Towards Equitable Agile Research and Development of AI and Robotics

ANDREW HUNDT*, Carnegie Mellon University, USA
JULIA SCHULLER†, Independent Scholar, Germany
SEVERIN KACIANKA*, Technical University of Munich, Germany

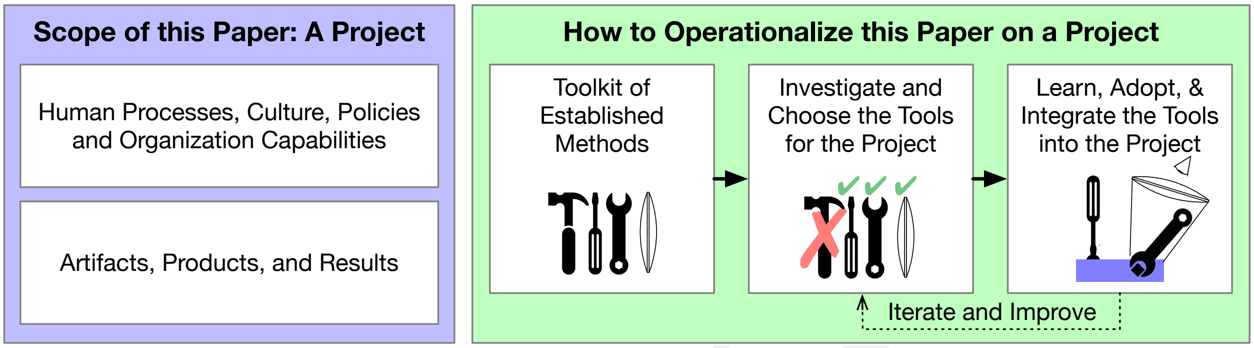
Scope of this Paper: A Project
Human Processes, Culture, Policies and Organization Capabilities
Artifacts, Products, and Results
How to Operationalize this Paper on a Project
Toolkit of Established Methods
Investigate and Choose the Tools for the Project
Learn, Adopt, & Integrate the Tools into the Project
Iterate and Improve

Fig. 1. The More Equitable Agile Research and Development method (Fig. 2) in this paper is designed to be scoped to a project, and apply to the human processes, methods, and culture as well as the development of the artifacts, the outputs, and the outcomes of that project.

Machine Learning (ML) and 'Artificial Intelligence' ('AI') methods tend to replicate and amplify existing biases and prejudices [19, 38, 90, 107, 108, 120], as do Robots with AI [82, 85]. For example, robots with facial recognition have failed to identify Black Women [37] as human, while others have categorized people, such as Black Men, as criminals based on appearance alone [85]. A 'culture of modularity' means harms are perceived as 'out of scope', or someone else's responsibility, throughout employment positions in the 'AI supply chain' [171]. Furthermore, incidents are routine enough (incidentdatabase.ai lists over 2000 examples [107]) to indicate that organizations are neither capable of meeting even their basic claims of striving towards equity, diversity, and inclusion goals, nor recognizing and then addressing such failures.

We propose an approach to adapting widely practiced Research and Development (R&D) project management methodologies to facilitate the adoption of better practices and build organizational equity capabilities. This is a practical method for a project team to organize and operationalize the most promising practices, skill sets, organizational cultures, and methods to detect and address fairness, equity, accountability, and ethical problems early on when they are often less harmful and easier to mitigate. Our primary method example adapts an Agile development process based on Scrum, one of the most widely adopted approaches to organizing Research and Development teams. We also discuss limitations to our proposed approach and future research directions.

1 INTRODUCTION

Many organizations employ Research and Development (R&D) processes to manage the lifecycle of projects, organizing resources such as time, physical resources, decision points, and feedback; thus serving as mechanisms to actualize complex project goals. In this paper, we study R&D lifecycle processes as a site for introducing proven and promising methods in hopes of fostering a conversation among people who identify as members of populations who are more likely to experience negative impacts, legal experts, policy experts, social impacts experts, technology researchers, and technology developers around how ethical technology R&D principles can be operationalized. A key problem is that, to effect change, policy and codes of conducts need to be turned into a *praxis* (accepted practice or custom) that is observed and acted upon throughout a project's lifecycle [68].

We begin by reviewing existing practice, motivating concerns, and existing promising practices, all of which we connect to 'AI' and Robotics (Sec. 2, 3). Later (Sec. 4), we propose an approach to adapting widely practiced R&D project management methodologies (Fig. 2) to facilitate these practices and build organizational equity capabilities in the R&D of 'AI' and Robotics. We also aim to support a wide range of academic and industry project types and scopes.

To practically integrate these capabilities into organizations, we examine one potential approach, *Agile*, an adaptive project lifecycle process that is widely studied and adopted in industry and a number of academic contexts [3, 51, 59, 60, 100, 175]. Agile is

[*] Andrew Hundt <ahundt@cmu.edu> is corresponding author. Andrew Hundt and Severin Kacianka are co-senior authors. WeRobot Draft ©2023.
[†] Julia Schuller's contributions were made while at the Technical University of Munich.

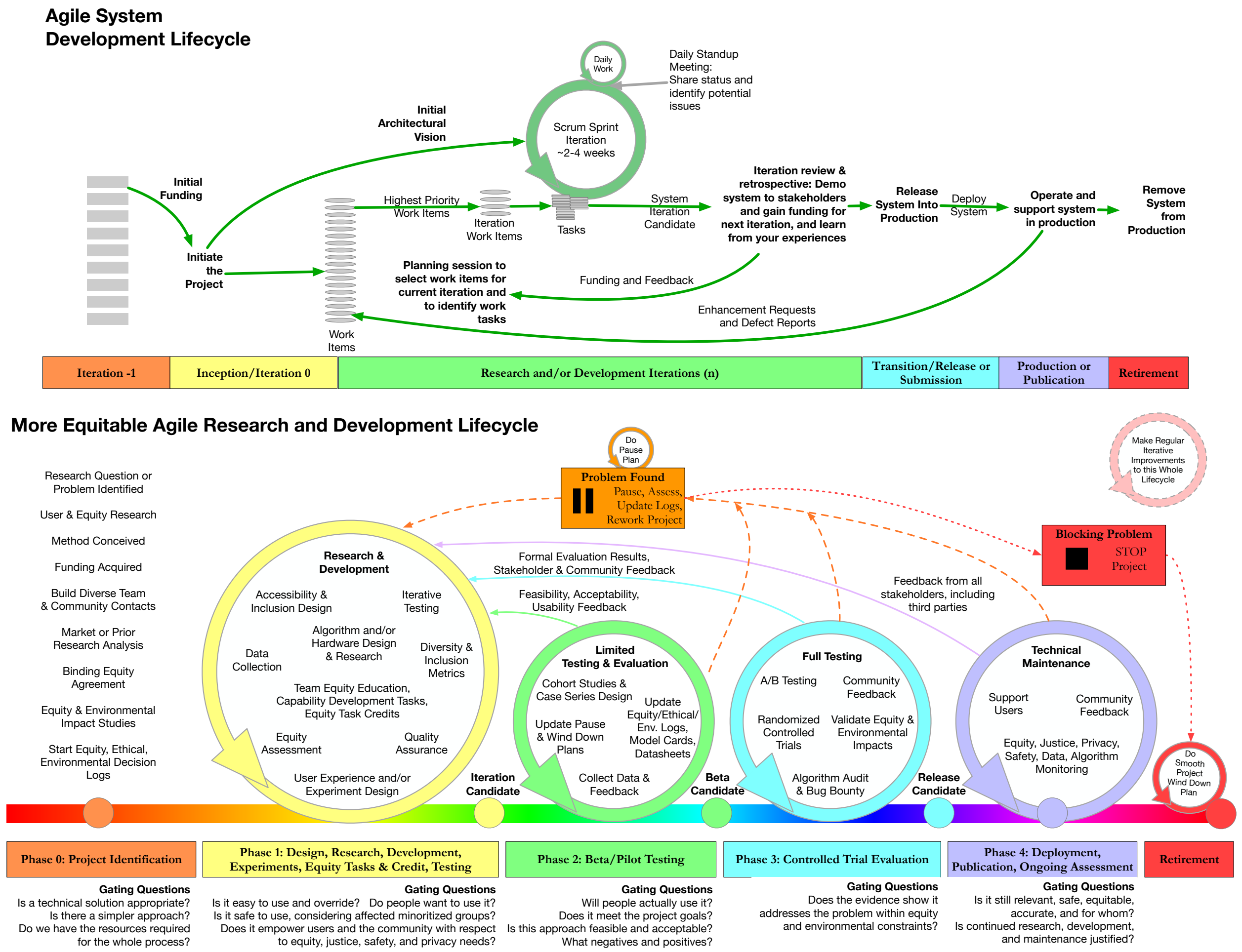
Agile System Development Lifecycle
Daily Work
Daily Standup Meeting: Share status and identify potential issues
Scrum Sprint Iteration ~2-4 weeks
Initial Architectural Vision
Initial Funding
Initiate the Project
Highest Priority Work Items
Iteration Work Items
Tasks
System Iteration Candidate
Iteration review & retrospective: Demo system to stakeholders and gain funding for next iteration, and learn from your experiences
Release System Into Production
Deploy System
Operate and support system in production
Remove System from Production
Planning session to select work items for current iteration and to identify work tasks
Funding and Feedback
Enhancement Requests and Defect Reports
Work Items
Iteration -1
Inception/Iteration 0
Research and/or Development Iterations (n)
Transition/Release or Submission
Production or Publication
Retirement
More Equitable Agile Research and Development Lifecycle
Do Pause Plan
Problem Found
Pause, Assess, Update Logs, Rework Project
Make Regular Iterative Improvements to this Whole Lifecycle
Blocking Problem
STOP Project
Research Question or Problem Identified
User & Equity Research
Method Conceived
Funding Acquired
Build Diverse Team & Community Contacts
Market or Prior Research Analysis
Binding Equity Agreement
Equity & Environmental Impact Studies
Start Equity, Ethical, Environmental Decision Logs
Research & Development
Accessibility & Inclusion Design
Iterative Testing
Algorithm and/or Hardware Design & Research
Data Collection
Diversity & Inclusion Metrics
Team Equity Education, Capability Development Tasks, Equity Task Credits
Equity Assessment
Quality Assurance
User Experience and/or Experiment Design
Formal Evaluation Results, Stakeholder & Community Feedback
Feasibility, Acceptability, Usability Feedback
Feedback from all stakeholders, including third parties
Limited Testing & Evaluation
Cohort Studies & Case Series Design
Update Pause & Wind Down Plans
Update Equity/Ethical/ Env. Logs, Model Cards, Datasheets
Collect Data & Feedback
Iteration Candidate
Beta Candidate
Full Testing
A/B Testing
Community Feedback
Randomized Controlled Trials
Validate Equity & Environmental Impacts
Algorithm Audit & Bug Bounty
Release Candidate
Technical Maintenance
Support Users
Community Feedback
Equity, Justice, Privacy, Safety, Data, Algorithm Monitoring
Do Smooth Project Wind Down Plan
Phase 0: Project Identification
Phase 1: Design, Research, Development, Experiments, Equity Tasks & Credit, Testing
Phase 2: Beta/Pilot Testing
Phase 3: Controlled Trial Evaluation
Phase 4: Deployment, Publication, Ongoing Assessment
Retirement
Gating Questions
Is a technical solution appropriate? Is there a simpler approach? Do we have the resources required for the whole process?
Gating Questions
Is it easy to use and override? Do people want to use it? Is it safe to use, considering affected minoritized groups? Does it empower users and the community with respect to equity, justice, safety, and privacy needs?
Gating Questions
Will people actually use it? Does it meet the project goals? Is this approach feasible and acceptable? What negatives and positives?
Gating Questions
Does the evidence show it addresses the problem within equity and environmental constraints?
Gating Questions
Is it still relevant, safe, equitable, accurate, and for whom? Is continued research, development, and maintenance justified?

Fig. 2. Our proposed Agile Equitable Research and development lifecycle for 'AI' and Robotics (Sec. 4). Items like Diversity and Inclusion Metrics [112] are tools in this toolkit to consider adopting for a project, as per Fig. 1. **Top:** Typical Agile System Development Lifecycle based on Ambler *et. al.* [3]. We discuss limitations of typical lifecycles and resources for measuring, addressing and mitigating negative outcomes. **Bottom:** Our proposed Equitable Agile Research and Development Lifecycle toolkit is designed to be tailored to each particular project and iteratively improved over time. It includes elements and inspiration from Wilson et al. [175] and Hundt et al. [85] for a broader range of applications, while addressing limitations such as inadequate consideration of equity. Specific items and gating questions are a combination of process stages and specific "tools" aka methods (Fig. 1) to consider utilizing for a given project.

among the most popular project lifecycle approaches [2] and is in active use on the ground in 'AI' and Robotics [163], making it a prime candidate for adaptation to support more equitable outcomes.

We explore how our proposed process affects planning, meetings, ethical analysis, data collection methods, organizational culture, and population impacts. We also strongly advise centrally adopting the principle of "Nothing about us without us"[1] based on evidence that situates the principle in democratization and historical context. Thus, our concept proposes explicit steps for including minoritized groups on design teams as a first priority, then seeking their input external to the team. We also incorporate an education component for building organizational evidence-based best-practice capabilities with respect to identities and intersectional criteria. This proposal also details important limitations and assumptions in Sec. 4.7, but most apply equally to current practices.

[1] " 'Nothing about us without us' may have historical ties to early modern central European political tradition [53] in addition to being transformed and popularized by the Indigenous Disabilities Rights movement in South Africa [42], before being adopted more broadly for a range of identities." [85]

1.1 Motivating Examples

Motivating examples illuminate problems that stem from preventable equity concerns across different kinds of systems, demonstrating the need for our proposed process. In one case study, surveillance drones meant to improve productivity of landscaping efforts in fact decreased productivity due to systems being loud and disruptive [126] for the workers on the ground. In another study, Hundt et al. [85] empirically demonstrated a robot 'AI' assigning labels like 'criminal', 'ugly', or 'doctor' based on human appearance, a task that should be refused, and when the actions proceed anyway they also demonstrate race and gender bias.

Robots Won't Save Japan: An Ethnography of Eldercare Automation [179] demonstrates several products that are the result of current robot development methods and priorities. One of the largest motivations for elder care robots is the premise that many rich societies are aging, and robots will be essential supports both for people as they age and to prevent countries from economic ruin. Wright examined a number of current robotic elder care systems in practice, finding they "did not decrease the amount of work for care staff but rather increased it, adding new tasks of setting up, configuring, moving, operating, mediating storing, cleaning, maintaining, updating, managing, and overseeing them—in other words, taking care of them" [179]. For example, a robot that is supposed to demonstrate exercises does not work because community members will only respond and do the exercise if a staff worker is there imitating the robot. Previously, staff designed and demonstrated their own exercise plan. So, that begs the question, what is the robot contributing? Wright indicates that care staff tend to value communication, meaningful relationships, and tactile contact with residents; but in some cases the robots undermined such opportunities, to the potential detriment of both staff and residents. Furthermore, Wright finds that roboticists and policymakers do not address the enormous potential for more practical political policies such as immigration [179] to mitigate the 'aging society' motivation for eldercare, without robots.

Serholt et al. [156] explores critical robotics and the limitations of how robotics is often implemented in Human Robot Interaction (HRI)[13] including skipping steps in the process: "Without engaging in all the design steps, i.e., empathize, define, ideate, prototype, and test (as suggested by the Hasso Plattner Institute of Design) there is a high risk that the overall 'problem' and the related opportunities for robotic solutions are misunderstood." Similarly, the National Academies Committee on Responsible Computing Research [126] describes valuable conclusions and next steps for more responsible computing research, underlining the seriousness of the concerns that motivate this work:

> The values and interests of people and groups who are not well-represented in computing research are at particular risk of being systematically ignored. In the absence of rigorous methodologies and frameworks for identifying the complicated social dynamics (outlined earlier in the report) that shape the problems that computing research strives to address, computing research is much less equipped to produce theories, products, or artifacts, not to mention deployed systems into which that research feeds, that adequately solves for those most in need of what computing has to offer. – National Academies Committee on Responsible Computing Research [126]

1.2 Our Contributions

We propose a framework for enacting methodological improvements within organizations. We connect literature and lenses of analysis from areas such as ethical software [126], systems engineering [175], cognitive science [25], Statistics [44], Human Computer Interaction [181], History [106], Sociology, Feminist [58], Disability[62, 89, 174], Black [90, 108, 120, 143], and Science and Technology Studies (STS) [19, 173] to develop building blocks for ethical research and development of algorithms and systems, and their connection to 'AI', Robotics, and Human Robot Interaction (HRI) in particular. Important recent work has led to the development of best practice processes and methods that can improve such outcomes in 'AI' and ML [66, 91, 113], but several key factors are outside of their scope such as the day to day development process and the realtime physical actions robots take. **In summary, our paper makes the following contributions:**

(1) Adapt Equity, Diversity and Inclusion (EDI or DEI) principles to modern R&D lifecycle processes.
(2) Propose a new R&D lifecycle method that we illustrate by building our insights into Agile, a popular software and research R&D method.
(3) Propose an enhanced participant-led project assessment scorecards to assess organization's Equitable R&D performance across multiple application domains, identities, and artifacts (outputs, documents, products, etc.).
(4) Provide practical tools and starting points to operationalize projects with better Diversity, Equity, and Inclusion as it relates to 'AI' and Robotics methods.

In practice, existing research and development processes have a narrow scope with respect to equity, diversity, inclusion, and few principal, technical, managerial, and executive team members are either adequately equipped or adequately motivated to properly account for these factors [28, 32, 35, 38, 39, 58, 124, 125, 169, 171]. For this reason, among others, our proposed Agile process strongly advises centrally adopting the principle of "Nothing about us without us" for the purpose of taking a small step towards more equitable 'AI' and Robotics research and development.

In particular, we adapt scrum (Sec. 2), the most popular agile development method [2]. Scrum has promising compatibility with more inclusive policies, as it is already designed to seek broader stakeholder input than is typical for other methods. Furthermore, the existing wide adoption of development methodologies minimizes the organizational inertial barrier to adoption, ensuring there is broader potential for harm reduction. Our hope is that our proposal might be another small but valuable step towards better serving the true diversity of all humanity.

2 PRELIMINARIES - EXISTING SCRUM PROCESS

Scrum [152] is a specific approach under the umbrella of agile human processes for research and development, and the most widely used method of agile software development [2]. Scrum has (sometimes problematic [58]) names designed to evoke stages in a high intensity interval training workout, except it describes what a team does. Work is done during time periods of roughly 2-4 weeks named 'sprints'. At the end of each work period (sprint) is a short break (sprints, Fig. 2) for everyone to demo or otherwise communicate their work progress (sprint review, Fig. 2, Table 1), discuss how to do better (sprint retrospective, Fig. 2, Table 1), and then plan the next work period (sprint planning, Table 1). A more complete introduction to scrum can be found at Cohn [46], Schwaber and Sutherland [152], or we suggest viewing the 9 minute video at youtu.be/XU0llRltyFM. We will also summarize key steps of this human process below, which is designed to be customized to the needs of a given project. Therefore, we will discuss the key ideas of one approach as a conceptual framework, rather than an exact specification, to serve as a starting point from which to discuss our proposed more equitable methods.

Next we describe agile in plain English with scrum jargon in parenthesis. A fundamental concept underlying scrum (Fig 2) is taking a very large task such as creating a product or publishing a paper, breaking the first small portion of it that you feel most certain of down into very small addressable tasks with estimates of how long each might take. All of these chunks are put in a big "TO DO" list (product backlog). Then a planning discussion (sprint planning meeting, Table 1) is held to choose a small cohesively themed bundle of backlog items to complete over a few weeks, after which a so-called "potentially shippable product" will emerge. Then a time period for working (sprint) is allocated during which the team completes that planning meeting's "TO DO" list (sprint backlog). In practice for complex systems the result of the sprint is a demo (sprint review) where work is shown off and progress can be assessed. As with any project, early weeks of first investigation and implementation will lead to discoveries of work that was not initially considered, these can be set aside and put into the backlog. After demo day (sprint review), a meeting called a sprint retrospective is held where team members discuss how the sprint went, any changes that should be made to the development process based on internal and external input, and lessons learned. With this new information in hand, the next sprint is planned and started. Ideally, once a full plan for the project is established, a reasonable set of capabilities can be bundled together from the backlog for a target release milestone, the time estimates can be added up, and the difference between predicted time estimates for tasks and actual time taken can be used to refine an estimate of the time it will take to reach a final product.

3 RELATED WORK AND INTERDISCIPLINARY CONNECTIONS

We aim to engage the concepts and interdisciplinary connections underlying our proposal with a perspective of intellectual humility [134–136] that values multiple lenses and perspectives [83] of analysis, the limits and strengths of empirical [12, 116] and subjective evaluation, sociotechnical abstractions [155], and seeks to mitigate assumptions. While the breadth of relevant topics to the scope of projects that contain 'AI' and/or Robotics may seem dramatic or perhaps even scattered, the fact is the world is not actually compartmentalized. In fact, disparate systems [109] tend to interact with incredible complexity. Thus, there is enormous opportunity to learn and improve technology by embracing concepts and meanings of terms from across fields with humility (Sec. A) and as they are defined in their own setting [101]. We will proceed through several topics and related works that include Agile and other human processes, ML and 'AI', robotics, ethics, their priorities and deliberation before proceeding to describe our proposed method in Sec. 4.

3.1 Human Process Methods such as Agile Development

Maturity Models [88, 151, 162, 177] predate agile and are a related topic which attempt to classify and improve organizational capability levels, but equitable product development is out of scope of these methods.

Agile is a widely studied and adopted development method in industry and a number of academic contexts, but equity is largely out of scope [3, 51, 59, 60, 100, 175]. In the healthcare space Bonten et al. [34] have reviewed development processes, and Wilson et al. [175] have devised a valuable agile process which briefly mentions equity and incorporates all research, development, product testing, deployment, product monitoring, through product retirement. A medical product developed with a process of this kind has also successfully completed all phases of research through deployment [118]. We go beyond these first steps to propose a more comprehensive model with respect to organizational equity capabilities, and cover a different scope of project types for academic and industry research and development of 'AI' and Robotics. We also focus on day-to-day and week to week steps. This work is complimentary to Model Cards [113], Datasheets for Datasets [66], Audit Processes [142], Diversity and Inclusion Metrics [112], and the EADS Ethical Agile toolkit [94], as each are steps that can address tasks and methods in Fig. 2, particularly during phases 1-3 (Fig. 2, Sec. 4). In fact, the reverse direction is complimentary too, that is, applying agile to the aforementioned processes, protocols, and toolkits. This is because the agile processes we describe can help facilitate the execution of other tasks and methods by breaking them down into concrete and actionable chunks.

Agile also comes with significant downsides as it is typically implemented today. For example, Dancy [50] develops a cognitive model and examines a sociotechnical perspective in the context of scrum to begin describing a human-process based explanation of how antiblackness and power structures are embedded into technical infrastructure, including agile methods themselves. Babb et al. [10] examines how agile development methods could be transformed with critique, reflection, renewal, and emancipatory thinking, although in that case without consideration of identity. Kroener et al. [98] discusses the ethical agile development of crisis management software. Agile Development For Vulnerable Populations [11] discusses lessons learned across multiple projects with different populations, finding that with some vulnerable populations rapid changes to interfaces can be disruptive and some of the typically tighter agile sprint loops are more appropriate to stretch to longer time frames due to factors such as certain kinds of Institutional Review Board (IRB) approval that mean change should be minimized after a certain stage.

3.2 Machine Learning and 'AI'

Past events in 'AI' are both worthy of consideration on their own, and as a way to bring risks to the field of robotics into focus as those methods are rapidly proposed for deployment. Hundt et al. [85] reviews how recent works have shown how products using machine learning (ML) and artificial intelligence (AI) on robots can be used to exclude and outright oppress marginalized groups. Prominent work examines the flaws and assumptions in 'AI' and its surrounding sociotechnical systems [19, 23, 24, 26, 33, 58].

AI research is often conducted with baseline assumptions that render the methods ineffective across many stages. There is a common assumption that 'AI' systems actually work, and the algorithm will often be taught and used as if it works in cases when it does not; Raji et al. [141] elaborates with reasons and examples. At the time of writing incidentdatabase.ai has over 2000 additional incident reports of AI harms and counting [107]. Some of the motivating 'AI' risks and negative outcomes include software that failed to recognize people with darker skin tones [38], face recognition software that led to wrongful arrests [80, 81], software used to "detect emotions" much like the discredited pseudosciences of phrenology and physiognomy [49, 70, 110, 158], hiring algorithms discriminate against otherwise qualified workers on random attributes such as them being caregivers, veterans or disabled [64]. A number of projects have been pulled due to ethical concerns, such as Microsoft's Tay conversational system that was pulled due to training on live discriminatory data and rapidly redeploying, as well as their emotion recognition component of facial recognition AI, which can be particularly harmful to disabled and marginalized populations when facial responses vary [22, 49, 78]. The audit in Birhane and Prabhu [30] led to the tiny images dataset being retracted, and to faces to be blurred in the imagenet dataset. Fully developing a poorly founded research concept over decades, only to have it pulled due to flawed fundamental assumptions and limitations with respect to scientific validity, responsible uses, or other aspects of the project, is far more costly and might have been easier to catch at an earlier stage with the processes we describe here. Furthermore, datasets used to train ML/AI models have been found to suffer from racial and gender bias [19, 30, 90] and needlessly waste power and thus CO_2 [49, 102, 103].

Madaio et al. [104] co-designed an excellent AI-fairness checklist with input from a range of practitioners and recommended the consideration of other sectors and roles, and our work, in turn, addresses how practitioners can operationalize such a checklist.

3.3 Robotics

Robotics as a field that is quickly maturing on some metrics such as economic viability, and robots are now deployed every day in the field via drones, household robots, security robots, and more. AI is already ubiquitous [58] for some segments of the world population, and new algorithms for robotics incorporate Deep Learning [86] or Reinforcement Learning [87]. Robotic task plan authoring systems for non-experts [132, 133] are also deployed and regularly operating in the field. Projects such as “Robots for Humanity” [43] are positive examples with user centered design and input driving the research questions from the beginning. Hundt et al. [85] has an extensive discussion of the collection, processing, and inclusion of a diverse range of people in robotics data and empirical analysis. Hundt et al. [85] connects datasets to project performance and demographics, as well as leading resources to address the underlying concerns. Scheuerman et al. [147] analyzes the political values encoded into datasets.

Identity Safety Framework Hundt et al. [85] demonstrates the need for an identity safety assessment framework based on the evidence that human biases that are performed in the world and in the development process become encoded across most stages of ‘AI’ product lifecycles [161], so algorithms incorporating human data must be assumed biased until proven safe, effective and just; Hundt et al. [85] section B outlines an approach to such a framework grounded in principles of safety culture [99, 117, 119, 144]. However, if adaptive robots that incorporate ‘AI’ become readily available for mass deployment, without intervention, these robots will reproduce and even amplify the harms they observe, train upon, and use as a basis for taking action [84, 85].

Agile Development in Robotics Iterative development approaches such as agile [14, 166, 170] are growing to become one of the most common development processes for modern robotic systems. Diebold and Dahlem [56] mapped out fields using agile, which includes companies utilizing the practices in automation. Kasauli et al. [97] also examined agile in general safety critical systems.

Externalized Costs Robotics research often imagines a future with ubiquitous robots [35] and since the entirety of the world economy is part of the biosphere [52] the external costs of that one possible future needs to be accounted for due to the intensive mining, pollution, and habitat destruction required to build hardware [49] as well as the potential for global displacement for marginalized populations such as in the global south.

Policy approaches can mitigate a range of concerns, Vasiliki et al. [167] outlines ‘AI’ and social robot policy regarding the rights of the child. Brandão et al. [36] on Responsible Robotics connects the motivation and importance of a broad range of framings for the integrated advancement of robotics across topics from “user studies and philosophical inquiry, to modeling, algorithmic, and governance methods”. Pasquale [130] examines policy proposals such as a license to practice and designing algorithms to augment human capabilities.

3.4 Ethics

The ethics of the impacts ‘AI’ will have on people with marginalized identities is under-studied compared to more abstract and traditionally western ethics concerns [32, 33], and the underlying concepts can be dismissed due to terminology or a percieved position in a hierarchy of knowledge [65], rather than on the strength of their merits. Birhane [24] establishes how technical solutions are not sufficient, and outlines a way forward that centers vulnerable groups, prioritizes understanding over prediction, describes how we must question the purpose of an algorithm, what type of society the algorithms enforce, and if they should be deployed the context of a given region and cultural space. Birhane [23] argues that western technical ‘AI’ solutions echo colonial exploitation by creating dependence.

In “From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance”, Mhlambi [111] discusses the exclusion of marginalized communities from the design of systems and the use of Ubuntu philosophy for analysis. Mhlambi [111] indicates that “the relational Sub-Saharan African philosophy of ubuntu reconciles the ethical limitations of rationality as personhood by linking one’s personhood to the personhood of others”. A specific example of exclusion is the dehumanization that many members of marginalized groups experience when they encounter a method or product that does not work, treats them dismissively, never accounts for them, talks down to them, deploys deficit narratives[160, 181], or harms them [111, 181]. Ubuntu philosophy gained international popularity through Suthu and Nguni of southern Africa, from which three out of the four first African Nobel Peace Prize laureats hail [111] at the time of writing.

Robot Ethics This extends to robotics, where Zawieska [182] contends that by disengaging “with roboethics, roboticists contribute to the tacit dehumanisation process emerging in and outside of robotics. An alternative approach includes ‘lived ethics' which involves not only incorporating formal ethical approaches into the roboticists' work but also 'being' ethical and actually engaging with ethical reflection and practice." Ornelas et al. [128] describe culture as an emergent behavior, rather than a static trait of certain groups like nationality, and thus "criticised the current treatment of culture in robotics, we advanced a conception of culture based on research in cognitive science, and we have explored which robot capacities researchers should focus on to realise this revised conception of culture in robotics" [128]. Ethical problems exist in robots with integrated vision components [37, 82, 85] but the topic has received much less investigation. Gogoll and Müller [67] discusses ethical considerations in Robotics for self driving vehicles. There is also global variation in values systems, priorities, and guidelines [9, 92], and work has discussed normative implications [146]. Our work considers how to enact ethical principles into practice. Morley et al. [115] reviews related research, dividing major topics of ethical concerns into evidence that is inconclusive, inscrutable, or misguided; unfair outcomes; transformative effects; and traceability.

Abstract Digital Ethics Digital ethics [8] has become a widely studied field of research over the last couple of years. On an abstract level, recurring ethical problems of algorithms and machine learning systems specifically are mapped, concrete application areas studied, and general principles and codes of conduct to be followed proposed. At the same time, low-level technical solutions (e.g., for the provision of transparency and preservation of privacy) are developed, audit approaches suggested, and design or project management methods adopted to allow for ethics and value-centred approaches [8].

However, the abstract academic debate and the solution-oriented suggestions are still largely detached from each other and a comprehensive, easily operationalizable methodology for practice is understudied. For example, codes of conduct have been shown to have no noteworthy effect on developers' daily work, as the values stated are too abstract for concrete realizations, or are simply ignored [114]. While design approaches such as value-sensitive design [63] break down abstract values and integrate them into the product conceptualisation, they often conclude with abstract requirements that developers are expected to use as the basis for their decisions without active ethical guidance. Practitioners also admit their lack of education, skills, and interest needed to embed ethical considerations into the systems they create and responsibility is not allocated [69, 171]. It is unclear which tools need to be used in which context, how ethical conflicts should be identified and resolved, as well as how and when to think about such issues. Our paper, therefore, aims to provide a holistic approach to the complete agile development process which allows tending to ethical considerations in meetings, artifacts, and at decision points with a focus on equity issues in which identity as an important factor. A number of methods have been developed to engage with communities in a manner that has shown potential to mitigate these issues, which we will describe next.

3.5 Participatory Design, Human Computer Interaction, User Centered Design, and Interaction Design

User Centered Design (UCD), Human Computer Interaction (HCI)[160], and Interaction Design [157] have exhibited harmful impacts with respect to disabled [181], racial [74], and queer [138] identity intersections. *Participatory methods* have been proposed as philosophies for addressing and preventing these harms, in a manner that goes beyond diversity and inclusion. Specifically, *Participatory Design* [145, 150] and *Design Justice* [47] emphasize empowering impacted populations at every stage of design and implementation processes, enabling many of the different causes of 'AI' bias and harms to be detected and addressed. Harding [73] is an introductory text for qualitative methods. Harrington et al. [75] conducted participatory research showing marginalized groups, such as Black older adults, need to do cultural code switching to interact with 'AI' systems like google home, and Ostrowski et al. [129] does a year long co-design process with older adults and social robots. Birhane et al. [27] illustrates how participatory design has potential to recognize and address such limitations if implemented thoughtfully or otherwise exacerbate harms and inequities. Winkle et al. [176] presents principles and a vision for introducing feminist practices into Human Robot Interaction research. Widder and Nafus [171] conducts a study demonstrating the prevalence of 'modularity culture' in 'AI' software development which prevents harms from being addressed because they are 'out of scope' (someone else's responsibility); they propose three possible ways forwards including working within modularity, strengthening interfaces, and rejecting modularity.

3.6 Ethical priorities and deliberation

Organizations currently choose to limit investment into the capability of following their own stated deontological (imperative, must-happen) values. For example, a self driving harvester or car company might set deontological goals of: (1) do not kill humans (hard red line), (2) mitigate harm to humans, (3) maintain a path to profitability and keep customers returning (i.e. Harvest/drive efficiently), (4) be as eco-friendly as possible provided (1), (2) and (3) are met, then non-imperative priorities: (5) Keep customers happy (6) Maintain a happy and healthy workforce.

We argue that the evidence [20, 38, 58, 71] detailed in the Introduction (Sec. 1) and Related Work (Sec. 3 indicates that organizations are not yet capable of meeting even their basic deontological goals. This is particularly true of Robot and/or 'AI' algorithms that suffer from small sample and other critically flawed methodolgical assumptions. We contend that organizations also do not possess the capability to recognize and then address failures, a necessary prerequisite to meeting their deontological goals. Furthermore, organizational pressures lead to priority inversions, for example (3) maintaining a path to profitability will sometimes in practice be prioritized over (1) not killing humans and (2) mitigating harm to humans [79].

The EDAP [94] protocol for ethical agile serves as an example of the complexity of ethical deliberation. EDAP makes several assumptions: (1) organizations are capable of meeting their basic deontological goals in correct priority order, (2) power dynamics are not a factor and individuals are "ideal actors"; (3) It assumes there is always a technical solution to the problem at hand. The evidence (Sec. 1, 3) we have already explored in this work indicates that these assumptions are a limitation of EDAP, and that most organizations are not yet capable of meeting even their primary deontological company goals in the general case. In other words, EDAP makes excellent contributions in their topic, but assumes greater organizational capabilities than actually exist.

We, in turn, propose methodological concepts to address those limitations. One of our contributions is to devise mechanisms to improve the scrum process so that organizations can build the capabilities and metrics necessary to get closer to meeting their primary deontological goals, and thus respect identity in practice.

3.7 Accountability

Accountability is a particularly pernicious challenge because individuals with organizational power have incentives to override real concerns. For example, authors informed their method has bias proven in a third party audit declined to similarly evaluate their latest methods, despite similar risks. Wieringa [172] surveys the literature of accountability, and Raji and Buolamwini [140] outlines the effectiveness of audits on real products. A strong project governance model can facilitate accountability, and Jernite et al. [91] is a practical large scale case study of more inclusive data collection and governance, where data was collected and stored by people from the regions to which it applies.

4 METHOD

Here we describe our approach to our More Equitable Agile Research and Development method (Fig. 1, 2) for which elements are based on Hundt et al. [85], Wilson et al. [175], and the Patient-led Research Collaborative [123] and adapted to a broader range of 'AI' and robotics contexts. The method is scoped to the duration of a single project as illustrated in Fig. 1 and 2. An essential component of this method is to build teams that are capable of adapting and maintaining an inclusive organization, culture and structure. This goal tightly interlocks with the goals for the artifacts, products, and/or research results (Fig. 1). A core reason for this are that a team which forces employees out when their communication style, culture, or medium differs due to their disability status, race, gender, or other identity factors is unlikely to [125, 136] conduct research or create a product that meaningfully accommodates those criteria.

4.1 Project Phases and Decision Points

Fig. 2 outlines the phases of our proposed research and development process. Our intention is to provide a toolkit that should be customized for a specific project (Fig. 1), so not every step will be appropriate for every project.

Phase 0: Project Identification Is when the initial project purpose, scope, budget and other criteria is established, the core team defined and funding secured, be it through corporate channels, grants, or startup funds. The list of items above phase 0 in Fig. 2 are preliminary preparation steps that can be considered to assess project viability and potential.

Phase 1: 'Sprint' Is the central research and development time allocation loop during which Equity Tasks and Credit, Development, Experiments, and Testing are typically assigned. One key difference in this phase for many projects that are developing physical hardware is that aspects of early development cannot be as dynamic as mobile health [175] applications. This is due to the additional time required to design, build, and validate that physical hardware as part of the process. However, significant progress can often be made with a combination of simulation, mock-ups, and off-the-shelf proof-of-concept hardware to be replaced with

Roles	**Meetings (Ceremonies)**	**Artifacts (documents or other outputs)**
Product owner	Sprint Planning	Product Backlog
ScrumMaster	Sprint Review	Sprint Backlog
Team	Daily scrum meeting	Burndown Chart
Key Stakeholders	Sprint Retrospective* (Sec. 4.3)	**Resource Usage & Env. Impact Records** (Sec. 4.6)
Participant	**Governance Meeting** (Sec. 4.3)	**Ethical + Equity Feedback and Decision Records** (Sec. 4.6)
		Participant-led Project Assessment Scorecards (Sec. 4.6, C)

Table 1. The scrum process from [46] (Sec. 2), with new proposed steps in bold. * The sprint retrospective has **new ethical & equity steps**.

final production hardware at a later stage. While the quality and intent can vary, Design and Research are intertwined tasks being done simultaneously as any product or research methods are being developed.

Phase 2: Limited Pilot/Beta Testing This phase begins once there is an initial prototype that has a usefully testable product or product component iteration. In robots with AI, this will involve an 'AI' model which should be documented via Model Cards [113] that characterize the properties, limitations, and performance with respect to various demographics.

Phase 3: Controlled Trial Evaluation This is the phase in which core experiments and testing for a new AI, robot, or application are run. The experiments and method should be designed with input from marginalized populations and the experiments and testing itself should incorporate input of a larger and highly diverse population when it is appropriate for the method. An example exception would be a one-off design to meet the needs of one person [77]. For larger projects it can be appropriate to do an Algorithm Audit [48, 142] and Algorithm Bug Bounty to detect and prevent harms at larger scale deployment. As Wilson et al. [175] indicates, A/B testing is an appropriate way to do evaluations at this phase, however we note additional precautions and broader demographics are required. We also incorporate the human override, pause, and wind down step to ensure the testing mechanisms are more inclusive [71] (Sec. 4.5).

Phase 4: Deployment, Publication, Ongoing Assessment This is the phase where the 'product', broadly construed, is out in the world. The product might be physical hardware, a service, a research paper, datasets, other artifacts, and so on. In most cases with ongoing work on the topic, there should be regular assessment to ensure it is functioning well and meeting people's needs, with particular attention to the effects on marginalized populations, to ensure it is a net benefit. The system should strive to maintain its status and adapt to ensure it is safe, effective, and just.

Project Wind Down or Retirement If a project is wound down at early phases, we recommend a technical report or research paper, as appropriate, that examines the reasons this decision is being made the outcome, and data when appropriate in a manner to benefit the community. An example of one positive step for that case and context in a real-world project is *Lessons Learned in Designing AI for Autistic Adults*[16], which was wound down after receiving negative participant feedback on their 'emotion recognition AI'. Another example is the Makani airborne wind energy platform project, which released a significant quantity of technical materials and a patent non-assertion pledge upon project wind down [4]. However, in other contexts such as some cases of marginalized communities public releases might not be an appropriate wind down strategy. If the case is a less complex pause and rework of some aspects of a project, the reason, options, and decisions should be recorded in appropriate decision logs for revisiting and reference in the future.

Gating Questions A series of sample Gating Questions [85, 175] to be adjusted for particular applications then applied as per Fig. 2. Gating Questions can serve the dual purposes of preventing unnecessary work and missteps:

> We recommend that future projects ask questions through technical, sociological, identity (which refers to factors such as race, indigenous identity, physical and mental disability, age, national origin, cultural conventions, gender, LGBTQIA+ status, and personal wealth), historical, legal, and a range of other lenses. Such questions might include, but are not limited to[2]: Is a technical method appropriate? Is there a simpler approach? [175] Whom does our method serve? Is our method easy to use and override? Have we respected the principle of "Nothing about us without us"? Is the data setting (Sec. A) appropriate? Does our method empower researchers and the community with respect to equity, justice, safety and privacy needs? What are the negatives and positives? Does the evidence show our method addresses the problem within equity and environmental constraints? Does the scope of method evaluation address the scope of algorithm inputs? Do any concerns indicate that we should pause, rework, or wind down the project?
> – Hundt et al. [85]

[2] The questions are sourced from Hundt et al. [85] and are inspired by Wilson et al. [175] Fig. 3.

Gating questions are appropriate to ask during transitions between phases including self-transitions, for example, from one iteration candidate to another. If the questions identify an area concern it might be appropriate to initiate a pause or stop process. We will describe the pause and stop process next.

Pause/Stop Process A pause and stop process should actively be supported for a very large proportion of stakeholders as is customary for kanban [159], or perhaps even to a greater extent when appropriate. In particular, we introduce the need for identity and demographic-based factors to be considered in the case of 'AI' projects. When a pause is triggered there should be a corresponding safety and ethical review appropriately scaled to the problem, which on small teams and projects in the case of minor concerns can be a brief discussion with minutes outlining the issue, options, potential courses of action, the course of action taken, and later the result of followup; however the possibility of proceeding to a stop should be a serious option. Pauses and stops could also be conducted for particular components of a system, for example, a vehicle's self driving lane keeping and steering functionality might in testing prove reliable on highways but not residential areas, so the feature might be geographically restricted until reliability is adequate. One limitation of this step is that it requires an inclusive and supportive safety culture [85, 99, 144] to be effective, as powerful organization members can easily override it.

Self Changing Process The process we outline here is subject to iterative improvements, as is method development, the meta-process of how methods are developed, the Agile process itself (an Agile Manifesto [15] mantra is to place individuals and interactions over process, we greatly broaden the range of individuals to consider), and other aspects of methodological practice. This matches the scientific consensus for STEMM inclusion [125] p148. Concretely, the Pause and Stop steps as well as the Retrospective Meetings outlined below are occasions specifically designed for the proposal of changes to the process, such as removing adding, or modifying the project or the process, and for the decision making necessary to make such changes. Furthermore, a governance model (Sec. 4.3) and metrics (Sec. 4.6, C) are designed to democratize the project direction and get input from new perspectives, the mechanism itself will be designed to support people to propose positive changes then get them improved and implemented.

4.2 Meetings, "Ceremonies" in scrum terminology

Brief daily meeting ("Standup") harm prevention step: In daily meetings each person can spend one minute to share something they learned about human populations, policies, DEI related research, auditing methods or papers they learned about, identities other than their own, etc. as a part of a continuing education process. Ethical concerns about the project or how the input of a broad range of stakeholders is being integrated can also be briefly coordinated during this meeting.

Sprint Retrospective and Planning Ethical and Equity Steps The retrospective and planning meetings can consider a checklist of questions about specific identity groups to mitigate negative outcomes:

(1) How do items for the last and next sprint affect minoritized groups? Consider each identity factor such as age, race, physical ability, mental ability, age, national origin, experience, cultural conventions, physical appearance of varied regions of the world, and combinations thereof, with particular attention to highly minoritized combinations.
(2) If answer for one group is "it won't", assume the process has missed something and create a priority sprint task to quantify it, gather information and feedback from such groups.
(3) What action, if any should we take based on scorecard assessments (Sec 4.6, C), audits, AI-fairness checklists [104], or other evaluations (Fig 2)?
(4) What have we missed?
(5) Listen to representatives who self identify at the meeting, while giving space and being considerate.
(6) Discuss what each team member has learned from their education component and what will be studied next.

Additionally, compare the findings of different assessment methods, participant-led project scorecards (Sec. 4.6, C), or add tasks to seek externalized costs [112] and assessment methods (e.g. [85]) that have not yet been evaluated on this project to find undiscovered method and/or premise flaws that need to be addressed. Next we will cover governance models which have potential to mitigate and even prevent many negative outcomes, while shifting project directions towards better positive outcomes.

4.3 Democratizing Governance Models

We examine a participant-led framework as a concrete illustrative example of governance models for the purposes of this proposal. However, particular projects might have significant and well-justified differences ranging from participatory methods (Sec. 3.5) to distributed governance [91], or more traditional operational mechanisms.

In our case, we consider a planned project governing committee consisting of 50% plus one participants with the remainder consisting of traditional stakeholders such as product owners, management, researchers, and/or developers. Participants are drawn from a broadly construed definition of stakeholders, including potential users and other directly and indirectly impacted demographics, broadly construed, excluding traditional stakeholders. The governing body can then operate according to organization-defined bylaws that provide significant powers to delegate project operations, make major ethical decisions (>50% vote) and record the outcome, set project direction, and crucially define and initiate partial or full pauses (>33% vote, or just 1 person in safety-critical circumstances), stop (>50% vote) and wind down (>66% vote) operations.

An appropriate schedule for the governing committee might be to meet periodically every two to four weeks (once each sprint, Sec. 2) with asynchronous communication and additional essential-purpose meetings as needed. These meetings can serve as opportunities to make and log ethical and governance decisions brought up at other meetings, collect updated participant-led project assessment scorecard results and policy change proposals to increase the overall score (Sec. C), try out demos and evaluate for governance concerns, propose and vote on changes to policies, adjust project direction, approve accountability mechanisms, approve pause and wind down policies, and delegate further work to individuals teams or subcommittees to ensure both more equitable outcomes and timely project progress. Then when the project is either eventually completed or otherwise reaches an end point it can initiate the smooth wind down plan for final project retirement.

4.4 Equity Context

When considering a project it is important to think about the equity context and data setting (Sec. A), including who might be affected, where a project will be deployed; to think on a large scale and small, the long term and short; and to add new lenses of analysis to examine problems from new angles. It can even be particularly beneficial to learn about and use analysis tools an individual or organization might disagree with to expand their perspective. Familiarity with a range of perspectives and methods of analysis can expand the capacity to prevent, detect and mitigate problems, even if any one particular method is not put into action.

One weakness common in 'AI' and Robotics projects is a demographic factor where, due to the high cost of robotic systems, alongside their tendency to be portable sensing systems, benefits and power are likelier to accrue to wealthier demographics [183]. This interplay of human systems and humans themselves is a particularly difficult and interesting problem. For example, 2.9 billion people remain offline [154] as of 2021. According to the OECD, an organization of mostly wealthy countries, "Around one-quarter of adults in all participating countries have no or only limited experience with computers or lack confidence in their ability to use computers" [122]. This reality motivates some of our gating questions (Fig. 2), because the best solution to a problem might not involve a technical approach, Robots, or 'AI' at all.

Another key illustrative example comes from the experiences of the Disability community [77, 89, 178]. The Disabled population is the world's largest minority representing about 1 in 6 people in the world [127], and is also an enormously diverse and capable group with varied and sometimes even conflicting essential human needs. This immense population means that all but the smallest projects will both have disabled people on the team and interacting with the project's outputs and artifacts. Furthermore, the expertise of marginalized populations is often dismissed [181], or even taken without proper credit [89] (Sec. 4.5, 4.6). Liz Jackson's *Disability Dongle* [89] concept elegantly captures the risk of ignoring such expertise, leading to "well intended and elegant, yet useless solution[s] to a problem we never knew we had". They illustrate disability dongles with stair climbing robotic wheelchairs that, in most circumstances, cannot match the effectiveness of well-established accessible architectural design [77, 89]. This case and the elder care robots in the Introduction (Sec. 1) is also where the influence of human power dynamics can come into play. Wheelchairs failures can in some circumstances impose devastating safety, financial, and life costs on marginalized individuals, even grinding their lives to a halt when essential mobility is unavailable. Modifications to adapt the built environment to be accessible via a social model of disability is often a more durably and inexpensively accessible approach that supports a broader range of people than designating the 'nonconforming' disabled person to be a problem and attempting to adapt them to the world [77, 89].

Inclusion Guidelines Experiment design can occur in multiple phases, and we recommend strengthening experimental design to be more inclusive of the world population across forms identity, as a claim that a method generalizes does not comport if the premise simply excludes millions or billions of people. Guidelines are available for a wide range of identities [6, 7, 139, 148] and could be leveraged in experimental, algorithm, meeting, product, and other design. We also propose scorecards adapted to the R&D context so that a range of participants can periodically evaluate project performance over time (Sec. 4.6, C). Equity impacts also extend to externalized costs such as Environmental Impacts, which we discuss below.

Environmental Impact All of our human processes are completely dependent on environmental processes as part of one larger system [52] among other externalized environmental costs are essential considerations to Equitable R&D, and both Robotics and 'AI', and the true extent of current visions of ubiquitous robots [35] and 'AI' carry real risks of devastation far in excess of the benefits with particularly harsh disproportionate impacts on marginalized populations. ML systems to estimate environmental consumption are emerging, and tracking of resource use like electrical consumption can be achieved with built-in or small tracking attachments to equipment [5, 61, 76, 102]. The full extent of environmental risks and impacts remains outside the scope for this paper, but is an essential topic for future study and quantification.

Concretely thinking about the full variety of people that actually exist, externalized costs, their crucial equity contexts, recognizing others' expertise, and to genuinely onboard and adapt to meaningful critiques is an essential opportunity to develop methods that truly generalize in the sense of supporting people across the full diversity of human needs.

4.5 Sprint Task Blocks

Continuing Education A promising and powerful consensus on best practices [39, 124, 125, 136] for steps established to improve organizational supports and outcomes with respect to Diversity Equity and Inclusion in Science, Technology, Engineering, Math, and Medicine (STEMM) is emerging. A key aspect of this is the known limitations and areas of active research that will need to be regularly updated over time.

The current lack of effective organizational capabilities, training, and meaningfully operational practices on Diversity, Equity, and Inclusion in STEMM means there is a lot of work to do [136] to build them. Therefore, we modify the agile process to directly incorporate a percentage of continuing education to develop such capabilities both with respect to internal organizational supports and to support better understanding of the equity context for the project itself (Sec. 4.4, 4.6, C). An example quantity of time is two hours a week, which can be allocated for studying learning about the tools of the trade, best practices for the application at hand, and better practices for supporting a diversity of people who are either on project teams, or may be impacted (see Sec. 1, 4.4, B, 4.7). Example starting points for organizations are the best practice reports [39, 124–126, 136]. Example starting points for 'AI' include Data Feminism [58] and ethics.fast.ai.

Credit for traditionally unpaid or uncredited labor. We propose a pool of sprint task blocks and job performance credit be made available for team members to claim on an as-needed basis to be used for self-determined purposes that need not be disclosed. The task blocks and credit are a way to bridge equity gaps such as for uncredited or uncompensated labor [125] (Sec. 4.7). This also includes some Disabled team members' need for additional time, in the sense of crip time (Sec. A), which is often essential due to differences in their bodymind (body and mind as one). We propose a large undisclosed blocks, group-level blocks as successfully demonstrated by Wu [180], or individualized time points be used (without penalty) for anonymous use, and extend its availability to everyone. Conceptually, another way this might be viewed as a kind of internal organizational work time insurance scheme, or even a by-the-hour version of unlimited paid time off— with credit. This has potential to help account for the manner in which, due to the normal course of human life, some people will need to engage in a disproportionate amount of uncredited but important work, while others receive credit for others' uncredited labor, and these factors should be accounted for in a manner that ensures promotions raises and job performance reviews are less affected by factors outside marginalized individuals' control [17, 40, 45].

We also consider counter-arguments to this concept. For example, it might lead to the indirect tracking of personal information, which they might not desire to be prominently visible, so careful thought about privacy and other intersectional concerns of operationalizing the concept will be needed for such a proposal to mature and be adopted. A number of concerns were raised when some companies introduced unlimited paid time off, which has its complexities and limitations, but has also been successfully maintained for many years at some organizations.

Human Overrides and The Big Red Button Many robots have a large red security stop button that will halt the robot in its tracks. This concept was made famous by the Toyota Kanban [159] system's centering of respect-for-humans, where any worker can halt production (known as Jidoka, see Sec. A). We propose an analogous override for all of, or parts of, the development process itself. Generally, Human overrides should be designed into every stage of the process, including the human research process. The process should also be designed to be iterative and deliberately consider marginalized community members. Human overrides can be considered for incorporation at several different levels such as on the physical machine, in the development process, in the data collection process, and everywhere else that harm can be introduced in the 'AI' lifecycle [161].

Human overrides are also an important component of physical product, as the diversity of human requirements and situations for public robotic systems means mechanisms are required to get robots out of the way, teach them new tasks, and successfully interact across the wide range of human capabilities and needs. For example, how will a sidewalk robot accommodate, communicate, and prioritize access to navigation resources of space, time, and equipment of a range of pedestrians using guide dogs, scooters, wheelchairs, canes, walkers, and so on? A process for incorporating modifications based on real detected issues on an ongoing basis is also necessary. For example, if the robot encounters someone traveling backwards in a wheelchair [164] and fails to stop, the mere existence of anomalies is unsurprising, so a cohesive iterative design update process that is designed in from the start should be activated to account for the situation in the future. Task work such as Amazon Mechanical Turk provides another compelling ethical example with critical technical impacts that we elaborate on in Sec. B.1.

4.6 Artifacts: Ethical + Equity Feedback and Decision Logs/Record; Participant-led Project Assessment Scorecards

Decision points can be recorded through a set of logs created and stored in a manner that is appropriate for the organizational complexity, in a text file or task tracking tool for very small organizations, or internal custom project management tools for larger organizations. The record should include: Checklist steps followed, Ethical decisions, Reasoning for decisions, Equity feedback, Changes based on feedback, Why those changes were made/not made, no-go decisions that are made, and why.

Equity log: Documentation and records of potential concerns across a range of demographics, a record of the reasoning, and records of (potentially anonymized) sources of input gathered both through interactive communication and through relevant written experiences. For example, an equity consideration of a large-scale self driving taxi or public transport system might be advised to consider the possibility of some demographics being excluded or denied access to transport, directly or indirectly, due to regional shifts or reductions in access to other forms of transport. Another potential downside is the exclusion of people due to transport needs such as public transportation and wheelchair accessibility. Another example for an equity log is to record the impacts on stakeholders, participants, and other demographics that considers externalized costs, a record of the specific feedback mechanisms for each group, as well as policies designed to account for potential harms due to power dynamics.

Ethical and environmental log: each record ethical and environmental items considered, respectively, possible approaches to address any issues raised as needed, mitigating steps chosen, and the outcome of those mitigating steps logged. An environmental log can also track resource impacts with tools such as carbontracker [5] for smaller projects, or more sophisticated metrics as appropriate, and resource consumption can be reported in public research and development materials.

Participant-led Project Assessment Scorecards: Our Participant-led scorecards are a measuring tool designed to assess organizational capabilities with respect to populations that might be impacted by a project directly or indirectly who are onboarded as participants to provide feedback on the capabilities and effectiveness of the Equitable Research & Development Lifecycle.

Our scorecards are directly adapted from the proven Patient-led Research Scorecards [123] created by the Council of Medical Specialty Societies (CMSS) and Patient-Led Research Collaborative [54, 55]. We reworked the scorecards to expand their scope beyond patient-led domains to multi-domain and interdiciplinary projects with a goal of measuring Participant-led organizational capabilities. Each scorecard is specified in Sec. C, and the cards cover assessments of: the burden the project imposes on participants, project governance, R&D organizational readiness (e.g. companies, universities, labs, nonprofits, or units thereof), integration of Participant priorities into the research process, and our new addition the Researcher and Staff Support scorecard. Each scorecard has several items that get rated from -2 to 2 denoting non-collaboration, minimal collaboration, acceptable collaboration, great collaboration, and ideal collaboration, respectively.

The scorecards are intended to be provided to and completed by organization members, stakeholders, and participants. Another reasonable option to consider not included on the cards is to add 'I don't know' and/or 'Not Applicable' options for items when people don't have information of other's perspective or don't individually relate to an item.

It might also be worthwhile to consider clarifying the following two points of conflict in the survey design. The first regards to whom survey answers are in relation to, such as themselves, another person, or something else. For example, for the Participant

Burden Scorecard Sec. C.1 Accessible Engagement, and the Researcher and Staff Support Scorecard Sec. C.5, while some participants, researchers, and staff might each need accommodations in some circumstances, others might not. Therefore, ensuring the scorecard can capture the difference among no accessibility need in that context, having such needs met, and having an unmet need might provide valuable actionable information. Second, there are currently some scenarios that are not accounted for, such as cases where open data or publishing results might entail a high risk of non-consensual misuse or abuse [105], as forewarned by scholars like Birhane and colleagues [29, 31], and we have integrated descriptions of mitigated vs unmitaged net negative impacts from open release into the Integration into Research Process Scorecard's item on Products, Artifacts and/or Publications (Sec. C.4).

When both traditional stakeholders, like developers, as well as a broader diverse range of participants provide feedback and submit participant-led project ratings that are very different from traditional stakeholder expectations, that provides a signal indicating where process improvements are needed. We also recommend the scorecard submission process also provide a way to elaborate in more detail in writing.

We speculate that it is likely that most active projects at the time of writing would get surprisingly high rates in which there is at least one item with a non-collaborative -2 rating, a rate which should be as close to 0 as possible. Nonetheless, we suggest that even a single score of -2 should be grounds for reflection and at least small action items to work towards determining and then mitigating the underlying cause whenever possible. Scorecard results should also be added to the equity log whenever feasible, as should the evaluation process, findings, steps to mitigate the problem, and the outcome. We also suggest organization members, stakeholders, and participants all periodically complete the scorecards too to track progress over time, and have the option to anonymously submit an additional scorecard at a moment's notice. Furthermore, provided the sample size is large enough, the scorecards can be analyzed according to likert survey statistical best practices [149] or via bayesian methods [44, 131, 165]. One limitation of the scorecards is that small projects might inadvertently deanonymize submissions due to unique incidents. A second limitation is the academic language style of the scorecards, so future work should consider developing a plain language or easy read version [6, 7].

Taken together, our Equitable Agile R&D process knits together a tapestry of capability building methods and harm mitigation strategies with the potential to get positive projects on track, as well as to catch and redirect effort away from harmful projects that would otherwise be likely to fall through the cracks. Next we will elaborate on the limitations of our method as a whole followed by an outline of future work and our conclusion.

4.7 Assumptions and Limitations

The processes and approach outlined in this paper includes a number of assumptions and limitations. We assume organizational and team buy in on equity concerns. One clear vulnerability of this approach is any ethical and other human processes are vulnerable to simply be overridden by individuals with organizational power and a focus on expediency. This is a particularly significant weakness because it is so uncommon at the time of writing for team members on the ground to be adequately equipped to assess equity issues [136].

Another significant challenge will be equipping teams to be genuinely willing to accept and take action consistent with communication or other signals indicating the best outcome would be to make a project pause or wind-down decision. Furthermore, there is a risk of backlash and inertia derailing more equitable R&D, regardless of the evidence. Examples from the range of possible reasons for rejecting such signals might include attachment to one's work or the sunk cost fallacy. Furthermore, processes we outline here are vulnerable to internal organizational discrimination against people based on identity, who are disproportionately likely to be forced out of organizations for calling attention to problems or injustices [1, 125].

Many of the individuals willing to embrace working towards more Equitable R&D will require step-by-step guidance and resources relevant to their field before they can meaningfully get started and build broader organic engagement. Even good-faith engagement and contributions can be hindered by human social network challenges. Many teams are likely to be hindered by a lack of connection to and understanding of communities that should be included. Some people will also have difficulty interacting with those communities in a manner that leads to desirable outcomes [89, 135, 153]. In cases where individuals meaningfully commit to working towards more Equitable R&D, many will remain ill equipped to build a diverse team or seek feedback, and even with an improved process some populations will still be excluded.

There can even be conflicting access needs between groups as identity factors vary, so while our proposal has potential to lead to improvements, there is no complete 'solution' for all situations. We also recognize the impossibility of representing all views.

People's lived experiences vary enormously, and population level discourse or trends do not represent all cases. Additionally, while case studies are a valuable sample, the full set of world experiences are extremely diverse and cannot be fully captured. There are cases and views that we do not know we don't know, and thus we welcome constructive critique and seek iterative methodological refinement.

Another potential critique of this work is that aspects of it repackages and perhaps even co-opts well-understood and widely practiced methods that have been available for decades in various communities and fields [93]. Another is that powerful actors might dismiss the approach or co-opt the system themselves, sometimes even doing so with good intentions [93].

Workplace and Education Environment A prerequisite to the development processes we describe here is a workplace environment that buys in and is willing to make coordinated focused effort [136] at multiple organizational levels (Fig. 1). It will not be effective in a toxic workplace environment, a detrimental problem that is dismayingly common and with proven persistence in both Industry and Academia [24, 93, 124, 125, 136]. For example, a National Academies Press Scientific Consensus report indicates "substantial research demonstrates that implicit and explicit biases discourage women from entering STEMM careers or influence their decision to leave STEMM after beginning their careers. These factors include a spectrum of explicit and implicit biases, as well as structural and interpersonal interactions that impede women's progress. These are factors across the career life cycle." [125] The process we describe is also vulnerable to a powerful stakeholder, such as an Executive or Principal Investigator simply squashing the system with their authority. New legislative frameworks and incentives might need to be developed to address such priorities, and we encourage legislators to advance and democratize deliberative processes, as in governance, but specific legislative recommendations are out of scope for this work.

The actual internal development process and organization team should also operate as equitably as possible, with cognizance of each others humanity and proper coordination and accommodations for the needs of each team member [58, 62, 125]. However, we recognize that individuals are human and meaningful changes can take time, so a pragmatic goal is to ensure ongoing iterative improvements in team capability for equity of interactions and the working environment. Broader discourse on this topic is out of the scope of this work, so we refer the reader to recent scientific consensus reports on evidence based best practices with respect to Responsible Computing [126] Research, reducing the sexual harassment of women [124], reducing the underrepresentation of women [125] in STEMM, mentorship [39], and the Universal Design for Learning Guidelines [41]. Each of these resources also considers identity as an important factor in their analysis and recommendations.

Future Work Questions to consider for future work include: How much engagement should there be across identities worldwide and what are the criteria to determine such decisions? How to shift more power to minoritized groups? How to guide educating the team, perhaps through self study or a specific education package? How to onboard people from minoritized groups and other groups? How to start seeking community feedback find people in other countries or groups? How to highly prioritize and integrate community feedback? How to change budget priorities? How best to prioritize at different budgets/team sizes? How to reach out to communities and seriously incorporate feedback? There is also potential for policy and legal changes at the level of institutions and government to more effectively mitigate potential concerns, such as required integration of independent ethics review steps and audits, or even a license to practice [130]. We leave this to future work.

5 CONCLUSION

We have proposed a toolkit that aims to serve as a step towards equitable agile research and development of 'AI' and Robotics, so organizations can build the capability to examine their particular problem and choose the right tools to ensure more equitable outcomes to their project. We draw connections between the topics of research, the methods being practiced, and people being included or excluded as part of the human process of developing research or products that contain 'AI' or Robotics. The applicability of particular tools in this toolkit will vary across projects and they should be chosen on a case by case basis. We introduced concrete steps such as ethical and equity discussion components of meetings and retrospectives, participant-led project assessment scorecards, a participant-led governance model, the time blocks for general use, and the time blocks for building organizational capabilities through education and experimentation. We hope this work will support projects and communities to thrive.

ACKNOWLEDGMENTS

We thank William Agnew, Anne Washington, Kari George, Di Wu, Christopher Dancy, and those not listed here for the valuable feedback and insightful discussions during various stages of this work. We thank Wilson et al. [175] for their compelling visualization of the traditional Agile System Development lifecycle as well as their mobile device clinical Health (mHealth) Agile Development & Evaluation Lifecycle. We adapted and expanded upon these under their Creative Commons Attribution 4.0 License (CC BY 4.0) to develop our Fig. 2's More Equitable Agile Research and Development Lifecycle. We thank the Council of Medical Specialty Societies (CMSS) and Patient-Led Research Collaborative (PLRC) for their brilliant Patient-led Research Scorecards [55, 123], which we directly paraphrased and then expanded from a medicine-specific assessment focus to scorecards capable of assessing a much broader range of application domains, identities, development artifacts, and licensing criteria. To the best of our knowledge, Sec. C (the scorecards), and the latex code for their table formatting, are the only section modified with Large Language Model (LLMs) based tools, specifically ChatGPT and Bing Chat. Finally, the Researcher and Staff Support Scorecard in Sec. C.5 is our not just partially new, but completely new addition to the scorecards.

REFERENCES

[1] Sara Ahmed. 2021. *Complaint!* Duke University Press, Durham. https://doi.org/10.1515/9781478022336

[2] Malek Al-Zewairi, Mariam Biltawi, Wael Etaiwi, and Adnan Shaout. 2017. Agile Software Development Methodologies: Survey of Surveys. *Journal of Computer and Communications* Vol.05No.05 (2017), 24. https://doi.org/10.4236/jcc.2017.55007

[3] Scott W. Ambler. 2008. Agile Software Development at Scale. In *Balancing Agility and Formalism in Software Engineering*, Bertrand Meyer, Jerzy R. Nawrocki, and Bartosz Walter (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12.

[4] Mark Anderson. [n.d.]. Exclusive: Airborne Wind Energy Company Closes Shop, Opens Patents. https://spectrum.ieee.org/exclusive-airborne-wind-energy-company-closes-shop-opens-patents accessed July 30, 2022.

[5] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. arXiv:2007.03051.

[6] Autistic Self Advocacy Network (ASAN). [n.d.]. Holding Inclusive Events: A Guide to Accessible Event Planning. https://autisticadvocacy.org/wp-content/uploads/2019/05/Accessible-Event-Planning.pdf

[7] Autistic Self Advocacy Network (ASAN). [n.d.]. One Idea Per Line: A Guide to Making Easy Read Resources. https://autisticadvocacy.org/wp-content/uploads/2021/07/One-Idea-Per-Line.pdf

[8] Mona Ashok, Rohit Madan, Anton Joha, and Uthayasankar Sivarajah. 2022. Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management* 62 (2022), 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

[9] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (2018), 59–64.

[10] Jeffry S Babb, Jacob Nørbjerg, David J Yates, and Leslie J Waguespack. 2017. The Empire Strikes Back: The end of Agile as we know it? *Communications of the Association for Information Systems. Selected Papers of the IRIS* 8 (2017), 43–59.

[11] Marcos Baez and Fabio Casati. 2018. Agile Development for Vulnerable Populations: Lessons Learned and Recommendations. In *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Society* (Gothenburg, Sweden) *(ICSE-SEIS '18)*. Association for Computing Machinery, New York, NY, USA, 33–36. https://doi.org/10.1145/3183428.3183439

[12] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[13] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers, and S. Šabanović. 2020. *Human-Robot Interaction: An Introduction*. Cambridge University Press. https://books.google.com/books?id=YibUDwAAQBAJ

[14] BERTHOLD Bauml and Gerd Hirzinger. 2006. Agile robot development (ard): A pragmatic approach to robotic software. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3741–3748.

[15] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. 2001. Manifesto for Agile Software Development. http://www.agilemanifesto.org/

[16] Andrew Begel, John Tang, Sean Andrist, Michael Barnett, Tony Carbary, Piali Choudhury, Edward Cutrell, Alberto Fung, Sasa Junuzovic, Daniel McDuff, Kael Rowan, Shibashankar Sahoo, Jennifer Frances Waldern, Jessica Wolk, Hui Zheng, and Annuska Zolyomi. 2020. Lessons Learned in Designing AI for Autistic Adults. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 6 pages. https://doi.org/10.1145/3373625.3418305

[17] Donica Belisle and Kiera Mitchell. 2018. Mary Quayle Innis: Faculty Wives' Contributions and the Making of Academic Celebrity. *Canadian Historical Review* 99, 3 (2018), 456–486. https://doi.org/10.3138/chr.2017-0108 arXiv:https://doi.org/10.3138/chr.2017-0108

[18] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[19] Ruha Benjamin. 2019 - 2019. *Race after technology : abolitionist tools for the New Jim Code*. Polity, Cambridge, UK ;.

[20] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 375, 19 pages. https://doi.org/10.1145/3411764.3445498

[21] James Bessen, Stephen Michael Impink, and Robert Seamans. 2022. The Cost of Ethical AI Development for AI Startups. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 92–106. https://doi.org/10.1145/3514094.3534195

[22] Sarah Bird. [n.d.]. Responsible AI investments and safeguards for facial recognition. https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/

[23] Abeba Birhane. 2020. Algorithmic Colonization of Africa. *Scriptorium* 17, 2 (2020), 389–409.

[24] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), 100205. https://doi.org/10.1016/j.patter.2021.100205

[25] Abeba Birhane. 2021. The Impossibility of Automating Ambiguity. *Artificial Life* 27, 1 (06 2021), 44–61. https://doi.org/10.1162/artl_a_00336 arXiv:https://direct.mit.edu/artl/article-pdf/27/1/44/1925148/artl_a_00336.pdf

[26] Abeba Birhane and Olivia Guest. 2020. Towards Decolonising Computational Sciences. *Kvinder, Køn and Forskning* 2 (2020), 60–73. https://arxiv.org/abs/2009.14258

[27] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3551624.3555290

[28] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 173–184. https://doi.org/10.1145/3531146.3533083

[29] Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. 2023. On Hate Scaling Laws For Data-Swamps. arXiv:2306.13141 [cs.CY] https://doi.org/10.48550/arXiv.2306.13141

[30] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

[31] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv* abs/2110.01963 (2021). https://arxiv.org/abs/2110.01963

[32] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 948–958. https://doi.org/10.1145/3531146.3533157

[33] Abeba Birhane and Jelle van Dijk. 2020. Robot Rights?: Let's Talk about Human Welfare Instead. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 207–213.

[34] Tobias N Bonten, Anneloek Rauwerdink, Jeremy C Wyatt, Marise J Kasteleyn, Leonard Witkamp, Heleen Riper, Lisette Jewc van Gemert-Pijnen, Kathrin Cresswell, Aziz Sheikh, Marlies P Schijven, and Niels H Chavannes. 2020. Online Guide for Electronic Health Evaluation Approaches: Systematic Scoping Review and Concept Mapping Study. *Journal of Medical Internet Research* 22, 8 (2020), 1–22.

[35] Martim Brandão. 2021. Normative roboticists: the visions and values of technical robotics papers. In *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*. 671–677. https://doi.org/10.1109/RO-MAN50785.2021.9515504

[36] Martim Brandão, Masoumeh Mansouri, and Martin Magnusson. 2022. Editorial: Responsible Robotics. *Frontiers in Robotics and AI* 9 (2022). https://doi.org/10.3389/frobt.2022.937612

[37] Joy Buolamwini. 2018. When the Robot Doesn't See Dark Skin. https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html

[38] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[39] Angela Byars-Winston and Maria Lund Dahlberg. 2019. *The Science of Effective Mentorship in STEMM. Consensus Study Report.* https://doi.org/10.17226/25568

[40] Emily Callaci. 2020. On Acknowledgments. *The American Historical Review* 125, 1 (02 2020), 126–131. https://doi.org/10.1093/ahr/rhz938 arXiv:https://academic.oup.com/ahr/article-pdf/125/1/126/32323240/rhz938.pdf

[41] CAST. 2018. Universal Design for Learning Guidelines version 2.2. (2018). http://udlguidelines.cast.org

[42] James I. Charlton. 1998. *Nothing about us without us : disability oppression and empowerment.* University of California Press, Berkeley.

[43] T. L. Chen, M. Ciocarlie, S. Cousins, P. M. Grice, K. Hawkins, Kaijen Hsiao, C. C. Kemp, Chih-Hung King, D. A. Lazewatsky, A. E. Leeper, Hai Nguyen, A. Paepcke, C. Pantofaru, W. D. Smart, and L. Takayama. 2013. Robots for humanity: using assistive robotics to empower people with disabilities. *IEEE Robotics and Automation Magazine* 20, 1 (2013), 30–39.

[44] Aubrey Clayton. 2021. *Bernoulli's fallacy: Statistical illogic and the crisis of modern science.* Columbia University Press. http://cup.columbia.edu/book/bernoullis-fallacy/9780231199957

[45] Donald A. Clelland. 2013. *Chapter Four. Unpaid Labor as Dark Value in Global Commodity Chains.* Stanford University Press, 72–88. https://doi.org/doi:10.1515/9780804788960-010

[46] Mike Cohn. 2005. *An Introduction to Scrum.* Technical Report. https://www.mountaingoatsoftware.com/presentations/an-introduction-to-scrum

[47] S. Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* MIT Press. https://mitpress.mit.edu/books/design-justice open access: https://design-justice.pubpub.org/.

[48] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583. https://doi.org/10.1145/3531146.3533213

[49] Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press, New Haven.

[50] Christopher L. Dancy. 2022. Using a Cognitive Architecture to consider antiblackness in design and development of AI systems *(International Conference on Cognitive Modelling (ICCM))*. arXiv. https://doi.org/10.48550/ARXIV.2207.00644

[51] Maya Daneva, Egbert Van Der Veen, Chintan Amrit, Smita Ghaisas, Klaas Sikkel, Ramesh Kumar, Nirav Ajmeri, Uday Ramteerthkar, and Roel Wieringa. 2013. Agile requirements prioritization in large-scale outsourced system projects: An empirical study. *Journal of Systems and Software* 86, 5 (2013), 1333–1353.

[52] Partha Dasgupta. 2021. *The Economics of Biodiversity: the Dasgupta Review.* HM Treasury. https://www.gov.uk/government/publications/final-report-the-economics-of-biodiversity-the-dasgupta-review

[53] Norman Davies. 2001. *Heart of Europe : the past in Poland's present.* Oxford University Press, Oxford ;.

[54] Hannah E. Davis, Gina S. Assaf, Lisa McCorkell, Hannah Wei, Ryan J. Low, Yochai Re'em, and et al. 2021. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine* 38 (Aug. 2021). https://doi.org/10.1016/j.eclinm.2021.101019

[55] Hannah E. Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J. Topol. 2023. Long COVID: Major Findings, Mechanisms and Recommendations. *Nature Reviews Microbiology* 21, 3 (March 2023), 133–146. https://doi.org/10.1038/s41579-022-00846-2

[56] Philipp Diebold and Marc Dahlem. 2014. Agile Practices in Practice: A Mapping Study. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, England, United Kingdom) *(EASE '14)*. Association for Computing Machinery, New York, NY, USA, Article 30, 10 pages. https://doi.org/10.1145/2601248.2601254

[57] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 135–143.

[58] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism.* The MIT Press, Cambridge, Massachusetts. http://data-feminism.mitpress.mit.edu/

[59] Kim Dikert, Maria Paasivaara, and Casper Lassenius. 2016. Challenges and success factors for large-scale agile transformations. *Journal of Systems and Software* 119 (2016), 87–108.

[60] Torgeir DingsØyr, Nils Brede Moe, Tor Erlend Fægri, and Eva Amdahl Seim. 2018. Exploring software development at the very large-scale: a revelatory case study and research agenda for agile method adaptation. *Empirical Software Engineering* 23, 1 (2018), 490–520.

[61] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the Carbon Intensity of AI in Cloud Instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1877–1894. https://doi.org/10.1145/3531146.3533234

[62] Jay T Dolmage. 2017. *Academic Ableism : Disability and Higher Education.* University of Michigan Press, Ann Arbor. https://www.press.umich.edu/9708722/academic_ableism

[63] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination.* The MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

[64] Fuller, Joseph B., Raman, Manjari, Sage-Gavin, Eva, and Hines, Kristen. 2021. Hidden Workers: Untapped Talent. https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf

[65] Timnit Gebru. 2021. Hierarchy of Knowledge in Machine Learning and Related Fields and Its Consequences. In *The Future is Intersectional: Black Women Interrogating Technology.* Spelman College Center of Excellence for Minority Women in STEM. https://youtu.be/OL3DowBM9uc Announcement: https://www.spelman.edu/coe-mws/news-events/the-future-is-intersectional-series/2021/04/14/default-calendar/the-hierarchy-of-knowledge-in-machine-learning-and-related-fields-and-its-consequences.

[66] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (nov 2021), 86–92. https://doi.org/10.1145/3458723

[67] Jan Gogoll and Julian F. Müller. 2017. Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics* 23, 3 (2017), 681–700.

[68] Jan Gogoll, Niina Zuber, Severin Kacianka, Timo Greger, Alexander Pretschner, and Julian Nida-Rümelin. 2021. Ethics in the software development process: from codes of conduct to ethical deliberation. *Philosophy & Technology* 34, 4 (2021), 1085–1108.

[69] Jan Gogoll, Niina Zuber, Severin Kacianka, Timo Greger, Alexander Pretschner, and Julian Nida-Rümelin. 2020. Ethics in the Software Development Process: From Codes of Conduct to Ethical Deliberation. *arXiv preprint arXiv:2011.03016* (2020).

[70] Stephen Jay. Gould. 1996. *The mismeasure of man* (rev. and expanded. ed.). Norton, New York.

[71] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt Publishing Company, Boston.

[72] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 449.

[73] Jamie Harding. 2018. *Qualitative data analysis from start to finish* (second edition ed.). SAGE Publications Ltd.

[74] Christina N. Harrington. 2020. The Forgotten Margins: What is Community-Based Participatory Health Design Telling Us? *Interactions* 27, 3 (apr 2020), 24–29. https://doi.org/10.1145/3386381

[75] Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 604, 15 pages. https://doi.org/10.1145/3491102.3501995

[76] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *Journal of Machine Learning Research* 21, 248 (2020), 1–43. http://jmlr.org/papers/v21/20-312.html

[77] S. Hendren. 2020. *What Can a Body Do?: How We Meet the Built World.* Penguin Publishing Group. https://www.penguinrandomhouse.com/books/561049/what-can-a-body-do-by-sara-hendren/

[78] Kashmir Hill. [n.d.]. Microsoft Plans to Eliminate Face Analysis Tools in Push for 'Responsible A.I.'. https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html?smid=url-share

[79] Kashmir Hill. 2019. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf

[80] Kashmir Hill. 2020. Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

[81] Kashmir Hill. 2020. Wrongfully Accused by an Algorithm. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

[82] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics* 24, 5 (2018), 1521–1536.

[83] Han-Yin Huang and Cynthia C. S. Liem. 2022. Social Inclusion in Curated Contexts: Insights from Museum Practices. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 300–309. https://doi.org/10.1145/3531146.3533095

[84] Andrew Hundt. 2021. *Effective Visual Robot Learning: Reduce, Reuse, Recycle.* Dissertation. Johns Hopkins University. https://jscholarship.library.jhu.edu/handle/1774.2/66803 Talk: https://youtu.be/R3dv3ARXpco.

[85] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. Robots Enact Malignant Stereotypes, In 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul, Republic of Korea). *FAccT*, 743–756. https://doi.org/10.1145/3531146.3533138 website: https://sites.google.com/view/robots-enact-stereotypes/home PDF with appendix: https://arxiv.org/pdf/2207.11569.pdf.

[86] Andrew Hundt, Varun Jain, Chia-Hung Lin, Chris Paxton, and Gregory D. Hager. 2019. The CoSTAR Block Stacking Dataset: Learning with Workspace Constraints. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1797–1804.

[87] Andrew Hundt, Benjamin Killeen, Nicholas Greene, Hongtao Wu, Heeyeon Kwon, Chris Paxton, and Gregory D. Hager. 2020. "Good Robot!": Efficient Reinforcement Learning for Multi-Step Visual Tasks with Sim to Real Transfer. In *IEEE Robotics and Automation Letters*, Vol. 5. 6724–6731. https://doi.org/10.1109/LRA.2020.3015448

[88] CMMI Institute. 2019. *Capability Maturity Model Integration (CMMI) v2.0.* Technical Report. CMMI Institute. https://cmmiinstitute.com/

[89] Liz Jackson, Alex Haagaard, and Rua Williams. 2022. Disability Dongle. (2022). https://blog.castac.org/2022/04/disability-dongle/ The term Disability Dongle was coined by Liz Jackson in 2019.

[90] Brian Jordan Jefferson. 2020. *Digitize and punish : racial criminalization in the digital age.* University of Minnesota Press, Minneapolis.

[91] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2206–2222. https://doi.org/10.1145/3531146.3534637

[92] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[93] Matthew Johnson. 2020. *Undermining Racial Justice: How One University Embraced Inclusion and Inequality.* Cornell University Press.

[94] Severin Kacianka and Alexander Pretschner. 2021. Designing Accountable Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 424–437.

[95] Alison Kafer. 2013. *Feminist, queer, crip.* Indiana University Press, Bloomington.

[96] Andrea Paola Hernández Nadine Freischlad Karen Hao, Heidi Swart. 2022. AI Colonialism. https://www.technologyreview.com/supertopic/ai-colonialism-supertopic/

[97] Rashidah Kasauli, Eric Knauss, Benjamin Kanagwa, Agneta Nilsson, and Gul Calikli. 2018. Safety-Critical Systems and Agile Development: A Mapping Study. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).* 470–477. https://doi.org/10.1109/SEAA.2018.00082

[98] Inga Kroener, David Barnard-Wills, and Julia Muraszkiewicz. 2021. Agile Ethics: An Iterative and Flexible Approach to Assessing Ethical, Legal and Social Issues in the Agile Development of Crisis Management Information Systems. *Ethics and Information Technology* 23, 1 (Nov. 2021), 7–18. https://doi.org/10.1007/s10676-019-09501-6

[99] Daniel Reid Kuespert. 2016. *Research Laboratory Safety.* De Gruyter. https://doi.org/doi:10.1515/9783110444438

[100] Craig Larman and Bas Vodde. 2010. *Practices for Scaling Lean and Agile Development: Large, Multisite, and Offshore Product Development with Large-Scale Scrum.*

[101] Yanni A. (Yanni Alexander) Loukissas. 2019 - 2019. *All data are local : thinking critically in a data-driven society.* The MIT Press, Cambridge, Massachusetts.

[102] Alexandra Sasha Luccioni and Alex Hernandez-Garcia. 2023. Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning. arXiv:2302.08476 [cs.LG] https://doi.org/10.48550/arXiv.2302.08476

[103] Sasha Luccioni, Victor Schmidt, Alexandre Lacoste, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning.* https://arxiv.org/abs/1910.09700

[104] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445 checklist website: https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/.

[105] Emanuel Maiberg. 2023. Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale — 404media.co. https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/. [Accessed 07-09-2023].

[106] Sarah Maza. 2017. *Thinking about history.* University of Chicago Press.

[107] Sean McGregor. 2020. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *AAAI.* 15458–15463. https://incidentdatabase.ai/

[108] Charlton D. McIlwain. 2019. *Black Software : the Internet and Racial Justice, from the AfroNet to Black Lives Matter.* Oxford University Press USA - OSO, Oxford.

[109] Donella H. Meadows and Diana Wright. 2009. *Thinking in systems : a primer.* Earthscan, London.

[110] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. AI Now 2019 Report. https://ainowinstitute.org/AI_Now_2018_Report.pdf

[111] Sabelo Mhlambi. 2020. From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. (2020). https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial

[112] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and Inclusion Metrics in Subset Selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 117–123.

[113] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 220–229.

[114] Brent Mittelstadt. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* 1, 11 (Nov. 2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[115] Jessica Morley, Luciano Floridi, Lane Kinsey, et al. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci Eng Ethics* 26 (2020), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

[116] Arvind Narayanan. 2022. The Limits Of The Quantitative Approach To Discrimination. In *James Baldwin Lecture Series* (Princeton, New Jersey, USA). Princeton University. https://www.cs.princeton.edu/~arvindn/talks/baldwin-discrimination/baldwin-discrimination-transcript.pdf Announcement: https://aas.princeton.edu/events/2022/james-baldwin-lecture-series-limits-quantitative-approach-discrimination.

[117] National Research Council. 2014. *Safe Science: Promoting a Culture of Safety in Academic Chemical Research.* The National Academies Press, Washington, DC. https://doi.org/10.17226/18706

[118] Benjamin W Nelson and Nicholas B Allen. 2019. Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study. *Jmir mhealth and uhealth* 7, 3 (2019).

[119] NMA. 2018. CORESafety TV: August 2018. National Mining Association (NMA). https://youtu.be/w3UrhyZ_StI?t=45 Swiss Cheese Model of Accident Causation.

[120] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, New York.

[121] OECD. 2019. *OECD SME and Entrepreneurship Outlook 2019.* 320 pages. https://doi.org/https://doi.org/10.1787/34907e9c-en

[122] OECD. 2019. *Skills Matter.* 130 pages. https://doi.org/https://doi.org/10.1787/1f029d8f-en

[123] Council of Medical Specialty Societies (CMSS) and Patient-Led Research Collaborative (PLRC). 2023. Patient-Led Research Scorecards. (2023). https://cmss.org/wp-content/uploads/2023/01/11231_CMSS_Plybk_Scorecards_FINAL.pdf https://patientresearchcovid19.com/storage/2023/02/Patient-Led-Research-Scorecards.pdf.

[124] National Academies of Sciences Engineering and Medicine. 2018. *Sexual Harassment of Women: Climate Culture and Consequences in Academic Sciences Engineering and Medicine. Consensus Study Report.* National Academies Press. https://doi.org/10.17226/24994

[125] National Academies of Sciences Engineering and Medicine. 2020. *Promising Practices for Addressing the Underrepresentation of Women in Science Engineering and Medicine: Opening Doors. Consensus Study Report.* National Academies Press. https://doi.org/10.17226/24994

[126] National Academies of Sciences Engineering and Medicine. 2022. *Fostering Responsible Computing Research: Foundations and Practices. Consensus Study Report.* National Academies Press. https://doi.org/10.17226/26507

[127] World Health Organization. 2023. Disability and Health Fact Sheet. https://www.who.int/en/news-room/fact-sheets/detail/disability-and-health. Accessed on 2023-09-26.

[128] Mark L. Ornelas, Gary B. Smith, and Masoumeh Mansouri. 2022. Redefining Culture in Cultural Robotics. *AI & SOCIETY* (June 2022). https://doi.org/10.1007/s00146-022-01476-1

[129] Anistasia K. Ostrowski, Christinia N. Harrington, Cynthia Breazeal, and Hae Won Park. 2021. Personal Narratives in Technology Design: The Value of Sharing Older Adults' Stories in the Design of Social Robots. *Frontiers in Robotics and AI* 8. https://doi.org/10.3389/frobt.2021.716581

[130] Frank Pasquale. 2020. *New Laws of Robotics.* Harvard University Press. https://doi.org/doi:10.4159/9780674250062

[131] Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics* 6 (12 2018), 571–585. https://doi.org/10.1162/tacl_a_00040 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00040/1567662/tacl_a_00040.pdf

[132] Chris Paxton, Andrew Hundt, Felix Jonathan, Kelleher Guerin, and Gregory D. Hager. 2017. CoSTAR: Instructing collaborative robots with behavior trees and vision. In *2017 IEEE International Conference on Robotics and Automation (ICRA).* 564–571.

[133] Chris Paxton, Felix Jonathan, Andrew Hundt, Bilge Mutlu, and Gregory D. Hager. 2018. Evaluating Methods for End-User Creation of Robot Task Plans. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 6086–6092.

[134] Tenelle Porter, Chayce R. Baldwin, Michael T. Warren, Elise D. Murray, Kendall Cotton Bronk, Marie J.C Forgeard, Nancy E. Snow, and Eranda Jayawickreme. 2021. Clarifying the Content of Intellectual Humility: A Systematic Review and Integrative Framework. *Journal of Personality Assessment* 0, 0 (2021), 1–13. https://doi.org/10.1080/00223891.2021.1975725 arXiv:https://doi.org/10.1080/00223891.2021.1975725 PMID: 34569872.

[135] Tenelle Porter, Abdo Elnakouri, Ethan A. Meyers, Takuya Shibayama, Eranda Jayawickreme, and Igor Grossmann. 2022. Predictors and Consequences of Intellectual Humility. *Nature Reviews Psychology* (June 2022). https://doi.org/10.1038/s44159-022-00081-9

[136] Julie R Posselt. 2020. *Equity in Science: Representation, Culture, and the Dynamics of Change in Graduate Education.* Stanford University Press, Redwood City.

[137] Margaret Price. 2011. *Mad at school : rhetorics of mental disability and academic life.* University of Michigan Press, Ann Arbor.

[138] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23).* Association for Computing Machinery, New York, NY, USA, 1882–1895. https://doi.org/10.1145/3593013.3594134

[139] Organizers of QueerInAI, A Pranav, MaryLena Bleile, Arjun Subramonian, Luca Soldaini, Danica J. Sutherland, Sabine Weber, and Pan Xu. 2021. How to Make Virtual Conferences Queer-Friendly: A Guide. In *Proceedings of the 2021 Workshop on Widening NLP.* Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic. queerinai.org/diversity-guide

[140] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19).* Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[141] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22).* Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

[142] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAccT '20).* Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[143] Victor Ray. 2021. *On Critical Race Theory: Why It Matters & Why You Should Care.* Penguin Random House. https://www.penguinrandomhouse.com/books/708739/on-critical-race-theory-by-victor-ray/

[144] J Reason. 1990. The Contribution of Latent Human Failures to the Breakdown of Complex Systems. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 327, 1241 (1990), 475–484. https://doi.org/10.1098/rstb.1990.0090

[145] Toni Robertson and Jesper Simonsen. 2012. Participatory Design: an introduction. In *Routledge international handbook of participatory design.* Routledge, 21–38.

[146] Mark Ryan and Bernd Carsten Stahl. 2020. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society* 19, 1 (2020), 61–86.

[147] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 317 (oct 2021), 37 pages. https://doi.org/10.1145/3476058

[148] Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, and Stacy M. Branham Foad Hamidi. 2020. HCI Guidelines for Gender Equity and Inclusivity. https://www.morgan-klaus.com/gender-guidelines.html

[149] Mariah L. Schrum, Michael Johnson, Muyleng Ghuy, and Matthew C. Gombolay. 2020. Four Years in Review: Statistical Practices of Likert Scales in Human-Robot Interaction Studies. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20).* Association for Computing Machinery, New York, NY, USA, 43–52. https://doi.org/10.1145/3371382.3380739

[150] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices.* CRC Press.

[151] Andreas Schumacher, Selim Erol, and Wilfried Sihn. 2016. A Maturity Model for Assessing Industry 4.0 Readiness and Maturity of Manufacturing Enterprises. *Procedia CIRP* 52 (2016), 161–166.

[152] Ken Schwaber and Jeff Sutherland. 2020. The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game. https://www.scrum.org/scrum-guide-2020. https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf [Online; accessed 5-September-2023].

[153] James C. Scott. 1998. *Seeing like a state : how certain schemes to improve the human condition have failed.* Yale University Press, New Haven.

[154] International Telecommunication Union Development Sector. 2021. *Measuring digital development: Facts and figures 2021.* ITUPublications. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf

[155] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[156] Sofia Serholt, Sara Ljungblad, and Niamh Ní Bhroin. 2022. Introduction: Special Issue—Critical Robotics Research. *AI & SOCIETY* 37, 2 (June 2022), 417–423. https://doi.org/10.1007/s00146-021-01224-x

[157] Helen Sharp, Jennifer Preece, and Yvonne Rogers. 2019. Interaction design: Beyond human computer interaction by Preece, Sharp and Rogers (2019). (2019), 656.

[158] Luke Stark and Jevan Hutson. 2021. Physiognomic Artificial Intelligence. *Available at SSRN 3927300* (2021). https://doi.org/10.2139/ssrn.3927300

[159] Y. SUGIMORI, K. KUSUNOKI, F. CHO, and S. UCHIKAWA. 1977. Toyota production system and Kanban system Materialization of just-in-time and respect-for-human system. *International journal of production research* 15, 6 (1977), 553–564. https://doi.org/10.1080/00207547708943149

[160] Cella M Sum, Rahaf Alharbi, Franchesca Spektor, Cynthia L Bennett, Christina N Harrington, Katta Spiel, and Rua Mae Williams. 2022. Dreaming Disability Justice in HCI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 5 pages. https://doi.org/10.1145/3491101.3503731

[161] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002 [cs.LG] https://arxiv.org/abs/1901.10002

[162] Cmmi Product Team. 2010. CMMI for Development, Version 1.3. (2010).

[163] Patrick Thibodeau. 2012. John Deere plows into agile. https://www.computerworld.com/article/2500298/john-deere-plows-into-agile.html

[164] Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI Fairness for People with Disabilities. *AI Matters* 5, 3 (dec 2019), 40–63. https://doi.org/10.1145/3362077.3362086

[165] Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2023. Learning from Crowds with Crowd-Kit. arXiv:2109.08584 [cs.HC] https://arxiv.org/abs/2109.08584

[166] Diego Armando Diaz Vargas, Rui Xue, Claude Baron, Philippe Esteban, Rob Vingerhoeds, Y Citlalih, and Chao Liu. 2018. Implementing SCRUM to develop a connected robot. *arXiv preprint arXiv:1807.01662* (2018).

[167] Charisi Vasiliki, Chaudron Stephane, Di Gioia Rosanna, Vuorikari Riina, Escobar Planas Marina, Sanchez Martin Jose Ignacio, and Gomez Gutierrez Emilia. 2022. *Artificial Intelligence and the Rights of the Child: Towards an Integrated Agenda for Research and Policy.* Scientific analysis or review KJ-NA-31048-EN-N (online). Luxembourg (Luxembourg). https://doi.org/10.2760/012329

[168] Salome Viljoen. 2020. Democratic Data: A Relational Theory For Data Governance. *Social Science Research Network* (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727562

[169] Sara Wachter-Boettcher. 2017. *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech.* WW Norton & Company. https://wwnorton.com/books/Technically-Wrong/

[170] Wendell Wallach and Gary Marchant. 2019. Toward the Agile and Comprehensive International Governance of AI and Robotics [point of view]. *Proc. IEEE* 107, 3 (2019), 505–508.

[171] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620. https://doi.org/10.1177/20539517231177620 arXiv:https://doi.org/10.1177/20539517231177620

[172] Maranke Wieringa. 2020. What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833

[173] Damien Patrick Williams. 2022. *Belief, Values, Bias, and Agency: Development of and Entanglement with "Artificial Intelligence".* Ph.D. Dissertation. Virginia Polytechnic Institute and State University, Blacksburg, Virginia. http://hdl.handle.net/10919/111528

[174] Rua M. Williams. 2021. I, Misfit: Empty Fortresses, Social Robots, and Peculiar Relations in Autism Research. *Techné: Research in Philosophy and Technology* 25, 3 (2021), 451–478. https://doi.org/10.5840/techne20211019147

[175] Kumanan Wilson, Cameron Bell, Lindsay Wilson, and Holly Witteman. 2018. Agile research to complement agile development: a proposal for an mHealth research lifecycle. *npj Digital Medicine* 1, 1 (2018), 1–6. https://doi.org/10.1038/s41746-018-0053-1

[176] Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. 2023. Feminist Human-Robot Interaction: Disentangling Power, Principles and Practice for Better, More Ethical HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 72–82. https://doi.org/10.1145/3568162.3576973

[177] Claes Wohlin, Per Runeson, Martin Hst, Magnus C. Ohlsson, Bjrn Regnell, and Anders Wessln. 2012. *Experimentation in Software Engineering.* Springer.

[178] Alice Wong. 2020. *Disability visibility.* Vintage Digital. https://www.penguinrandomhouse.com/books/617802/disability-visibility-by-alice-wong/

[179] James Wright. 2023. *Robots Won't Save Japan.* Cornell University Press, Ithaca, NY. https://doi.org/doi:10.1515/9781501768064

[180] Di Wu. 2023. Good for tech: Disability expertise and labor in China's artificial intelligence sector, Vol. 28. https://doi.org/10.5210/fm.v28i1.12887

[181] Anon Ymous, Katta Spiel, Os Keyes, Rua M. Williams, Judith Good, Eva Hornecker, and Cynthia L. Bennett. 2020. "I Am Just Terrified of My Future" – Epistemic Violence in Disability Related Technology Research. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3334480.3381828

[182] Karolina Zawieska. 2020. Disengagement with Ethics in Robotics as a Tacit Form of Dehumanisation. *AI & SOCIETY* 35, 4 (Dec. 2020), 869–883. https://doi.org/10.1007/s00146-020-01000-3

[183] Shoshana Zuboff. 2019. *The age of surveillance capitalism : the fight for a human future at the new frontier of power* (first edition. ed.). PublicAffairs, New York.

A DEFINITIONS

For inclusive working definitions of identity, race, ethnicity, sex, gender, data setting, and dissolution model in an 'AI' and Robotics context see Hundt et al. [85]. Many definitions have multiple different perspectives and evolve over time, so our definitions are intended to serve as a useful starting point.

Data Settings "Rather than talking about datasets, [data studies scholar Yanni Loukissas [101]] advocates that we talk about data settings—his term to describe both the technical and the human processes that affect what information is captured in the data collection process and how the data are then structured." [58]

Crip Time "Recognizing how expectations of how long things should take account [of a range of] types of minds and bodies" so that we can 'bend the clock' rather than bending bodies [[95]]. Margaret Price suggests that crip time is the 'flexible approach to normative time frames' ([[137]]). Crip time has generally been interpreted as responsive: a way to impose critical delay through the refusal to follow strict schedules (schedules that might be normative, ableist, medically rehabilitative, and so on)." - Dolmage [62] (p. 179)

Universal Design for Learning (UDL) "An approach to curricula and teaching methods that strives to be more inclusive than American with Disabilities Act guidelines." [39] See also: CAST UDL Guidelines [41].

Intellectual Humility The "meta-cognitive ability to recognize the limitations of one's beliefs and knowledge [... which] might be particularly important for scientists for its role in enabling scientific progress" Porter et al. [135]. Porter et al. [134] surveys intellectual humility definitions.

Cultural Humility "An emerging idea that highlights the ability to maintain openness to others, engage in self-critique, redress power asymmetries, and work in partnerships to advocate for others." [136].

Jidoka "The term Jidoka as used at Toyota means 'to make the equipment or operation stop whenever an abnormal or defective condition arises'." [159]

In this work "identity" collectively refers to factors such as race, indigenous identity, physical ability, mental ability, age, national origin, experience, cultural conventions, gender identity, LGBTQIA+ identity, etc. Altogether, there is comparatively little investigation into ML and 'AI' with respect to identity factors when compared to algorithm performance on other metrics which can be problematic when misused [18, 24, 38, 58]. For example, "accuracy" can leave segments of a large population completely excluded or subject to harm as 'outliers'.

B ADDITIONAL DETAILS

B.1 Task Work Example of Equity Context and Human Override

Task work such as Amazon Mechanical Turk is one of the myriad powerful examples of the way human overrides and an ethical goal of understanding and prioritizing marginalized perspectives is simultaneously fundamental to technical goals. Gray and Suri [71] and others [57, 57, 72, 180] have conducted research into task work. They have found that these positions can be very precarious, to the point where it can be difficult for task workers to take care of their bills, their family, and their mental health; journalists have described cases with a striking resemblance to colonialism [96]. Task workers build varied expertise, and can even be engineers and PhDs [71].

Task instructions can be very unclear and designed without functional feedback mechanisms, such that they are nearly impossible to complete. To handle this and other aspects of their job, many have created their own meeting webpages, facebook groups, forums, etc, independently of the company operating the platform. They discuss the outcomes of various tasks, how to submit or organize or label data in a way that will get them paid. Side channel communication can variously improve the task results, undermine them, or even invalidate them without the knowledge of task hosts, depending on the case, purpose, requirements, and type of communication. For example, if a task worker makes a guide that improves the data collection process, this will be beneficial to results. However, this labor will go uncredited without two way communication, despite their work being a contribution that might have been worthy of co-authorship had they done the same exact work in person as part of a research role. Invalid results are possible in across-user studies where aspects of the task are differentially hidden as part of an experiment, communication of critical details might break the experiment if task hosts neglect the way the human processes actually behave, and design accordingly.

"Good for tech: Disability expertise and labor in China's artificial intelligence sector" provides a window into labeling as source of empowerment and freedom; it elaborates on the complexities of labeler expertise that task hosts and AI researchers often do not understand, and even occasionally impede. Gray and Suri [71] describes steps and approaches that mitigate these considerations for more effective data collection, including and supporting the task workers on teams, and creating interactive development processes.

B.2 Method Adoption and Team Size

Our method aims to scale from micro (1-9 people) up through large organizations (250+ people), as defined by the OECD [121]. The number and depth of processes can grow over time as the organization grows. Organizations with a small budget and team are expected to adopt our toolkit in a manner that is substantively different from those with a large budget allocation.

People working on micro projects (1-9) can learn essentials of sociotechnical impacts [58], tailor a project based on existing published resources, conduct quick assessments, adopt inclusive guidelines, and seek comments from different communities and affinitiy groups in person or online, provided legal and policy requirements are met such as Institutional Review Boards (IRBs). Micro organizations tend to have the tightest resource limitations, but even small steps can make a big difference, for example the basics of Model Cards [113] take a few days to learn and can help develop an appropriate scope and plan for a project, and by carefully detailing the purpose of any data collected [66], this can help prevent misuse. Artifacts (Sec. 4.6) can be managed as simply as a plain text file in a code base or a versioned text document. Detailed studies can be sourced from existing research, and new questions might be initially addressed by reaching out to experts or posting thoughtful questions to appropriate online affinity groups.

The resources in Sec. 3.5 on Participatory design also detail a range of methods that are effective from micro (1-10 people) up through medium scale projects. As projects increase in headcount through large sale projects (250+ people), additional systematization, adoption of more of the toolkit, more representative deliberative governance, and more comprehensive investigation of underlying issues becomes feasible. A detailed treatment is out of scope for this paper, so we recommend Jernite et al. [91] for the governance of large scale 'AI' projects.

One opportunity on new teams and projects is there is more flexibility to define an inclusive and participatory culture, and to create a diverse team. Existing Agile teams and projects have a body of knowledge about how to conduct adaptable project planning, so the new task and time elements can be easier to integrate on a scheduling basis. This Agile process is designed for integration with many other processes such as datasheets [66], model cards [113], Audits [142], Diversity metrics [112], EDAP ethical agile [94], and Participatory Design (Sec. 3.5).

B.3 Financial Considerations

Viljoen [168] argues that maximizing financial gain from or to data subjects misses the point, and that the purpose of data is to put people into person based relations with one another. Bessen et al. [21] conducts a survey that outlines the interplay of economic pressures and 'AI' ethics policies in startups that are under 10 years old. Caveated by the limitations of their methodology, they find:

> "that 58% of these startups have established a set of 'AI' principles. Startups with data-sharing relationships with high-technology firms or that have prior experience with privacy regulations are more likely to establish ethical 'AI' principles and are more likely to take costly steps, like dropping training data or turning down business, to adhere to their ethical 'AI' policies. Moreover, startups with ethical 'AI' policies are more likely to invest in unconscious bias training, hire ethnic minorities and female programmers, seek expert advice, and search for more diverse training data. Potential costs associated with data-sharing relationships and the adherence to ethical policies may create tradeoffs between increased 'AI' product competition and more ethical 'AI' production."

C PARTICIPANT-LED PROJECT ASSESSMENT SCORECARDS

The scorecards are discussed in Sec. 4.6 with the cards themselves below. Adapted from Patient-led Research Scorecards [123].

C.1 Participant Burden Scorecard

Participant Burden Scorecard	
Score	**Accessible Engagement**
-2	**Non-collaboration:** Research & Development organization dictates engagement avenues with no consideration of the participant population's access needs. Full participation may be impossible; carry a high time, effort, or monetary cost; or cause participants harm.
-1	**Minimal collaboration:** Research & Development organization considers the participant population when designing engagement avenues, but rarely provides additional accommodations when requested.
0	**Acceptable collaboration:** Research & Development organization designs engagement avenues to offer sufficient time and accessibility for the participant population's needs, and provides individuals with additional accommodations upon request.
1	**Great collaboration:** Research & Development organization designs engagement avenues to offer sufficient time and accessibility for the participant population's needs, ensures participants can easily request additional accommodations, and works with participants to co-design systemic updates in response to requests.
2	**Ideal collaboration:** Participants co-create engagement avenues from the outset to ensure that full participation is accessible and minimally harmful across impacted sub-populations.
Score	**Trauma Informed Practices**
-2	**Non-collaboration:** Research & Development organization does not consider possible trauma burdens for participants. Participants may experience discrimination, hostility, new or recalled trauma, or other harms as a result of participation. No trauma-informed practices are in place, and participants receive no resources or support for the trauma caused by participation.
-1	**Minimal collaboration:** Research & Development organization is aware of a possible trauma burden, but no systemic trauma-informed practices are in place, and participants receive no resources or support for their trauma.
0	**Acceptable collaboration:** Research & Development organization recognizes potential trauma burdens, and some trauma-informed practices are in place. Resources and support are provided to participants upon request.
1	**Great collaboration:** Research & Development organization implements trauma-informed practices throughout the study, and collaborates with participants to co-design adjustments to those practices during the project. Requests for resources and support are honored at a systemic level for all participants.
2	**Ideal collaboration:** A diverse array of participants, representative of the project's sub-populations, collaborates from the outset to co-create a safe, inclusive, mutually respectful environment; implement and adjust trauma-informed practices throughout the research process; and ensure all participants proactively receive sufficient, comprehensive resources and support.
Score	**Responsiveness to Participants**
-2	**Non-collaboration:** No formal channels for participant input are established. Research & Development organization does not address participant feedback, and may exclude or retaliate against participants who voice concerns.
-1	**Minimal collaboration:** Participants find channels for input to be unclear, difficult to access, or unsafe from retaliation. Participant feedback may be acknowledged, but rarely results in changes to the current project.
0	**Acceptable collaboration:** Research & Development organization creates clear, accessible, safe channels for participant input only after the research process has begun. Participant feedback is acknowledged, resulting in changes to analysis, presentation, or communication; and ad-hoc changes to the current project.
1	**Great collaboration:** Research & Development organization creates clear, accessible, safe channels for participant input throughout the research process; acknowledges participant feedback; and establishes mechanisms for participants to co-design systemic changes to the current project.
2	**Ideal collaboration:** Participants co-lead the project from end to end, including creating clear, accessible, safe channels for input, using that input to inform the research process, and acknowledging its impact. Members of the Research & Development organization are excited about and fully engaged in participant collaboration.
Score	**Compensation**
-2	**Non-collaboration:** Participants are compensated below market rate for their domain expertise and experience level, with no or limited options for when and how they are paid. Expenses, harm, and risk assumed from participation are not compensated.
-1	**Minimal collaboration:** Participants are compensated at market rate for their expertise and experience, with no or limited payment options. Expenses, harm, and risk are not compensated.
0	**Acceptable collaboration:** Research & Development organization sets participant compensation at market rate for their expertise and experience; and for anticipated expenses, harm, and risk. Multiple payment options are offered upfront. Requests for additional compensation and/or payment options are honored ad-hoc.

Continued on next page

Table C.1 – continued from previous page

Participant Burden Scorecard	
1	**Great collaboration:** Research & Development organization sets participant compensation at or above market rate for their expertise and experience; and for anticipated expenses, harm, and risk. Multiple payment options are offered upfront. Requests for additional compensation and/or payment options result in systemic changes that benefit all participants.
2	**Ideal collaboration:** Participants have decision-making roles in setting and adjusting compensation. Participants are compensated at or above market rate for their expertise and experience; and for anticipated expenses, harm, and risk; in the method and timing of their choice. Requests benefit all participants. Participants receive non-monetary compensation in the form of visibility, professional development, authorship, and awareness of their impact.

C.2 Participant/Partner Governance Scorecard

Participant/Partner Governance Scorecard	
Score	**Meaningful Decision making between groups**
-2	**Non-collaboration:** Decision-making for significant decisions (funding, project design, publication, etc.) is not communicated transparently and/or the Research & Development organization decides the decision-making process without participant input.
-1	**Minimal collaboration:** Decision-making process for significant decisions (funding, project design, publication, etc.) is not communicated and/or agreed upon. Participants have limited or not meaningful decision-making power.
0	**Acceptable collaboration:** Decision-making process for significant decisions (funding, project design, publication, etc.) is well communicated and agreed upon between participants and the Research & Development organization.
1	**Great collaboration:** Decision-making for significant decisions (funding, project design, publication, etc.) is well communicated and agreed upon between the participant and partner group, with deference given to the participant group.
2	**Ideal collaboration:** Decision-making for significant decisions (funding, project design, publication, etc.) is well communicated and agreed upon between the participant and partner group, with deference given to the participant group with sufficient support to make the decisions.
Score	**Accountability between groups**
-2	**Non-collaboration:** There is a lack of understanding of the rules of engagement/culture between groups with no written agreement and no defined consequences for not following through.
-1	**Minimal collaboration:** There is an understanding of the rules of engagement/culture but no written agreement and/or defined consequences for not following through between groups.
0	**Acceptable collaboration:** There is a shared understanding and written agreement of the rules of engagement/culture with defined consequences for not following through between groups.
1	**Great collaboration:** Shared understanding and written agreement of the rules of engagement/culture with defined consequences for not following through between groups. Deference is given to participant groups to define the engagement.
2	**Ideal collaboration:** Shared understanding and written agreement of the rules of engagement/culture with defined consequences for not following through between groups. Deference is given to participant groups to define the engagement with sufficient support.

C.3 Research & Development Organization Readiness Scorecard

Research & Development Organization Readiness Scorecard	
Score	**Recognition of Biases**
-2	**Non-collaboration:** Research & Development organization does not recognize bias and ignores feedback from participants.
-1	**Minimal collaboration:** Research & Development organization has limited awareness of its own biases and listens to some feedback from participants.
0	**Acceptable collaboration:** Research & Development organization is aware of its own biases, is open to feedback from participants, and implements some of the feedback.
1	**Great collaboration:** Research & Development organization is aware of its own biases, is open to feedback from participant groups, and actively iterates on feedback given.
2	**Ideal collaboration:** Research & Development organization is aware of its own biases and is open to listening to feedback from participant groups. It actively iterates on feedback given. Other participant groups can attest to a positive working relationship. The Research & Development organization has a systemic process for accepting input from participants and participant groups.
Score	**Collaboration Process**
-2	**Non-collaboration:** Research & Development organization has no dedicated infrastructure for collaborating with participants.
-1	**Minimal collaboration:** Research & Development organization has minimal resources/infrastructure for collaborating with participants.
0	**Acceptable collaboration:** Research & Development organization has dedicated some resources and infrastructure for collaborating with participants (e.g., participant panels); has at least one coordinating personnel focused on meeting the participant group's needs; conducts limited training to build skills to engage with participants.
1	**Great collaboration:** Research & Development organization has an established infrastructure and process for collaborating and co-designing with participants, including at least one dedicated person focused on meeting the participant group's needs and advocating to the rest of the Research & Development organization. It conducts routine training to build skills to engage with participants.
2	**Ideal collaboration:** Research & Development organization has an established infrastructure and process for collaborating with participants that has been vetted by other participants/participant groups. It has at least one dedicated person who is focused on meeting the participant group's needs. The partner is recognized as a participant ally vetted by other participants and participant groups with a background in justice as it applies to impacted identities (e.g. disability justice, racial justice, etc.). The organization conducts extensive training on meaningful engagement with participants.
Score	**Knowledge of Impacted Identities**
–	*Please provide a score for each impacted identity about which you wish to provide feedback: Race, physical disability, cognitive disability, age, national origin, cultural, gender, LGBTQIA+ status, wealth status, income status, an intersection of multiple identities, or specify one or more of your own identities.*
-2	**Non-collaboration:** Research & Development organization has no knowledge/experience with the impacted identities.
-1	**Minimal collaboration:** Research & Development organization has minimal knowledge/experience (less than one year) with the impacted identities.
0	**Acceptable collaboration:** Research & Development organization has at least one year worth of knowledge/experience with the impacted identities.
1	**Great collaboration:** Research & Development organization has more than one year worth of knowledge/experience of the impacted identities.
2	**Ideal collaboration:** Research & Development organization has extensive knowledge and direct experience with the impacted identities, and those with knowledge are in decision-making roles. The Research & Development organization has a systemic way to keep on top of information from the participant community as well as the latest research findings.

C.4 Integration Into Research Process Scorecard

Integration Into Research Process Scorecard	
Score	**Hypothesis Generation**
-2	**Non-collaboration:** Research goals are siloed from participants' priorities. Participants' questions and experiences are not included and/or are dismissed when generating research hypotheses.
-1	**Minimal collaboration:** Research goals attempt to involve participants' priorities but are limited by communication or collaboration. Participants' inquiries and lived experiences are rarely included when generating research hypotheses. Participants may have suggested the research question with no further involvement.
0	**Acceptable collaboration:** Research goals take into account participants' priorities. Participants' inquiries and lived experiences are included when generating research hypotheses.
1	**Great collaboration:** Research goals proactively address participants' priorities with sufficient ongoing collaboration. Participant organization's inquiries and lived experiences are included when generating research hypotheses. Participant organizations work with participants to co-design research hypotheses.
2	**Ideal collaboration:** Research goals are based on participants' priorities and co-written by participant organization or participant-researchers. Participants' inquiries and lived experiences share equal weight with Research & Development organization's interests when generating research hypotheses.
Score	**Project Design**
-2	**Non-collaboration:** Research & Development organization does not include participants in the project design process. Participants do not have the opportunity to provide input on the project design. Participant groups are utilized for recruitment purposes only, if at all.
-1	**Minimal collaboration:** Research & Development organization does not include participants in the project design process. Participants may be invited to review the project design, but feedback is rarely incorporated, and no functioning accountability system is in place.
0	**Acceptable collaboration:** Select participant voices are approached to inform the project design. Participants are invited to review the project design and have an impact on the project design.
1	**Great collaboration:** Participant organizations and their community's input are proactively invited to help inform the project design. Participant organizations are invited to co-design and review the project design, and participant feedback changes the project design.
2	**Ideal collaboration:** Project design is co-written and reviewed by a diverse array of participant-researchers representative of the project's sub-populations. If applicable, protocol testing is done by the participant community.
Score	**Analysis**
-2	**Non-collaboration:** Participants do not have input on what data to prioritize for analysis and methods of analysis.
-1	**Minimal collaboration:** Participants are asked to review manuscript drafts and prototypes, as applicable, but have little say in what data to prioritize for analysis and methods of analysis.
0	**Acceptable collaboration:** Participants are involved in interpreting data and carrying out analysis in some capacity.
1	**Great collaboration:** Participants or participant organizations are invited and involved in interpreting data and carrying out analysis anywhere in the project.
2	**Ideal collaboration:** Participant-researchers co-lead on the interpretation and analysis and/or work concurrently with the partner organization's research team to carry out the analysis.
Score	**Products, Artifacts, and/or Publications**
-2	**Non-collaboration:** Project results are inaccessible to participants and/or behind an academic paywall. Findings are not communicated in lay terms and/or products are not available. Data with a probable risk of an overall negative impact is released anyway.
-1	**Minimal collaboration:** Research & Development organization summarizes findings in lay terms, but project results are inaccessible to participants, are behind avoidable cost barriers such as an academic paywall. Data with a foreseeable risk of an overall negative impact is released anyway, with consent but no discussion of the risks.
0	**Acceptable collaboration:** Project results are freely accessible to participants and the public. Findings are summarized in lay terms in ways that are informative to the participant and impacted populations. Physical and Digital artifacts as well as licenses are available on fair, reasonable, and non-discriminatory (FRAND) terms. Data with a potential net negative impact is released with minimally informed consent.
1	**Great collaboration:** Project results are freely accessible to participants and the public. Findings are summarized in lay terms and are actively disseminated to the participant and impacted populations. Participant-researchers co-write the interpretation and analysis. Good faith dissenting assessments are welcome. Physical and Digital artifacts as well as licenses are available on fair, reasonable, and non-discriminatory (FRAND) terms. Potential net negative impacts from data release considered before data release with informed consent.

Continued on next page

Table 1 – continued from previous page

	Integration Into Research Process Scorecard
2	**Ideal collaboration:** Project results are freely accessible to participants and the public. Findings are summarized in lay terms and are actively disseminated to the participant and impacted populations. Participant organizations invite participants to co-write findings and reports. A channel of communication is available for participants to ask questions of the Research & Development organization. Good faith dissenting assessments are welcome. Physical and Digital artifacts are available for free or at cost with open licenses (where applicable) in a manner better than fair, reasonable, and non-discriminatory (FRAND) terms. Potential net negative impacts from data release are carefully mitigated, and data is not released if the risk is too high, or substantive steps are taken to inform participants and obtain full informed consent.
Score	**Attribution**
-2	**Non-collaboration:** Participants' work is attributed to others and/or participants are not attributed at all.
-1	**Minimal collaboration:** Participants are listed as being involved without a description of how they were involved. Participants were not consulted on how they prefer to be attributed.
0	**Acceptable collaboration:** Participants are acknowledged/credited in major public-facing communication (press, announcements, papers), to the extent that participants wish to be named. Participants were consulted on how they prefer to be attributed.
1	**Great collaboration:** Participant group is credited in all public-facing communication and included as authors on papers or products, to the extent that the participant group wishes to be named. Participant group was consulted on how they prefer to be attributed.
2	**Ideal collaboration:** Participants are acknowledged specifically for what they did throughout the engagement process, are credited in all public-facing communication, and included as authors on papers or products, to the extent that the participant group wishes to be named. Participant group was consulted on how they prefer to be attributed.

C.5 Researcher and Staff Support Scorecard

Researcher and Staff Support Scorecard	
Score	**Accessibility**
-2	**Non-collaboration:** Support services and resources are difficult to find, access, or use. Researchers and staff face barriers such as lack of information, complex procedures, technical issues, or limited availability.
-1	**Minimal collaboration:** Support services and resources are somewhat accessible but require significant effort or time from researchers and staff. Researchers and staff encounter some challenges such as insufficient information, unclear procedures, occasional glitches, or limited options.
0	**Acceptable collaboration:** Support services and resources are adequately accessible but could be improved. Researchers and staff can access them with reasonable effort and time. Researchers and staff experience few difficulties such as outdated information, inconsistent procedures, minor errors, or limited flexibility.
1	**Great collaboration:** Support services and resources are easily accessible and user-friendly. Researchers and staff can access them with minimal effort and time. Researchers and staff have no major problems such as inaccurate information, confusing procedures, frequent failures, or rigid policies.
2	**Ideal collaboration:** Support services and resources are highly accessible and customized. Researchers and staff can access them with ease and convenience. Researchers and staff have a positive experience such as comprehensive information, streamlined procedures, reliable performance, or adaptive solutions.
Score	**Work Life Balance**
-2	**Non-collaboration:** Researchers and staff have poor work-life balance. They face excessive workload, stress, or pressure that negatively affect their health, well-being, or personal life. They have no flexibility or autonomy in their work schedule or location, and cannot work remotely if needed. They struggle to complete their work efficiently and effectively, are dissatisfied with their work quality and outcomes, and receive no or negative feedback or recognition for their work.
-1	**Minimal collaboration:** Researchers and staff have low work-life balance. They face high workload, stress, or pressure that moderately affect their health, well-being, or personal life. They have limited flexibility or autonomy in their work schedule or location, and face barriers to work remotely if needed. They find it hard to complete their work efficiently and effectively, are unhappy with their work quality and outcomes, and receive little or mixed feedback or recognition for their work.
0	**Acceptable collaboration:** Researchers and staff have adequate work-life balance. They face manageable workload, stress, or pressure that slightly affect their health, well-being, or personal life. They have some flexibility or autonomy in their work schedule or location, and can work remotely if needed with some support. They are able to complete their work efficiently and effectively, are satisfied with their work quality and outcomes, and receive adequate or positive feedback or recognition for their work.
1	**Great collaboration:** Researchers and staff have a strong work-life balance. They face balanced workload, stress, or pressure that enhance their health, well-being, or personal life. They have a lot of flexibility or autonomy in their work schedule or location, and can work remotely if needed with ease. They can complete their work efficiently and effectively with ease, are proud of their work quality and outcomes, and receive frequent or constructive feedback or recognition for their work.
2	**Ideal collaboration:** Researchers and staff have optimal work-life balance. They face optimal workload, stress, or pressure that boost their health, well-being, or personal life. They have full flexibility or autonomy in their work schedule or location, and can work remotely if needed with creativity. They excel at completing their work efficiently and effectively with creativity, are delighted with their work quality and outcomes, and receive consistent or exceptional feedback or recognition for their work.
Score	**Attribution**
-2	**Non-collaboration:** Researchers' and staff's work is attributed to others and/or researchers and staff are not attributed at all. Researchers and staff face discrimination, exclusion, or exploitation based on their identity, role, or contribution. Researchers, staff scientists, staff engineers, service staff, contractors, family, students, assistants, and colleagues who contributed with material, editing, logistical, research, or in other capacities are not consulted on how they prefer to be attributed, recognized, or credited. There may be uncredited ghostwriters or other uncredited people who made research contributions.
-1	**Minimal collaboration:** Researchers' and staff's work is acknowledged in a generic or vague way, without specifying their individual or collective contributions. Researchers and staff are not given equal opportunities or recognition based on their identity, role, or contribution. Researchers, staff scientists, staff engineers, service staff, contractors, family, students, assistants, and colleagues who contributed with material, editing, logistical, research, or in other capacities are rarely consulted on how they prefer to be attributed, recognized, or credited.

Continued on next page

Table continued from previous page

Researcher and Staff Support Scorecard	
0	**Acceptable collaboration:** Researchers' and staff's work is acknowledged in a fair and respectful way, with some indication of their individual or collective contributions. Researchers and staff are given equal opportunities and recognition based on their identity, role, or contribution. Researchers, staff scientists, staff engineers, service staff, contractors, family, students, assistants, and colleagues who contributed with material, editing, logistical, research, or in other capacities are consulted on how they prefer to be attributed, recognized, or credited.
1	**Great collaboration:** Researchers' and staff's work is acknowledged in a detailed and specific way, with clear indication of their individual or collective contributions. Researchers and staff are given equal opportunities and recognition based on their identity, role, or contribution. Researchers and staff are involved in the decision-making process regarding the attribution of credit. Researchers, staff scientists, staff engineers, service staff, contractors, family, students, assistants, and colleagues who contributed with material, editing, logistical, research, or in other capacities are consulted on how they prefer to be attributed, recognized, or credited.
2	**Ideal collaboration:** Researchers' and staff's work is acknowledged in a comprehensive and inclusive way, with explicit indication of their individual or collective contributions. Researchers and staff are given equal opportunities and recognition based on their identity, role, or contribution. Researchers and staff share the decision-making power regarding the attribution of credit. Researchers, staff scientists, staff engineers, service staff, contractors, family, students, assistants, bystanders, and practitioners who provide their expertise in the field (e.g., nurses, medical care staff, practitioners in any application industry), task workers (e.g., Amazon Mechanical Turk), and colleagues who contributed with ideas, material, editing, logistical, research, or in other capacities are consulted on how they prefer to be attributed, recognized, or credited. The attribution of credit is transparent, accountable, and respectful of the diversity and dignity of all contributors. Everyone involved is credited specifically for what they did throughout the engagement process (e.g., providing encouragement, funding, transportation, data analysis, etc.), credited in all public-facing communication (paper, speech, website, social media), and included as authors or acknowledgments on papers or products when appropriate, with a statement of their significance.