen 2021; Wang et al. 2020; De Silva and Alahakoon 2022). Across domains, effective lifecycle management depends on the unobstructed transfer of artefacts, assumptions, and performance evidence between stages.

Participation in the AI lifecycle

Participatory design emerged in Scandinavian labor contexts and has since influenced a wide range of socio-technical projects (Asaro 2000; Birhane et al. 2022). Contemporary methods such as workshops, role-playing, scenario building, prototyping, and other dialogic practices aim to build consensus among diverse stakeholders as they navigate complex and often contested design problems (Flanagan and Nissenbaum 2016; Sloane et al. 2022; Demelenne et al. 2020).

In the field of AI, participatory approaches have been employed in initiatives such as Collective Constitutional AI (Huang et al. 2024), the PRISM Alignment Dataset (Kirk et al. 2024), and MID-Space (Nayak et al. 2024), each of which integrates stakeholder input into the development and evaluation of AI systems. Empirical projects like Project Dorian (Berditchevskaia, Peach, and Malliaraki 2021) and We-BuildAI (Lee et al. 2019) demonstrate the practical value of iterative feedback loops, showing how repeated engagement can shape system outcomes in meaningful ways.

Studies further suggest that involving frontline users in the design and development process can improve task performance and uncover risks that may not be visible to developers alone. This is particularly evident in domains such as logistics, agriculture, public-sector recommender systems, and clinical decision support (Berditchevskaia, Malliaraki, and & Peach 2021; Zicari et al. 2021; Donia and Shaw 2021).

Despite these advances, most participatory efforts remain limited in scope. Often, stakeholders are consulted during a single phase or brought in for short-term input, while critical aspects such as problem definition, deployment, and post-deployment governance remain in the hands of experts (Gerdes 2022; Aizenberg and & van den Hoven 2020; Gerdes 2018; Wang, Ramaswamy, and Russakovsky 2022; Ravanera and Kaplan 2021). This restricted engagement reduces opportunities for meaningful knowledge exchange and limits accountability to the communities most affected by AI systems (Helbing et al. 2023; De Silva and Alahakoon 2022).

Diversity, equity, and inclusion within AI practice

DEI frameworks hold that individual experiences and identities mediate access to resources and shape technology adop-

Risk Category	Description
Social Responsibility	Ethical duty to consider AI's societal impacts and promote public benefit.
Vulnerable Populations	Risk of exacerbating harm to marginalized or disadvantaged groups.
Transparency	Need for clear, understandable AI processes and decisions.
Misuse & Hostile Use	Potential for malicious or unethical applications of AI.
Human Rights	Threats to privacy, expression, and freedom from discrimination.
Deception & Manipulation	Use of AI to mislead or covertly influence behavior.
Inequalities	Risk of deepening social and economic divides.
Workforce Diversity	Importance of inclusive development to avoid narrow perspectives.
AGI & Existential Risk	Long-term concerns over uncontrolled advanced AI.
Cultural Sensitivity	Need to respect diverse norms in AI design and use.
Trust	Building confidence in AI's reliability and ethical use.
Psychological Impacts	Effects on mental health, identity, and human purpose.
Unemployment	Job loss risks from automation-driven displacement.
Misinformation	Generation or spread of false information via AI.
Economic Growth	AI's role in boosting innovation and productivity.
Explainability	Clarity on how AI makes decisions.
Privacy	Safeguarding personal data from misuse or leaks.
Safety & Reliability	Ensuring systems function correctly and safely.
Human Control	Maintaining human oversight of AI actions.
Accountability	Assigning responsibility for AI outcomes.
Consent & Autonomy	Respecting individuals' control over data and impact.

Table 1: Risks and ethical considerations in AI development and use

tion (Ashley et al. 2022; Calabrese Barton and & Tan 2020; de Hond, van Buchem, and Hernandez-Boussard 2022). Reviews of AI deployments reveal persistent bias, limited demographic representation, and uneven privacy impacts, prompting calls for process interventions grounded in DEI principles (Cachat-Rosset and & Klarsfeld 2023; Forum 2022). Public, private, and civil-society actors espouse distinct ethical priorities; government and non-profit bodies emphasize broad societal effects and participatory governance, whereas corporate initiatives often concentrate on product-specific risk mitigation (Schiff et al. 2021a). Multidisciplinary collaboration offers one response, aligning technical expertise with normative, legal, and social analysis, yet empirical work documents persistent disciplinary silos and communication barriers (Bikakis and & Zheng 2015; Bisconti et al. 2023; Dwivedi et al. 2021; Beaudouin et al. 2020).

Rationale for an augmented AI lifecycle

The "empirical turn" in technology ethics reframed design as a moral practice, inspiring approaches such as value-sensitive design and ethics-by-design (Umbrello and van de Poel 2021; Gerdes and & Frandsen 2023). Responsible-AI principles (fairness, accountability, transparency) are now reflected in policy instruments and corporate standards, but operational guidance remains coarse. Competing mathematical definitions of fairness complicate implementation, and high-level frameworks rarely specify mechanisms for citizen participation throughout system development (Barocas, Hardt, and Narayanan 2023; Jobin, Ienca, and Vayena 2019; Université de Montréal 2018; European Commission 2018; NIST 2023). Comparative analyses show that broad ethical coverage, adaptability across contexts, and iterative governance are necessary but insufficient without practical procedures and documentation that enable stakeholder input (Schiff et al. 2021b). Catalogues of harms (discrimination, misinformation, privacy erosion, and others) compile evidence of risks that exceed the remit of traditional lifecycles (Arslan 2017; Brayne 2017; Galaz et al. 2021; Gerdes 2018; Golpayegani, Pandit, and & Lewis 2022; Koseki et al. 2022; Mushkani, Berard, and Koseki 2025; Mohamed, Png, and Isaac 2020; Alim and Adebayo 2022; Gowaikar et al. 2024). Table 1 consolidates these issues. Expansive-learning theory posits that durable solutions emerge when intersecting communities collectively reconstruct their activity systems, while design-justice scholarship centers those most affected by design outcomes (Roth 2004; Engeström 2014; Jordan 2023). Together, these perspectives motivate the augmented lifecycle proposed in this work, which embeds co-production, DEI, and iterative knowledge exchange across all phases to address the limitations identified above.

Materials and Methods

Literature Review

Between October 2023 and May 2024, we conducted a scoping review of academic and gray literature published from January 2013 to May 2024. Our goal was to identify concepts, methods, and empirical findings relevant to ethical and inclusive AI to inform a series of multidisciplinary workshops. We searched Scopus, PubMed, Web of Science, and Google Scholar, covering fields across computer science, the social sciences, and the humanities. We applied the following Boolean expression to titles, abstracts, and keywords:

```
  ("ethical AI" OR "AI ethics")
AND (fairness OR transparency OR
accountability OR bias OR "algorithmic
fairness" OR "explainable AI" OR
"participatory AI" OR "public
participation" OR co-production OR
co-creation OR "AI ethics guidelines"
OR "AI ethics frameworks" OR
"human-in-the-loop" OR "responsible
innovation" OR "ethical dilemmas"
```

```
OR "AI for good" OR "value-sensitive
design" OR "ethics by design" OR "AI
lifecycle" OR "product lifecycle"
OR inclusivity OR autonomy OR
interpretability).
```

We retained only English-language records. The initial search returned 330 documents. After removing duplicates and screening titles and abstracts for relevance to AI ethics, inclusivity, and lifecycle processes, we narrowed the pool to 147 records. To include a source at the full-text stage, it had to explicitly address (i) ethics and AI, (ii) co-creation or co-production, (iii) lifecycle or process models, and (iv) design-justice perspectives. We also added industrial standards and policy frameworks through targeted searches of organizational repositories. In total, we synthesized 76 sources to prepare the workshop materials.

Workshop Design

Between January and May 2024, we conducted four three-hour workshops in Montréal, Canada. Recruitment followed a purposive sampling strategy designed to ensure disciplinary breadth and institutional diversity. Across the four sessions, we enrolled twenty participants (5–9 per workshop). Participants were affiliated with diverse organizations, including Mila–Quebec AI Institute, Université de Montréal, the Institute for Data Valorization (IVADO), *Institut national de la recherche scientifique* (INRS), and the International Observatory on the Societal Impacts of AI and Digital Technology (OBVIA). Disciplinary backgrounds spanned computer science, social science, law, philosophy, and diversity–equity–inclusion practice. Nine participants identified primarily as researchers, six as industry practitioners, and five as civil-society advocates.

Each workshop began with a summary of key findings from the literature review, followed by a moderated discussion based on three core questions:

1. How can the ethical challenges identified in the literature be addressed in the design and use of AI systems?
2. Which methods have demonstrated efficacy in producing ethical AI?
3. Which methodological scenarios can guide future ethical AI development?

We recorded audio and took detailed notes, then transcribed and anonymized the data for analysis.

Data Analysis

We used an inductive–deductive approach to code the transcripts thematically. We began with codes informed by expansive-learning and design-justice theory, then added new categories as they emerged until we reached saturation. We resolved coding disagreements through peer debriefing. Once we finalized the codebook, we applied it across all workshop data. The resulting themes form the basis of the augmented AI lifecycle described later in the paper.

Theoretical Orientation

Our analysis draws on expansive-learning and activity theory, which treat learning as a collective transformation of activity systems and boundary crossing between communities (Engeström and Sannino 2010; Engeström 2014). We also apply design-justice principles, which center decision-making authority with those most affected by technological outcomes (Costanza-Chock 2020; Jordan 2023). By combining these frameworks, we treat co-production as a normative requirement rather than an optional supplement (Aizenberg and & van den Hoven 2020; Helbing et al. 2023).

Synthesis Procedure

We triangulated insights from the literature review and workshop data to develop an augmented AI lifecycle with five stages: co-framing, co-design, co-implementation, co-deployment, and co-maintenance. We iterated on this model with workshop participants to ensure it reflected both their contributions and the study's theoretical foundations.

Results

Thematic analysis of the four workshops produced four interdependent themes that shaped the design of the augmented AI lifecycle. Each theme is grounded in direct participant testimony and reflects recurring patterns observed across sessions.

Distributed authority. Participants consistently argued that decision rights should reside with the communities that will bear the consequences of an AI system. One attendee stated, "Decision-making power has to move closer to the communities who will live with the outcomes; otherwise the system will reproduce existing hierarchies" [Workshop 2, P5]. This view aligns with design-justice scholarship and informed the decision to embed community veto rights and shared governance checkpoints throughout the lifecycle.

Iterative knowledge exchange. The workshops highlighted the importance of repeated, dialogic learning. A participant explained that lay contributors acquire AI literacy "through participating in such learning circles where people from different background show up" [Workshop 3, P2]. Consequently, the lifecycle specifies cyclic feedback mechanisms—such as community review sessions and shared artifact repositories—that maintain context as teams and project phases evolve.

Contextual privacy. Privacy practices were viewed as inseparable from local social norms. As noted by one participant, "The privacy solution must match local norms; differential privacy alone does not answer the cultural questions" [Workshop 1, P4]. This observation supports a layered privacy strategy calibrated to cultural expectations and data sensitivity, extending contextual-integrity arguments (Nissenbaum 2010).

Resource constraints. Sustained engagement was deemed feasible only when budgets addressed tangible costs borne by community members. One participant observed, "We can sustain monthly check-ins only if the

budget covers childcare and travel for community members" [Workshop 4, P7]. To ensure meaningful involvement rather than tokenistic participation, the engagement lifecycle must include dedicated funding for logistical support and clearly defined role charters (Sloane et al. 2022).

These empirically grounded themes directly informed the tasks, artifacts, and checkpoints detailed in Section .

The Augmented AI Lifecycle

Grounded in design-justice principles (Costanza-Chock 2020), expansive-learning theory (Engeström 2014), and DEI scholarship (de Hond, van Buchem, and Hernandez-Boussard 2022; Calabrese Barton and & Tan 2020), the augmented AI lifecycle operationalizes co-production across five interdependent phases: *co-framing*, *co-design*, *co-implementation*, *co-deployment*, and *co-maintenance*. Each phase integrates citizens, domain specialists, and technologists as joint decision-makers, thereby addressing gaps identified in conventional lifecycles (Section Rationale for an augmented AI lifecycle). Figure 1 visualizes the overall process, while Figure 2 maps design versus co-design phase-specific risks and mitigation strategies derived from our workshops and prior studies (De Silva and Alahakoon 2022; Mitchell et al. 2019).

Co-framing

Co-framing establishes a shared problem definition, an initial risk register, and a participation plan. Drawing on Arnstein's ladder of participation (Arnstein 1969) and the boundary-crossing mechanisms of expansive learning (Engeström and Sannino 2010), the project team convenes citizens likely to experience the system's outcomes, alongside domain and DEI experts. Structured workshops, semi-structured interviews, and deliberative forums surface contextual knowledge often absent from expert-centric scoping (Koseki et al. 2022; Haakman et al. 2021). Key questions include:

1. Which communities and activity systems are affected, and how are impacts geographically distributed?
2. What comparable systems exist, and what ethical failures have been documented?
3. Which engagement methods (e.g., open calls, community partnerships) align with project constraints and DEI objectives?
4. How will citizen perspectives reshape the initial problem statement?

Co-design

In this phase, participants select data sources, model families, and interface concepts aligned with co-framed objectives. Participatory prototyping sessions and scenario walkthroughs translate workshop insights into technical specifications (Demelenne et al. 2020; Sloane et al. 2022). Comparative risk assessments evaluate trade-offs among pipeline options, clarifying how choices distribute benefits and burdens (Berditchevskaia, Malliaraki, and & Peach 2021; Zicari et al. 2021). Guiding questions include:

1. How do data-collection strategies affect representation and privacy?
2. Which foundational models, if any, meet performance requirements without compromising interpretability or fairness?
3. How does the overall pipeline architecture support autonomy and privacy while minimizing environmental impact?
4. What documentation makes design decisions accessible to non-technical stakeholders?
5. How will participatory feedback be incorporated when reusing pre-trained components?

Co-implementation

This phase includes data acquisition, feature engineering, model training/fine-tuning, and iterative validation. A multidisciplinary team records decisions in version-controlled artifacts and model cards to ensure traceability to co-design deliberations (Mitchell et al. 2019). Citizen partners review intermediate outputs, such as data summaries and error reports, to check conformance with DEI commitments (Cachat-Rosset and & Klarsfeld 2023). Guiding questions include:

1. Does the implementation team reflect heterogeneous expertise and lived experience?
2. How are privacy safeguards (e.g., differential privacy, k-anonymity) applied and communicated (Pawar, Ahirrao, and Churi 2018; Véliz 2021)?
3. Which artifacts (code, configuration, audit trails) are required for downstream accountability?
4. How is knowledge transferred to citizen partners to support informed oversight?

Co-deployment

Deployment introduces the system into non-stationary social contexts where latent risks may emerge (Koseki et al. 2022). Citizen representatives conduct user-acceptance testing with domain experts, while a governance charter assigns responsibility for emergent harms. Performance dashboards and recourse mechanisms enable affected individuals to contest outputs (Galaz et al. 2021). Central questions include:

1. Have evaluation metrics captured distributional impacts across stakeholder groups?
2. What safeguards prevent mission creep and unauthorized secondary use?
3. How long is the model considered valid, and what triggers rollback or redesign?
4. What transparency measures communicate real-time system behavior to the public?

Co-maintenance

Co-maintenance treats the system as a living artifact subject to concept drift, regulatory change, and evolving norms (De Silva and Alahakoon 2022). Periodic audits (technical, ethical, and participatory) evaluate alignment with original

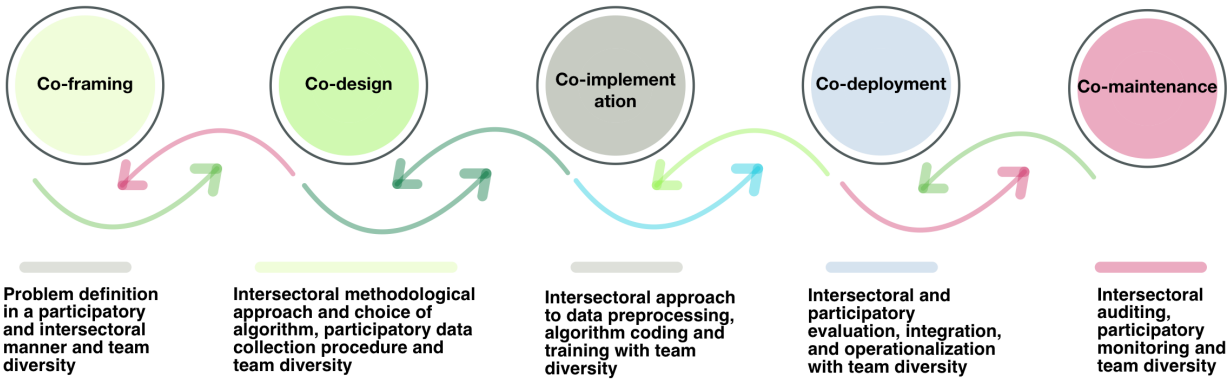

Figure 1: The augmented AI lifecycle. Five co-production phases are linked by continuous knowledge exchange and shared accountability.

objectives. Citizen assemblies or standing committees may recommend updates, suspension, or decommissioning (Helbing et al. 2023). Key questions include:

1. What cadence and scope govern multidisciplinary audits, and who funds them?
2. How are updates communicated, and how is consent re-established when functionality changes?
3. What dispute-resolution processes address tensions between commercial aims and public interest?
4. How is institutional memory preserved as team membership changes?

Cross-cutting Considerations

Across all phases, successful co-production depends on (i) recruitment practices that meet DEI goals (Calabrese Barton and & Tan 2020), (ii) anonymization protocols that uphold contextual integrity (Nissenbaum 2010; Véliz 2021), and (iii) facilitation methods that reduce power asymmetries (Turnhout, Van Bommel, and Aarts 2010; Jordan 2023). Our workshop participants shared a common understanding that participatory practices are feasible within typical resource constraints, provided that roles, incentives, and communication channels are clearly defined from the outset.

Overall, the augmented lifecycle embeds shared accountability and iterative learning throughout AI development, aligning empirical methods with the normative goals of policy frameworks and design-justice scholarship. The following sections assess its applicability across domains and propose avenues for scalable participatory governance.

Discussion

This discussion interprets the findings of the workshops using the conceptual frameworks of design justice, expansive learning, and DEI. These lenses help situate how the proposed lifecycle addresses the methodological and normative commitments surfaced during the study. Rather than suggesting deterministic outcomes, the discussion outlines contingent pathways through which co-production may support negotiated responses to AI-related harms.

Engaging Participants Through Design-Justice Recruitment

Design-justice literature positions affected communities as essential actors in shaping technology. Workshop discussions highlighted the importance of beginning engagement with the identification of three domains: *area of impact*, *area of interest*, and *project constraints*. Participants described this as a prerequisite for identifying appropriate community stakeholders and clarifying expectations among developers. They further emphasized that early transparency regarding project scope and resource constraints can mitigate asymmetries in authority (Helbing et al. 2023; Mushkani et al. 2025a). Several recruitment practices, such as open calls, partnerships with community organizations, and collaboration with labor unions, were discussed as potentially viable, contingent on alignment with DEI goals and access to compensation mechanisms.

Participation Modalities as Expansive-Learning Interventions

Expansive-learning theory conceptualizes participation as an encounter between intersecting activity systems. In the workshops, participants assessed various engagement formats in relation to these boundary-crossing dynamics. In-person sessions were described as productive for surfacing tacit knowledge, although resource-intensive (Creswell 2013). Online tools expanded geographic reach but were noted to risk excluding participants with limited connectivity (Goggin and & Soldatic 2022). Other modalities, such as surveys and deliberative assemblies, were associated with distinct trade-offs concerning inclusivity, cost, and depth of interaction (Goggin, Ellis, and & Hawkins 2019; Lewis et al. 2020; Turnhout, Van Bommel, and Aarts 2010). These insights suggest that no single method guarantees equitable participation. Instead, formats may need to be selected or

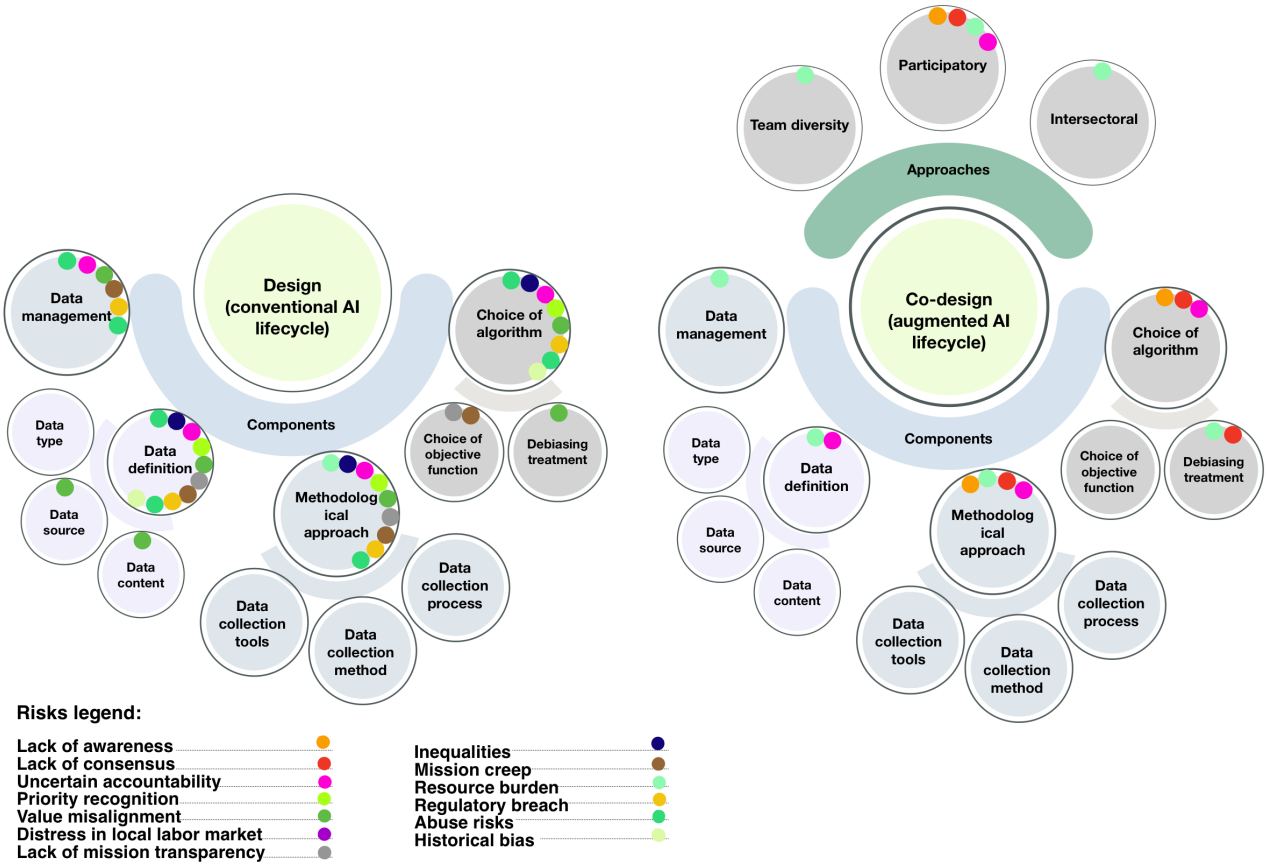

Figure 2: Phase-specific risks in a conventional AI lifecycle and corresponding mitigation via the co-design phase.

combined based on contextual factors, including institutional capacity and participant access.

Sustaining Partnerships Across the Lifecycle

Long-term legitimacy depends on structures that sustain dialogue beyond the initial design phase. Participants recommended institutionalizing "boundary spaces" such as public innovation labs, recurring town hall meetings, and virtual forums. These venues support ongoing monitoring of concept drift and shifts in social norms, aligning with the principles of the co-maintenance phase. As illustrated in Figure 3, Arnstein's ladder places such arrangements at the "partnership" level, where citizen influence extends to shared authority over strategic decisions (Arnstein 1969). Establishing clear charters that define roles, compensation, and audit rights is essential for maintaining stakeholder engagement over multi-year timeframes (Chambers et al. 2022).

Multidisciplinary Collaboration and Team Diversity

Workshop testimony indicated that projects requiring both social and technical competencies often involve coordination across disciplines, including computer science, DEI, social science, and law (Bisconti et al. 2023; de Hond, van Buchem, and Hernandez-Boussard 2022). Participants described the benefits of this arrangement as situational and contingent. While some reported gains in creativity and reflexivity, others noted the challenge of negotiating divergent epistemologies and terminologies (Sloane et al. 2022). Participants proposed strategies such as rotating facilitation, shared glossaries, and just-in-time expert input to manage friction points (Avellan, Sharma, and Turunen 2020; Cachat-Rosset and & Klarsfeld 2023).

Building consensus can be achieved throughout the discourse and decision-making process by allowing adequate time for all participants to express their opinions—for example, through the use of Indigenous tools like talking sticks or methods such as voting. Moreover, treating each partner equally, valuing their input, and avoiding power imbalances can promote consensus and agreement during the discourse (Lewis et al. 2020; Turnhout, Van Bommel, and Aarts 2010). Building consensus strengthens the legitimacy and acceptance of outcomes through inclusive dialogue and shared decision-making. It is considered crucial in conflict resolu-

Figure 3: Arnstein's ladder of citizen participation. A visual framework showing eight levels of citizen involvement, from nonparticipation to full citizen control in decision-making.

tion and community-based planning but requires patience, open-mindedness, and often, neutral facilitation to navigate differing viewpoints (Chambers et al. 2022).

Privacy, Anonymity, and Trust

Sustained collaboration likely requires robust local privacy guarantees. Participants conveyed a shared understanding that a layered approach—beginning with informed consent and progressing through k-anonymity, l–diversity, t–closeness, pseudonymization, and differential privacy—can be calibrated to the sensitivity of the data (Pawar, Ahirrao, and Churi 2018; Véliz 2021; Nissenbaum 2010). Onion-routing techniques protected communication channels during remote sessions (Obaidat et al. 2020). These measures, combined with regular transparency reports, appeared to strengthen trust and reduce attrition among citizen partners. Self-identification questionnaires enabled demographic monitoring without disclosing individual identities, aligning data stewardship practices with DEI objectives (Saunders et al. 2018).

Limitations and Future Work

This study is subject to three main limitations that circumscribe the evidentiary strength and external validity of its findings. **First**, data collection relied on four workshops held in Montréal, Canada, with twenty self-selected participants drawn from research, industry, and civil-society organizations. Lay citizens and communities situated outside the local AI ecosystem were not included. The resulting sample is small, geographically concentrated, and professionally homogeneous, which restricts transferability to less-resourced or culturally distinct settings and increases the likelihood that a limited number of voices shaped the consensus (Sloane et al. 2022; Bisconti et al. 2023). Social-desirability pressures may have further muted dissent despite anonymized transcription.

Second, the scoping review covered English-language literature. While we strived to include a broad range of sources, in-depth discussion of some literature did not occur. Due to the format of the workshops, we couldn't engage the literature systematically. No formal quality appraisal of individual studies was performed; therefore, the reliability of specific contributions varies across sources.

Third, the augmented lifecycle remains a conceptual framework whose practical feasibility, cost profile, and performance implications have not been empirically evaluated. The workshops provided preliminary face validity, but they do not establish that co-production systematically reduces algorithmic harms or enhances technical metrics. Potential conflicts between community veto rights and organizational accountability mechanisms identified in prior work (Helbing et al. 2023; De Silva and Alahakoon 2022) were not examined in operational environments.

Future research should address these constraints by (i) conducting longitudinal field trials that involve citizen cohorts—including historically marginalized groups—in multiple jurisdictions, (ii) extending the evidence base to incorporate non-English and gray sources, and (iii) collecting quantitative and qualitative data on fairness, privacy, cost, and project efficiency when the lifecycle is applied in production settings. Comparative studies across regulatory regimes will clarify how the framework interacts with emerging legislation such as the EU AI Act (European Commission 2018). Systematic cost–benefit analyses are also required to determine whether the additional governance overhead yields proportional societal value. While conceptual, this lifecycle offers a structured response to a widely acknowledged gap in operational AI ethics.

While conceptual, the proposed lifecycle offers a structured response to a widely acknowledged gap in operational AI ethics—namely, the absence of participatory, accountable processes that center the voices of those most affected by AI systems.

Risks and Challenges

Analysis of the workshops suggests that risks can emerge at various stages of AI development and are not evenly distributed across social groups. Participants, drawing on design-justice principles, emphasized that communities with limited institutional power often experience disproportionate exposure to these risks and should be included in risk-mitigation processes from the outset. Expansive-learning theory provides a framework through which these processes may be understood as opportunities for collective reframing of the AI activity system. Commitments to diversity, equity, and inclusion (DEI) inform which stakeholders are necessary for participation in these reframing processes.

Existing guidelines and empirical studies identify recurring patterns such as feedback loops that reinforce structural inequities, system opacity that limits contestability, and mission drift that expands AI use beyond its intended scope (Koseki et al. 2022). Leslie categorizes the resulting harms into six types: bias and discrimination; non-transparency; social isolation; denial of autonomy, recourse, and rights; unreliable or unsafe outcomes; and privacy invasion (Leslie 2019). Mohamed et al. contextualize these harms within broader post-colonial power asymmetries (Mohamed, Png,

Title	Principles	DEI Participation	Multidisciplinarity	Team Diversity
The Montreal Declaration for Responsible AI	Well-being			
	Respect for autonomy			
	Protection of privacy and intimacy			
	Solidarity			
	Democratic participation			
	Equity			
	Diversity inclusion			
	Caution			
	Responsibility			
	Sustainable development			
Microsoft Responsible AI Standard	Accountability			
	Transparency			
	Fairness			
	Reliability & safety			
	Privacy & security			
	Inclusiveness			
IEEE Global Initiative	Human rights			
	Well-being			
	Data agency			
	Effectiveness			
	Transparency			
	Accountability			
	Awareness of misuse			
	Competence			
AI Risk Management Framework — NIST	Valid and reliable			
	Safe			
	Secure and resilient			
	Accountable and transparent			
	Explainable and interpretable			
	Privacy-enhanced			
	Fair with harmful bias managed			
Ethics Guidelines for Trustworthy AI — EC	Human agency and oversight			
	Technical robustness and safety			
	Privacy and data governance			
	Transparency			
	Diversity, discrimination, & fairness			
	Environmental & societal well-being			
	Accountability			

Table 2: Alignment of selected AI ethics guidelines with dimensions of co-production in the AI lifecycle. Shaded cells indicate levels of alignment: light gray denotes high correspondence, while light green indicates areas needing further research. This comparison highlights opportunities for future inquiry and practice.

Impact Category	Description
Enhanced AI Reliability	By incorporating diverse perspectives from the co-framing to the co-maintenance phase, AI systems are likely to be more robust and reliable.
Minimized Biases	Active involvement of citizens, especially from marginalized communities, can help identify and rectify inherent biases in AI algorithms.
Contextual Solutions	AI systems co-produced with citizens are likely to be contextually relevant, addressing real-world challenges effectively.
Empowered Communities	The approach not only democratizes the AI development process but also empowers communities by giving them a voice in technology creation.

Table 3: Impact of the augmented AI lifecycle

and Isaac 2020). The workshop discussions were consistent with these classifications and further identified context-specific concerns, including labor displacement in localized economies and the implications for Indigenous data sovereignty.

The augmented AI lifecycle proposes mechanisms for addressing such risks by incorporating co-production checkpoints at each developmental phase. In earlier stages, inclusion of citizen and domain experts facilitated the surfacing of latent value conflicts. Iterative knowledge exchange supported the identification and modification of proxy variables that were associated with socioeconomic status. During co-implementation, community feedback on misclassification logs led to additional data collection intended to address representational imbalances. These instances demonstrate how mechanisms associated with expansive learning—such as boundary crossing, negotiation of meaning, and iterative model adjustment—can serve as instruments for operationalizing design-justice principles.

The co-production approach introduces several challenges. Diverse perspectives can complicate consensus formation and increase the resource demands of a project (Sloane et al. 2022; Varanasi and Goyal 2023). Meeting DEI-related participation goals often requires sustained outreach and adequate compensation, which may be under-prioritized in institutional planning. Multidisciplinary collaboration may be affected by epistemic divergence and differences in disciplinary language, leading to elevated coordination costs (Bisconti et al. 2023; Bikakis and & Zheng 2015). Ambiguities in role definitions can give rise to participation washing, where engagement is formal rather than substantive (Sloane et al. 2022). Addressing these issues may require governance frameworks, including decision-making charters, facilitation tools to support cross-disciplinary communication, and periodic audits to assess the outcomes of participatory interventions (Birhane et al. 2022).

Table 3 presents a summary of workshop participants' observations regarding the impacts and trade-offs associated with the augmented lifecycle. This framework provides one possible approach for managing tensions between technical performance goals and social accountability. Table 2 outlines its alignment with established ethical guidelines.

Conclusion

This study integrates design-justice scholarship, expansive-learning theory, and DEI research into an augmented AI lifecycle that operationalizes co-production across five interdependent phases. A scoping review and four multidisciplinary workshops provided empirical grounding for the model and demonstrated concrete mechanisms (co-framing, co-design, co-implementation, co-deployment, and co-maintenance) that redistribute decision-making authority toward affected publics.

The lifecycle remains conceptual until validated in applied settings. Future work should therefore implement the framework in domains such as health, finance, and public administration, measuring its effects on system performance, participant satisfaction, and long-term accountability. Comparative studies are required to evaluate scalability across organizational contexts and to analyze cost–benefit trade-offs relative to conventional pipelines. Further investigation of privacy, labor, and innovation outcomes will refine governance recommendations and inform alignment with evolving regulatory standards.

By situating technical choices within participatory structures that reflect diverse expertise and lived experience, the augmented AI lifecycle offers a practical pathway toward AI systems that are both contextually responsive and socially just.

References

Aizenberg, E.; and & van den Hoven, J. 2020. Designing for human rights in AI. Big Data & Society, 7(2), 2053951720949566. https://doi.org/10.1177/2053951720949566.

Alim, S.; and Adebayo, A. O. 2022. Ethics in artificial intelligence: Issues and guidelines for developing acceptable AI systems. Global Journal of Engineering and Technology Advances, 11(3), 37–44.

Ameri, F.; and Dutta, D. 2005. Product Lifecycle Management: Closing the Knowledge Loops. *Computer-Aided Design and Applications*, 2(5): 577–590.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. In *Ethics of Data and Analytics*, chapter 37, 254–264. Auerbach Publications.

Anthropic. 2023. Collective constitutional AI: Aligning a language model with public input. Technical Report.

Arnstein, S. R. 1969. A Ladder Of Citizen Participation. Journal of the American Institute of Planners, 35(4), 216–224. https://doi.org/10.1080/01944366908977225.

Arrieta, A. B.; Díaz-Rodríguez, N.; Ser, J. D.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58: 82–115.

Arslan, F. 2017. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, by Cathy O'Neil. Journal of Information Privacy and Security, 13(3), 157–159. https://doi.org/10.1080/15536548.2017.1357388.

Asaro, P. M. 2000. Transforming society by transforming technology: The science and politics of participatory design. Accounting, Management and Information Technologies, 10(4), 257–290. https://doi.org/10.1016/S0959-8022(00)00004-7.

Ashley, A. J.; Loh, C. G.; Bubb, K.; and Durham, L. 2022. Diversity, Equity, and Inclusion Practices in Arts and Cultural Planning. *Journal of Urban Affairs*, 44(4–5): 727–747.

Assal, H.; and Chiasson, S. 2018. Security in the Software Development Lifecycle. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS 2018)*, 281–296. Baltimore, MD: USENIX Association. ISBN 978-1-939133-10-6.

Avellan, T.; Sharma, S.; and Turunen, M. 2020. AI for All: Defining the What, Why, and How of Inclusive AI. In *Proceedings of the 23rd International Conference on Academic Mindtrek*, 142–144. New York, NY, USA: ACM. ISBN 978-1-4503-7774-4.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. ISBN 978-0-262-04861-3.

Beaudouin, V.; Bloch, I.; Bounie, D.; Clémençon, S.; d'Alché Buc, F.; Eagan, J.; Maxwell, W.; Mozharovskyi, P.; and Parekh, J. 2020. Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach. *arXiv preprint arXiv:2003.07703*.

Beck, K.; Beedle, M.; van Bennekum, A.; Cockburn, A.; Cunningham, W.; Fowler, M.; Grenning, J.; Highsmith, J.; Hunt, A.; Jeffries, R.; Kern, J.; Marick, B.; Martin, R. C.; Mellor, S.; Schwaber, K.; Sutherland, J.; and Thomas, D. 2001. Manifesto for Agile Software Development. https://agilemanifesto.org/.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J. T.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development*, 63(4/5): 4:1–4:15.

Berditchevskaia, A.; Malliaraki, E.; and & Peach, K. 2021. Participatory AI for humanitarian innovation: A briefing paper. Nesta, London. https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/.

Berditchevskaia, A.; Peach, K.; and Malliaraki, E. 2021. Participatory AI for Humanitarian Innovation. London: Nesta.

Bikakis, A.; and & Zheng, X. 2015. Multi-disciplinary Trends in Artificial Intelligence: 9th International Workshop, MIWAI 2015, Fuzhou, China, November 13-15, 2015, Proceedings (Vol. 9426). Springer International Publishing. https://doi.org/10.1007/978-3-319-26181-2.

Binns, R. 2020. On the apparent conflict between individual and group fairness. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 514–524. https://doi.org/10.1145/3351095.3372864.

Birhane, A.; Isaac, W.; Prabhakaran, V.; Díaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. 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)*, 6:1–6:8. New York, NY, USA: Association for Computing Machinery.

Bisconti, P.; Orsitto, D.; Fedorczyk, F.; Brau, F.; Capasso, M.; Marinis, L. D.; Eken, H.; Merenda, F.; Forti, M.; Pacini, M.; and Schettini, C. 2023. Maximizing team synergy in AI-related interdisciplinary groups: An interdisciplinary-by-design iterative methodology. *AI & Society*, 38(4): 1443–1452.

Brayne, S. 2017. Big Data Surveillance: The Case of Policing. American Sociological Review, 82(5), 977–1008. https://doi.org/10.1177/0003122417725865.

Buolamwini, J.; and & Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91.

Burema, D.; Debowski-Weimann, N.; von Janowski, A.; Grabowski, J.; Maftei, M.; Jacobs, M.; van der Smagt, P.; and Benbouzid, D. 2023. A Sector-Based Approach to AI Ethics: Understanding Ethical Issues of AI-Related Incidents Within Their Sectoral Context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 705–714. New York, NY, USA: Association for Computing Machinery.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512. https://doi.org/10.1177/2053951715622512.

Cachat-Rosset, G.; and & Klarsfeld, A. 2023. Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines. Applied Artificial Intelligence, 37(1), 2176618. https://doi.org/10.1080/08839514.2023.2176618.

Calabrese Barton, A.; and & Tan, E. 2020. Beyond Equity as Inclusion: A Framework of "Rightful Presence" for Guiding Justice-Oriented Studies in Teaching and Learning. Educational Researcher, 49(6), 433–440. Scopus. https://doi.org/10.3102/0013189X20927363.

Chambers, J. M.; Wyborn, C.; Klenk, N. L.; Ryan, M.; Serban, A.; Bennett, N. J.; Brennan, R.; Charli-Joseph, L.; Fernández-Giménez, M. E.; Galvin, K. A.; Goldstein, B. E.; Haller, T.; Hill, R.; Munera, C.; Nel, J. L.; Österblom, H.; Reid, R. S.; Riechers, M.; Spierenburg, M.; Tengö, M.; Bennett, E.; Brandeis, A.; Chatterton, P.; Cockburn, J. J.; Cvitanovic, C.; Dumrongrojwatthana, P.; Durán, A. P.; Gerber, J.-D.; Green, J. M. H.; Gruby, R.; Guerrero, A. M.; Horcea-Milcu, A.-I.; Montana, J.; Steyaert, P.; Zaehringer, J. G.; Bednarek, A. T.; Curran, K.; Fada, S. J.; Hutton, J.; Leimona, B.; Pickering, T.; and Rondeau, R. 2022. Coproductive Agility and Four Collaborative Pathways to Sustainability Transformations. *Global Environmental Change*, 72: 102422.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047.

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

Creswell, J. W. 2013. Qualitative Inquiry and Research Design: Choosing Among Five Approaches. SAGE.

Dastin, J. 2022. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women *. In Ethics of Data and Analytics. Auerbach Publications.

de Hond, A. A. H.; van Buchem, M. M.; and Hernandez-Boussard, T. 2022. Picture a Data Scientist: A Call to Action for Increasing Diversity, Equity, and Inclusion in the Age of

AI. *Journal of the American Medical Informatics Association*, 29(12): 2178–2181.

De Silva, D.; and Alahakoon, D. 2022. An artificial intelligence life cycle: From conception to production. Patterns, 3(6), 100489.

Demelenne, A.; Gou, M.-J.; Nys, G.; Parulski, C.; Crommen, J.; Servais, A.-C.; and Fillet, M. 2020. Evaluation of Hydrophilic Interaction Liquid Chromatography, Capillary Zone Electrophoresis and Drift Tube Ion-Mobility Quadrupole Time-of-Flight Mass Spectrometry for the Characterization of Phosphodiester and Phosphorothioate Oligonucleotides. Journal of Chromatography A, 1614, 460716. DOI: 10.1016/j.chroma.2019.460716.

Deming, W. E. 1986. Out of the crisis. MIT Press.

Divakaran, A.; Sridhar, A.; and Srinivasan, R. 2023. Broadening AI Ethics Narratives: An Indic Art View. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2–11. https://doi.org/10.1145/3593013.3593971.

Donia, J.; and Shaw, J. A. 2021. Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211065248.

Dwivedi, Y. K.; Hughes, L.; Ismagilova, E.; Aarts, G.; Coombs, C.; Crick, T.; Duan, Y.; Dwivedi, R.; Edwards, J.; Eirug, A.; Galanos, V.; Ilavarasan, P. V.; Janssen, M.; Jones, P.; Kar, A. K.; Kizgin, H.; Kronemann, B.; Lal, B.; Lucini, B.; Medaglia, R.; Meunier-FitzHugh, K. L.; Meunier-FitzHugh, L. C. L.; Misra, S.; Mogaji, E.; Sharma, S. K.; Singh, J. B.; Raghavan, V.; Raman, R.; Rana, N. P.; Samothrakis, S.; Spencer, J.; Tamilmani, K.; Tubadji, A.; Walton, P.; and Williams, M. D. 2021. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57: 101994.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226). ACM. https://doi.org/10.1145/2090236.2090255.

Engeström, Y. 2014. Learning by Expanding: An Activity-Theoretical Approach to Developmental Research (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139814744.

Engeström, Y.; and Sannino, A. 2010. Studies of expansive learning: Foundations, findings and future challenges. Educational Research Review, 5(1), 1–24. https://doi.org/10.1016/j.edurev.2009.12.002.

European Commission. 2018. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation.

Flanagan, M.; and Nissenbaum, H. 2016. Values at Play in Digital Games. https://mitpress.mit.edu/9780262529976/values-at-play-in-digital-games/.

Forum, W. E. 2022. A Blueprint for Equity and Inclusion in Artificial Intelligence. World Economic Forum.

Galaz, V.; Centeno, M. A.; Callahan, P. W.; Causevic, A.; Patterson, T.; Brass, I.; Baum, S.; Farber, D.; Fischer, J.; Garcia, D.; McPhearson, T.; Jimenez, D.; King, B.; Larcey, P.; and Levy, K. 2021. Artificial intelligence, systemic risks, and sustainability. *Technology in Society*, 67: 101741.

Gerdes, A. 2018. An Inclusive Ethical Design Perspective for a Flourishing Future with Artificial Intelligent Systems. European Journal of Risk Regulation, 9(4), 677–689. https://doi.org/10.1017/err.2018.62.

Gerdes, A. 2022. A participatory data-centric approach to AI Ethics by Design. Applied Artificial Intelligence, 36(1), 2009222. https://doi.org/10.1080/08839514.2021.2009222.

Gerdes, A.; and & Frandsen, T. F. 2023. A systematic review of almost three decades of value sensitive design (VSD): What happened to the technical investigations? Ethics and Information Technology, 25(2), 26. https://doi.org/10.1007/s10676-023-09700-2.

Goggin, G.; and & Soldatic, K. 2022. Automated decision-making, digital inclusion and intersectional disabilities. New Media and Society, 24(2), 384–400. Scopus. https://doi.org/10.1177/14614448211063173.

Goggin, G.; Ellis, K.; and & Hawkins, W. 2019. Disability at the centre of digital inclusion: Assessing a new moment in technology and rights. Communication Research and Practice, 5(3), 290–303. https://doi.org/10.1080/22041451.2019.1641061.

Golpayegani, D.; Pandit, H. J.; and & Lewis, D. 2022. AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. In A. Dimou, S. Neumaier, T. Pellegrini, & S. Vahdati (Eds.), Studies on the Semantic Web. IOS Press. https://doi.org/10.3233/SSW220008.

Gowaikar, S.; Berard, H.; Mushkani, R.; and Koseki, S. 2024. From Efficiency to Equity: Measuring Fairness in Preference Learning. arXiv:2410.18841.

Haakman, M.; Cruz, L.; Huijgens, H.; and van Deursen, A. 2021. AI lifecycle models need to be revised. Empirical Software Engineering, 26(5), 95. https://doi.org/10.1007/s10664-021-09993-1.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

Helbing, D.; Mahajan, S.; Hänggli, R.; Musso, A.; Hausladen, C. I.; Carissimo, C.; Carpentras, D.; Stockinger, E.; Sánchez-Vaquerizo, J. A.; Yang, J. C.; Ballandies, M. C.; Korecki, M.; Dubey, R. K.; and Pournaras, E. 2023. Democracy by Design: Perspectives for Digitally Assisted, Participatory Upgrades of Society. *Journal of Computational Science*, 71: 102061.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; van den Driessche, G.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Rae, J. W.; Vinyals, O.; and Sifre, L. 2022. Training Compute-Optimal Large Language Models. https://arxiv.org/abs/2203.15556.

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

Hummer, W.; Muthusamy, V.; Rausch, T.; Dube, P.; El Maghraoui, K.; Murthi, A.; and Oum, P. 2019. ModelOps: Cloud-Based Lifecycle Management for Reliable and Trusted AI. 2019 IEEE International Conference on Cloud Engineering (IC2E), 113–120. https://doi.org/10.1109/IC2E.2019.00025.

Ibrahim, R.; and Paulson, B. C. 2008. Discontinuity in organisations: Identifying business environments affecting efficiency of knowledge flows in Product Lifecycle Management. International Journal of Product Lifecycle Management, 3(1), 21–36. https://doi.org/10.1504/IJPLM.2008.019972.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Jordan, R. 2023. Design Justice: Community-Led Practices to Build the Worlds We Need: by S. Costanza-Chock, Massachusetts, MIT Press, 2020, 360 pp., $25.00 (paperback), Open Access, ISBN: 9780262043458. Technical Communication Quarterly, 32(1), 114–116. https://doi.org/10.1080/10572252.2022.2130671.

Kiritsis, D.; Bufardi, A.; and Xirouchakis, P. 2003. Research issues on product lifecycle management and information tracking using smart embedded systems. Advanced Engineering Informatics, 17(3), 189–202. https://doi.org/10.1016/j.aei.2004.09.005.

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. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019.

Korobenko, D.; Nikiforova, A.; and Sharma, R. 2024. Towards a Privacy and Security-Aware Framework for Ethical AI: Guiding the Development and Assessment of AI Systems. Proceedings of the 25th Annual International Conference on Digital Government Research, 740–753.

Koseki, S.; Jameson, S.; Farnadi, G.; Denis, J.; Regis, L.; Rolnick, C.; Lahoud, C.; Pienaar, J.; Thung, I.; Owigar, J.; Sommer, K.; Nkuidje, L.; Pennanen-Rebeiro-Hargrave, P.; Westerberg, P.; Sietchiping, R.; Yousry, S.; Prud'homme, B.; Landry, R.; Sagar, A. S.; and L'Archeveque, S. 2022. AI and Cities: Risk, Applications and Governance. UN Habitat.

Lee, M. K.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; Lee, S.; Psomas, A.; and Procaccia, A. D. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Lember, V.; Brandsen, T.; and Tõnurist, P. 2019. The potential impacts of digital technologies on co-production and co-creation. Public Management Review, 21(11), 1665–1686. https://doi.org/10.1080/14719037.2019.1619807.

Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; and Vinck, P. 2018. Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges. Philosophy & Technology, 31(4), 611–627. DOI: 10.1007/s13347-017-0279-x.

Leslie, D. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Zenodo. https://doi.org/10.5281/ZENODO.3240529.

Lewis, J. E.; Abdilla, A.; Arista, N.; Baker, K.; Benesiinaabandan, S.; Brown, M.; Cheung, M.; Coleman, M.; Cordes, A.; Davison, J.; Duncan, K.; Garzon, S.; Harrell, D. F.; Jones, P. L.; Kealiikanakaoleohaililani, K.; Kelleher, M.; Kite, S.; Lagon, O.; Leigh, J.; and Whaanga, H. 2020. Indigenous Protocol and Artificial Intelligence Position Paper. Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research.

Li, F.; and Lu, Y. 2022. Human-AI interaction and ethics of AI: How well are we following the guidelines. Proceedings of the Tenth International Symposium of Chinese CHI, 96–104. https://doi.org/10.1145/3565698.3565773.

McBride, K.; Nikiforova, A.; Lnenicka, M.; Kempeneer, S.; and Wolswinkel, J. 2023. The role of open government data and co-creation in crisis management: Initial conceptual propositions from the COVID-19 pandemic. Info. Pol., 28(2), 219–238. https://doi.org/10.3233/IP-220057.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology, 33(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8.

Mohammed, N. M.; Niazi, M.; Alshayeb, M.; and Mahmood, S. 2017. Exploring software security approaches in software development lifecycle: A systematic mapping study. Computer Standards & Interfaces, 50, 107–115. https://doi.org/10.1016/j.csi.2016.10.001.

Mushkani, R.; Berard, H.; Cohen, A.; and Koseki, S. 2025a. Position: The Right to AI. arXiv:2501.17899v1.

Mushkani, R.; Berard, H.; and Koseki, S. 2025. Negotiative Alignment: Embracing Disagreement to Achieve Fairer Outcomes – Insights from Urban Studies. arXiv:2503.12613.

Mushkani, R.; Nayak, S.; Berard, H.; Cohen, A.; Koseki, S.; and Bertrand, H. 2025b. LIVS: A Pluralistic Alignment Dataset for Inclusive Public Spaces. arXiv:2503.01894.

Nayak, S.; Mushkani, R.; Berard, H.; Cohen, A.; Koseki, S.; and Bertrand, H. 2024. MID-Space: Aligning Diverse Communities' Needs to Inclusive Public Spaces. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Nissenbaum, H. 2010. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

NIST. 2023. Safeguarding International Science: Research Security Framework, National Institute of Standards and Technology. NIST Interagency/Internal Report 8484. DOI: 10.6028/NIST.IR.8484.

Obaidat, M. A.; Obeidat, S.; Holst, J.; Al Hayajneh, A.; and Brown, J. 2020. A Comprehensive and Systematic Survey on the Internet of Things: Security and Privacy Challenges, Security Frameworks, Enabling Technologies, Threats, Vulnerabilities and Countermeasures. Computers, 9(2), 44. https://doi.org/10.3390/computers9020044.

Pawar, A.; Ahirrao, S.; and Churi, P. P. 2018. Anonymization Techniques for Protecting Privacy: A Survey. 2018 IEEE Punecon, Pune, India, 1–6. https://doi.org/10.1109/PUNECON.2018.8745425.

Qian, M.; Lee, W. D.; and Schwanen, T. 2021. The Association Between Socioeconomic Status and Mobility Reductions in the Early Stage of England's COVID"{19} Epidemic. Health & Place, 69, 102563. DOI: 10.1016/j.healthplace.2021.102563.

Ravanera, C.; and Kaplan, S. 2021. An Equity Lens on Artificial Intelligence. https://www.gendereconomy.org/artificial-intelligence/.

Roth, W.-M. 2004. Introduction: "Activity Theory and Education: An Introduction". Mind, Culture, and Activity, 11(1), 1–8. https://doi.org/10.1207/s15327884mca1101_1.

Saunders, B.; Sim, J.; Kingstone, T.; Baker, S.; Waterfield, J.; Bartlam, B.; Burroughs, H.; and Jinks, C. 2018. Saturation in qualitative research: Exploring its conceptualization and operationalization. Quality & Quantity, 52(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8.

Schiff, D.; Borenstein, J.; Biddle, J.; and Laas, K. 2021a. AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. IEEE Transactions on Technology and Society, 2(1), 31–42. https://doi.org/10.1109/TTS.2021.3052127.

Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; and Lennon, M. 2021b. Explaining the Principles to Practices Gap in AI. IEEE Technology and Society Magazine, 40(2), 81–94.

Sculley, D.; Holt, G.; Golovin, D.; et al. 2015. Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503–2511.

Sloane, M.; Moss, E.; Awomolo, O.; and & Forlano, L. 2022. Participation Is not a Design Fix for Machine Learning. ACM International Conference Proceeding Series. Scopus. https://doi.org/10.1145/3551624.3555285.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; Althoff, T.; and Choi, Y. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org.

Terzi, S.; Bouras, A.; Dutta, D.; Garetti, M.; and Kiritsis, D. 2010. Product lifecycle management – from its history to its new role. International Journal of Product Lifecycle Management, 4(4), 360–389. https://doi.org/10.1504/IJPLM.2010.036489.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. Accessed: 2025-05-16.

Turnhout, E.; Van Bommel, S.; and Aarts, N. 2010. How Participation Creates Citizens: Participatory Governance as Performative Practice. Ecology and Society, 15(4). https://www.jstor.org/stable/26268213.

Umbrello, S.; and van de Poel, I. 2021. Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3.

Université de Montréal. 2018. Montréal Declaration for a Responsible Development of Artificial Intelligence. Accessed: 2025-05-16.

Varanasi, R. A.; and Goyal, N. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3544548.3580903.

Véliz, C. 2021. Privacy is Power: Why and How You Should Take Back Control of Your Data. Penguin Books.

Wang, A.; Ramaswamy, V. V.; and Russakovsky, O. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. 2022 ACM Conference on Fairness, Accountability, and Transparency, 336–349. https://doi.org/10.1145/3531146.3533101.

Wang, D.; Ram, P.; Weidele, D. K. I.; Liu, S.; Muller, M.; Weisz, J. D.; Valente, A.; Chaudhary, A.; Torres, D.; Samulowitz, H.; and Amini, L. 2020. AutoAI: Automating the End-to-End AI Lifecycle with Humans-in-the-Loop. Proceedings of the 25th International Conference on Intelligent User Interfaces Companion, 77–78.

Wong, R. Y.; Madaio, M. A.; and Merrill, N. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. Proceedings of the ACM on Human–Computer Interaction, 7(CSCW1), Article 145. https://doi.org/10.1145/3579621.

Wu, S. T.-I.; Demetriou, D.; and Husain, R. A. 2023. Honor Ethics: The Challenge of Globalizing Value Alignment in AI. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 593–602. https://doi.org/10.1145/3593013.3594026.

Yu, K.-H.; Beam, A. L.; and Kohane, I. S. 2018. Artificial Intelligence in Healthcare. *Nature Biomedical Engineering*, 2(10): 719–731.

Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; and Wu, X. 2019. Object Detection with Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3212–3232.

Zicari, R. V.; Sheraz, A.; Amann, J.; Braun, S. A.; Brodersen, J.; Bruneault, F.; Brusseau, J.; Campano, E.; Coffee, M.; Dengel, A.; Düdder, B.; Gallucci, A.; Gilbert, T. K.; Gottfrois, P.; Goffi, E.; Haase, C. B.; Hagendorff, T.; Hickman,

E.; Hildt, E.; Holm, S.; Kringen, P.; Kühne, U.; Lucieri, A.; Madai, V. I.; Moreno-Sánchez, P. A.; Medlicott, O.; Ozols, M.; Schnebel, E.; Spezzatti, A.; Tithi, J. J.; Umbrello, S.; Vetter, D.; Volland, H.; Westerlund, M.; and Wurth, R. 2021. Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. *Frontiers in Human Dynamics*, 3.