EASY WINS AND LOW HANGING FRUIT. BLUEPRINTS, TOOLKITS, AND PLAYBOOKS TO ADVANCE DIVERSITY AND INCLUSION IN AI

TOMASZ HOLLANEK AND MAYA INDIRA GANESH

The emergence of AI has been accompanied by myriad moral and philosophical questions. Some of these are speculative intuition pumps to activate a philosopher's muscles of reasoning: 'which way should the train trolley with the failed brakes be directed—jeopardizing the life of one person working on the track, or five?' There are more complex and pressing real-world questions as well, like: should commercial art and design companies be pointing their employees to generative AI image-making tools to speed up the ideation process? And, similarly: should large language models be trained by psychologists and psychiatrists to deliver therapeutic services to people with mental health conditions, thus making mental health services cheaper and more accessible? In tandem with the emergence of AI technologies and the ethical and moral challenges they animate, come lists of high-level ethics principles to be prioritized by developers to ensure their products' desirable social impact: human rights, fairness, non-discrimination, privacy, transparency, accountability. New legal requirements such as the European Union's AI Act set out bright lines around applications at high levels of risk, including credit scoring, targeted profiling, facial recognition, and automated hiring. These bright lines have emerged thanks to documentation and analysis of risks and harms in various contexts. However, the speed and scale at which AI technologies function, the social and cultural complexities emerging at the sites of their application, and potential harms emergent therein present challenges for the software development community. How should the ethical concerns associated with AI be managed and mitigated? How will they be framed, broken down into manageable parts, and addressed through computational, social, and policy actions? How can development teams be eased into openly and consciously engaging in ethical deliberation as part of the design process? The question here is not about *whether* ethical reasoning *could*, in fact, be turned into something modular, formulaic; into a pattern to be adopted by designers. Nor is it about the consequences of translating 'ethics' into user-friendly forms and formats that developers can immediately recognize and, therefore, operationalize. If *convenience* is the 'condition we inhabit within contemporary capitalism,'[1] the key question is *how*? How can inconvenient questions about the trade-offs and conflicts of interest be posed in ways that are both legible and bearable to those in the position to transform the development pipeline?

Enter the toolkit. A toolkit is a design staple, a set of ready-to-use practices to solve a problem or achieve a specific goal. A toolkit's promise is scaling and continuity; that a set of instructions, practices, or workflows will deliver a consistent, desired result. Sometimes that result is just

1 Joshua Neves and Marc Steinberg, 'The Cultural Politics of In/Convenience', *Global Emergent Media: In Progress*, January 2023, https://www.globalemergentmedia.com/in-progress/the-cultural-politics-of-in%2Fconvenience.

a *process,* rolling out across diverse spatial and temporal locations. As such, toolkits are a favored methodology when things are to be made collaboratively and collectively, and when the vagaries of time and place introduce discrepancies or inconsistencies. It is precisely the assumption that AI can be *designed* to adhere to sets of values to avoid perpetuating harm that brings AI ethics into the ambit of design. And it is the prevalent belief that following a predetermined process can ensure that the end-product is ethical and responsible that establishes the toolkit as the primary instrument of AI ethics. The ongoing *toolkitification* of AI ethics, the subject of this essay, reformulates ethical practice as frictionless, modular, as something that *can* and *should* be made convenient and scalable, and, as such, responds to the demands of *convenience* as a means of grappling with the complexity of our contemporary condition.

This complexity is software itself; its chains of supply and demand, infrastructural politics and extractivism; its code—a palimpsest of social, economic, and cultural norms and values; its power that, to its developers, is not a mysterious unaccountable force but something to be tamed, trained, or taught. Software is supposed to make life and work easier. Software is supposed to make the building of more software easier; modularity makes this possible. Modularity is a central organizational logic of software and is about the division of labor;[2] it enables reach and, in this way, makes building elaborate projects more convenient—both to imagine and to execute. In their study of how software developers assess their own professional accountability for the ethical harms and lapses of AI and algorithmic technologies, David Gray Widder and Dawn Nafus argue that, just as modularity helps to 'minimize friction as the code passes through many hands',[3] ethical challenges are similarly 'encapsulated into a module of work' so as not to 'introduce friction into the development process'.[4] It is the kind of division of labor that underlies 'convenient media'.[5]

Yet, managing, overseeing, tending to, and patching software requires its own elaborate system of systems. The 'countervailing tendencies' of elaborate software as simultaneously modular and Byzantine, unknowable and yet accessible, serve a specific challenge for an ideological or values-driven social or cultural project;[6] in this case, the work of correcting and reshaping institutional and technological systems, like AI, to be inclusive and equitable. For values and ideologies are also simultaneously highly contextual, shifting, broad, diverse, and yet also translatable into specific actions, positions, and normative rules. Translation requires engagement beyond individuals, with institutional and structural actors and norms; this draws us yet again to the vastness of systems. There's a mirroring here between software, and ideologies and institutions that design toolkits propose to intervene in but remain trapped by.

2 Lev Manovich, *The Language of New Media*, Cambridge, MA: The MIT Press, 2001, pp. 30-31.
3 David Gray Widder and Dawn Nafus, 'Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility,' *Big Data & Society* 10:1 (2023): 2.
4 Widder and Nafus, 'Dislocated accountabilities in the "AI supply chain"', 7-8.
5 Neves and Steinberg, 'The Cultural Politics of In/Convenience'.
6 Miriam Posner, 'Breakpoints and Black Boxes: Information in Global Supply Chains,' *Postmodern Culture* 31:3 (2021), https://www.pomoculture.org/2021/12/01/breakpoints-and-black-boxes-information-in-global-supply-chains/.

Toolkitification: the Making-convenient of 'Ethics' in and for AI Design

There is no agreed-upon definition of a design toolkit as the term might refer to both technical resources and educational brochures, to static text documents and interactive, web-based applications.[7] It is often used interchangeably with a singular tool, guideline, method, or blueprint. And yet, we can identify toolkit-ification as an industry-wide phenomenon occurring in response to the growing awareness of the risks and liabilities related to AI development. The OECD's Catalogue of Tools & Metrics for Trustworthy AI,[8] the biggest collection of its kind featuring over seven hundred toolkits (at the time of writing), speaks to the scale of this trend.[9] The landscape of AI ethics toolkits is even wider, encompassing radical design ideation tools and wiki-style web pages, such as the Intersectional AI Toolkit by Sarah Ciston.[10] So wide, in fact, that 'toolkit-scoping', the act of comparing available toolkits and testing their usefulness for AI professionals, has become a sub-genre of AI ethics research.[11] What this toolkit-scoping work makes clear is that the toolkit paradigm privileges certain kinds of information, worldviews, and practices in its organization and presentation, and in doing so discursively re-shape what (AI) ethics *is*.[12] Specifically, the toolkit implies that ethical conflicts and challenges associated with AI can be managed and that ethical practice is feasible and approachable. Toolkits for trustworthy, safe, responsible, and ethical AI make a promise: that the tools they contain are easily *adoptable* within existing workflows and *adaptable* to a particular team's needs, and that ensuring AI is made responsibly doesn't imply a procedural revolution—only selecting and applying an appropriate tool at the right stage of the design process.

The modularity of software development meets attempts at 'translating' the 'theory' of AI ethics into development 'practice'. Just like software is composed of smaller, constituent parts that can be swapped out, reconstituted, and re-assembled, 'ethics work' is embraced in terms of a similar organizing principle, as sets of actionable practices that can be stacked on and slotted in. The most evocative example of an ethical issue getting turned into a development module is the matter of AI bias; 'technical tools to remove bias' are, for instance, among the most sought after by the users of the OECD's Catalogue of Tools for Trustworthy AI.[13] This particular approach to toolkitification of AI ethics that frames ethics work as technical work has already been criticized for de-emphasizing the social, collective, and cultural value of diverse stakeholders' knowledge and engagement with AI—flattening and decontextualizing ethics by

7 Dorian Peters, Lian Loke, and Naseem Ahmadpour, 'Toolkits, cards and games – a review of analogue tools for collaborative ideation', *CoDesign* 17:4 (2020): 410-434.

8 OECD, 'Catalogue of Tools & Metrics for Trustworthy AI', *OECD.AI Policy Observatory*, https://oecd.ai/en/catalogue/tools.

9 Tomasz Hollanek, 'The Ethico-politics of Design Toolkits: Responsible AI Tools, From Big Tech Guidelines to Feminist Ideation Cards,' forthcoming.

10 Intersectional AI Toolkit, https://intersectionalai.miraheze.org/wiki/Intersectional_AI_Toolkit.

11 Hollanek, 'The Ethico-politics of Design Toolkits'.

12 Richmond Y. Wong, Michael A. Madaio, and Nick Merrill, 'Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics,' *Proceedings of the ACM on Human-Computer Interaction* 7, Issue CSCW1 (2023).

13 OECD, 'Catalogue of Tools & Metrics for Trustworthy AI'.

proposing generalizable and scalable practices.[14] But Widder and Nafus also highlight in their study the kind of ethics work that no one even attempts to turn into a development module; they show that this sort of work is 'frequently left undone or cast as low status work, offloaded to contractors' or turned into 'administrative labor no one else want[s] to do'.[15] Our interest lies precisely in the challenges that are seemingly 'untranslatable' into neat and prepackaged work modules, into additional steps conveniently fitting the already established development pipeline, and how design toolkits nonetheless promise to facilitate this difficult, inconvenient work—to help their users address complex issues, such as discrimination of marginalized groups perpetuated by AI systems, comprehensively and systematically. To think through the effects of toolkitification as a process of making 'ethics' convenient in and for AI design, we will look at a set of toolkits that aim to ensure AI is equitable and inclusive. These toolkits move beyond the 'remove-bias' type of work and indeed acknowledge the complexity of the ethical issues at hand. Yet they also promise to make these issues more approachable and bearable, and the processes of addressing them not only manageable but also, at times, *fun*.

Diversity, Equity, and Inclusion in AI

DEI or Diversity, Equity, and Inclusion (also referred to as DEI) is positioned as both a challenge and solution to AI's problems. While unfair, biased, and discriminatory outcomes along the lines of gender, race, class and intersections of these have emerged from the large-scale applications of algorithmic and automated technologies, these problematic consequences of AI deployment have also led to the establishment of the field of public, industrial, and academic inquiry into the ethics of algorithms and ethics of AI. Incident databases and registers aggregate the various biased outcomes of algorithmic decision-making.[16] Well-known cases include: Amazon's infamous Rekognition tool that negatively discriminated against women's CVs; the ProPublica investigation that revealed racially biased outcomes in a recidivism prediction algorithm; the UK government's disastrous A-level algorithm to predict school-leaving grades that delivered results along the lines of class and postcode. Yet, the business magazine Forbes reports that, according to startups in HR and recruitment, AI applications can enable DEI by identifying biased or stereotypical use of language in job advertisements and by identifying patterns of marginalization or disconnectedness among existing workers.[17] DEI is positioned as a solution to the 'white guy problem' in AI: the lack of gender, racial, and social diversity among AI's most powerful designers and developers is seen as a major influence on AI being biased and discriminatory in the first place.[18] DEI, understood as an equity measure, corrects the profit and power imbalance associated with AI to a wider community.

14 Wong, Madaio, and Merrill, 'Seeing Like a Toolkit'. See also, Thilo Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines,' *Minds & Machines* 30 (2020): 99-120.

15 Widder and Nafus, 'Dislocated accountabilities in the "AI supply chain"', 8.

16 See AI Incident Database, https://incidentdatabase.ai; AIAAIC Repository, https://www.aiaaic.org/aiaaic-repository.

17 Rebekah Bastian, 'AI Brings Opportunities And Risks To Workplace DEI Efforts,' *Forbes*, 8 May 2023, https://www.forbes.com/sites/rebekahbastian/2023/05/08/ai-brings-opportunities-and-risks-to-workplace-dei-efforts/?sh=4614ed8b4b2a; Jia Rizvi, 'How AI Can Be Leveraged For Diversity And Inclusion,' *Forbes*, 19 November 2023, https://www.forbes.com/sites/jiawertz/2023/11/19/how-ai-can-be-leveraged-for-diversity-and-inclusion/?sh=6565f7af4ee9.

18 Kate Crawford, 'Artificial Intelligence's White Guy Problem', *The New York Times*, 25 June 2016, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

If the very imagination of AI mirrors the aspirations of the white cis-male heteronormative elite that populated the universities and military industrial complexes that AI and computing emerged from,[19] then DEI in AI assumes that pre-existing, data-driven algorithmic bias might be spotted (better, earlier?) if people from marginalized and minoritized communities were involved in the high-level design and development of AI. Moreover, such a diverse workforce of decision makers might dilute the Silicon Valley monoculture that currently predominates AI futures.[20] Some organizations are bringing participatory, community-driven, embodied, and scientific approaches to DEI. These include: Black in AI, a network of Black data scientists working in AI; Our Data Bodies, a community-based research organization that investigates how digital information from marginalized communities are collected, stored, and used by governments and corporations; Data for Black Lives, a nonprofit focused on using data science for positive change in the lives of Black people; and the Carceral Tech Resistance Network, a campaign coalition against the experimental adoption and testing of technologies police, prisons, and border enforcement.[21]

Despite these efforts, aspirations to DEI in AI are also met with fatigue or scorn as principled high-level commitments betray reality. This happens when the optics of diversity becomes a proxy for actual diversity. We are reassured of DEI when we can see it, when it is visible: spotting, for instance, women speakers on a panel, or on the web page of an organization's leadership team. It is in response to this misinterpretation of DEI goals that new AI initiatives have begun to emerge. Rosebud.Ai, for instance, works in gaming, branding, and digital marketing, and offers to cut through the cost and effort of a photo shoot—and finding diverse human models—by creating synthetic images of diverse people for websites and gameworlds. In the same vein, the organizer of a 2023 tech conference created fake profiles of speakers using generative AI tools to suggest that his conference was gender-diverse.[22] The case of generative AI being used to generate fake diversity is troubling and instructive here: it alerts us to a minimization, a coloring-by-the-numbers approach to DEI, a shortcut that speaks to DEI being, in fact, something at the end of a drop down menu or a box that has to be ticked off.

19 The work of Alison Adam, whose early critical work pioneered feminist engagement with AI, is exemplary here since she asks key questions of where our notions of intelligence and rationality come from, and situates the relationship between place, embodiment, and knowing. See Alison Adam, *Artificial Knowing: Gender and the Thinking Machine*, London: Routledge, 1998.

20 In the past half decade there has been a veritable flourishing of alternative, diverse adoptions *and* refusals of AI by artists, organizers, designers, and scholars. Popular books by academics include Joy Buolamwini, *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*, New York: Random House, 2023; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press, 2018; Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge, MA: The MIT Press, 2018; and Ruha Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code*, Cambridge, UK: Polity Press, 2019.

21 Sarah T. Hamid, 'Community Defense: Sarah T. Hamid on Abolishing Carceral Technologies,' *Logic(s)* 11: Care (2020), https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/.

22 Natalie Lung and Ella Ceron, 'Developer Conference Axed After Fake Female Profiles Outcry,' *Bloomberg*, 27 November 2023, https://www.bloomberg.com/news/articles/2023-11-28/tech-conference-faces-backlash-on-claims-of-fake-women-speakers.

Even when AI companies do commit to DEI efforts to push beyond surface-level change, these are met with resentment when business decisions undermine the original pledges. For instance, when Kay Cole James, a vocal anti-LGBTQ campaigner, was appointed to Google's AI advisory council, the choice was vigorously opposed by the company's employees,[23] who argued the appointment made clear that Google's 'version of "ethics" value[d] proximity to power over the wellbeing of trans people, other LGBTQ people, and immigrants'.[24] In another example, the high-profile firing of Timnit Gebru from Google's Ethical AI team in December 2020 was seen as an attack on one of the few highly decorated black women computer scientists in AI vocal about the negative social and environmental consequences of development.[25]

Sara Ahmed writes in *On Being Included: Racism and Diversity in Institutional Life* that the work of inclusion in institutions is unrewarded, unrecognized labor often done by the very people who are affected by the lack of real DEI.[26] Diversity work is institutional transformation work, says Ahmed; she offers us rich language to consider what diversity work is and what its workers must do. She uses hard, material, infrastructural terms referring to 'sedimented' institutional practices that must be unsettled through confrontation of discrimination and lack of diversity and equity;[27] to working as 'institutional plumbers'; to being the person(s) who moves 'against the flow' of the everyday.[28] The flow of 'business as usual' that diversity work disrupts is not actually a flow, she says; those experiencing discrimination, bias, and a lack of inclusion and equity experience 'flow' as something solid. Diversity work, in other words, is about working with immobility and immobilization. And it is usually the work of those who do not quite fit into pre-existing norms, hence the requirement of DEI in the first place. Diversity work is hard work. It is precisely this type of work that cannot be encapsulated into a software development module, seamlessly fitting existing workflows. It is the kind of work that, as Widder and Nafus demonstrate, is turned into 'administrative labor no one else want[s] to do'.[29] And it is also the kind of work that new ethical AI toolkits that we analyze in what follows promise to make easier, manageable, frictionless—in other words, convenient.

Toolkitificiation of DEI in AI

In this section, we compare several blueprints, toolkits, and playbooks that aim to help AI providers meet the goals of inclusivity and equity in AI development, deployment, and governance.

23 Jillian D'Onfro, 'Google Employees Protest 'Anti-LGBTQ' Conservative's Appointment To AI Ethics Council,' *Forbes*, 1 April 2019, https://www.forbes.com/sites/jilliandonfro/2019/04/01/google-employees-protest-anti-lgbtq-conservatives-appointment-to-its-ai-ethics-council/?sh=776ce37413e1.

24 Googlers Against Transphobia, 'Googlers Against Transphobia and Hate', *Medium*, 1 April 2019, https://medium.com/@against.transphobia/googlers-against-transphobia-and-hate-b1b0a5dbf76.

25 Tom Simonite, 'What Really Happened When Google Ousted Timnit Gebru,' *Wired*, 8 June 2021, https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/.

26 Sara Ahmed, *On Being Included: Racism and Diversity in Institutional Life*, Durham, NC: Duke University Press, 2012.

27 Ahmed, *On Being Included*, pp. 175-176.

28 Ahmed, *On Being Included*, p. 186.

29 Widder and Nafus, 'Dislocated accountabilities in the "AI supply chain"', 8.

As mentioned, there is no single, agreed-upon definition of a *toolkit* that would distinguish it from other formats, such as a guideline or blueprint—terms often used interchangeably. So, for ease, we refer to all of the following as toolkits. We search for commonalities in their framings of the difficult work that Ahmed refers to as 'institutional plumbing' to examine how toolkitification makes what is uncomfortable approachable, the complex manageable, and the irresolvable frictionless.

The first among these is the *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook* (2020), produced by the Center for Equity, Gender, and Leadership at the Haas School of Business (University of California, Berkeley). The Playbook is a 62-page PDF document introducing AI business decision-makers to the matters of DEI in AI. It is accompanied by a website from which users can download the titular 'plays'—identifying necessary 'moves' ('Enable diverse and multi-disciplinary teams working on algorithms and AI systems'), relevant 'players' (such as 'C-suite' or 'Human Resources'), and useful 'tools' (such as the Alan Turing Institute's Diversity Dashboard). Next is the *Action Toolkit on Inclusive AI* (2021), developed by the Women4AI Daring Circle of the Women's Forum for the Economy & Society, in collaboration with UNESCO, the Oxford Internet Institute (OII), Shearman & Sterling, and Price Waterhouse Coopers. While the Action Toolkit is, similarly, a PDF document, it is slightly shorter (only 30-pages long), and aimed at both business and technical professionals, making a case for inclusive AI and introducing some initial measures of success, as well as linking to further, specialized tools. Then, we refer to *A Blueprint for Equity and Inclusion in Artificial Intelligence* (2022), by the Global Future Council on Artificial Intelligence for Humanity under the World Economic Forum (WEF), a 30-page long white paper presenting readers with issues related to equity and inclusion at various stages of the development cycle, and linking to both DEI-relevant tools, as well as application case studies. Finally, we also analyze *A Blueprint for Equitable AI: Building and Distributing Artificial Intelligence for Equitable Outcomes* (2023), from the Aspen Institute's Science & Society team with sponsorship from Google DeepMind. While this last, 34-page document does present ideas for potential strategies to meet DEI goals in AI, it is in fact a report summarizing insights from a series of workshops convened by the Aspen Institute team, and, as such, is the most general and least 'action-oriented' of the selected toolkits. We refer to them as the Haas, Women4AI, WEF, and Aspen toolkits, respectively.

Despite differences in the approaches proposed by these toolkits, we find similarities in how they structure the relationships between different stakeholder groups—business decision-makers, developers, policymakers, as well as users of AI systems—and individual tools and methods to achieve the goals of DEI in AI. They frame the complexity of DEI in AI in terms of several dialectical tensions: they aspire to comprehensiveness while being comprehensible to various stakeholders in the development process; being research-informed, easily digestible and jargon-free; actionable but not simplistic; and necessary yet playful.

Comprehensive (and Comprehensible)

The first common feature of the toolkits we selected is their intended audience: they are designed to be used, at least seemingly, by everyone in the AI ecosystem – not only design-

ers, data scientists, software engineers, but also business decision-makers, board members, and policymakers. The WEF toolkit, for instance, explicitly addresses 'managers and teams responsible for the different stages of AI development, as well as decision-makers from all sectors part of the AI ecosystem' and also includes suggestions for governments, while the Women4AI 'instrument' has been 'created for C-suite executives, technologists, HR managers, board members, developers, engineers and anyone who wants to change practice, policy, strategy and attitudes within their organization towards ethical, inclusive AI'. Only the Haas toolkit targets business decision-makers, but these are understood broadly, to include 'a CEO, a board member, an information / data / technology officer, a department head, a responsible AI lead, a project manager'. And yet, despite this broad spectrum of the toolkits' intended users, all four toolkits address only those who are already present at the metaphorical design table, rather than those who are still missing—the stakeholders who are most likely to bear the brunt of the negative impact of AI development and whose involvement in the design process the toolkits are supposed to encourage.

Related to this seemingly broad spectrum of intended users is another consistent feature of our selected DEI toolkits: they all acknowledge that the challenge of making AI more equitable and inclusive extends beyond technical questions of data bias and needs to be addressed comprehensively at different stages of the development process and in different parts of the AI ecosystem. This is key because the conception of DEI that the selected toolkits embody does not amount to 'bias-eradication'—the toolkits are nowhere near as simplistic and take DEI work seriously, recognizing that it is hard work that must happen at various stages of the design, development, and deployment process. While the WEF toolkit aims to 'paint a comprehensive picture of challenges and opportunities for improvements in equity and inclusion across the AI development life cycle and governance ecosystem', the Haas toolkit notes it is precisely because of the need to intervene at all the stages of the production pipeline that 'addressing bias in AI is an issue for business leaders' rather than technical workers, requiring changes in hiring practices, among others. Women4AI similarly promises to guide users through the necessary steps in transforming both the 'organizational culture' and aspects of the design and development practice, encouraging design teams to be 'as close as possible to the populations likely to use it or be affected by it', while the Aspen toolkit provides suggestions ranging from concrete tips for changing the AI production process, such as 'embedding the topic of inclusivity into training for development team members', to more general ones, such as 'preparing young people for AI through ethical tech education'. Because of this promise of comprehensiveness, the toolkits conflate design with policy: they merge different genres, methods, and perspectives to ensure that they remain the only necessary entry point to the question of inclusive and equitable AI for anyone—business decision-makers, regulators, or software developers.

As the toolkits aspire to comprehensively tackle issues of equity and inclusivity, it is crucial to highlight that each toolkit serves as a meta-toolkit, linking to more specialized tools, methods, and guidelines. For instance, the WEF toolkit links its users to the *AI Fairness Global Library*, where 'other knowledge resources from leading institutions can be found to deepen the topics presented', while the Women4AI toolkit includes a list of other toolkits for 'technical audiences that seek to improve the ethical and inclusive practices of AI systems', including

the Microsoft's Responsible Innovation Toolkit and PwC's Responsible AI Toolkit. The apparent convenience of equitable AI toolkits is related precisely to this conception of comprehensiveness: they are meant to be designed for *everyone*, include information on *all aspects* of the equity and inclusivity challenge, and gather (or link to) *everything*—all the necessary tools and methods—one requires to tackle this challenge. This *comprehensiveness* the toolkit creators have in mind is meant to acknowledge the complexity of the issues at stake, but not foreclose their *comprehensibility*; on the contrary, *comprehensiveness* in this context becomes synonymous with *comprehensibility*. It implies not information-overload, but total parse-ability, making the questions of DEI 'accessible' to actors not usually burdened with DEI-related work.

Informed (but 'jargon-free')

The selected toolkits are informed by a vast amount of research and consultations with experts, and the toolkits' creators make this critical work purposefully explicit, detailing the processes that led to the toolkits' creation. This suggests rigor. The Aspen toolkit, for instance, presents the insights drawn from discussions of 'two diverse groups of experts', including the legal scholars Lilian Edwards and Sandra Wachter, and the data scientist Cathy O'Neil—well known for their work on technology regulation, privacy, and feminist data science. The Haas toolkit similarly draws from 'academic literature and experts across disciplines – spanning sociology, philosophy, engineering and more', including the sociologist Gina Neff, the computer scientist Stuart Russell, and the Managing Director of the AI Now Institute Sarah Myers-West. The WEF toolkit, in turn, was created, by the Global Future Council on Artificial Intelligence for Humanity, whose members include Angie Abdilla, specializing in indigenous knowledges and their relation to technology production, and Safiya Umoja Noble, the author of *Algorithms of Oppression* (2018). The critical work of these scholars who are dedicated to exploring how machine learning-based technologies reproduce or exacerbate social inequities would certainly be considered 'inconvenient' by some decision-makers within technology companies because they are radical in the sense of identifying the *root* of inequity, bias, and discrimination. To pull things out by the roots is the sort of work that Sara Ahmed refers to as confronting the 'sedimentation' of institutional practice that does not acknowledge or make room for diversity.

Here comes the toolkit with its promise of 'translation' between disciplines and negotiation between different, potentially conflicting goals. The WEF toolkit, for example, promises to map 'the vast amount of equity and inclusion challenges' in the AI production and governance ecosystem to then integrate them into 'a *digestible* framework' (our emphasis), while the Haas toolkit sets itself apart from other available tools by promising to do the 'crucial translational work' and present 'conversations around "bias" in AI'—which can be, as it turns out, 'muddled and mean or refer to various concepts'—in a format that is 'jargon-free and comprehensive'. The toolkit may refer to key AI ethics experts and institutions to legitimize itself as an instrument of pro-justice change in AI. Yet, the very act of 'translation' or 'adaptation' of the critical insights derived from critical AI ethics work for the purposes of corporate change can, inadvertently, result in the 'taming' of critical, and often radical, positions; likely ridding them of their transformative potential. The logic of the toolkit is that inconvenient or upsetting perspectives—including the views of critical AI scholars whose views informed the development of the equity toolkits in the first place—are toned down, made appealing, and bearable.

Necessary (but Playful); Actionable (but not Simplistic)

The selected toolkits strategically frame the work they facilitate as *essential*, rather than optional. The Aspen toolkit says, '[p]ausing technological development and deployment until all concerns are addressed is not feasible', yet it is 'critical to ensure that processes and institutions exist to champion and implement efforts toward achieving equitable outcomes.' This necessity extends beyond societal value; as the toolkits suggest, making AI more inclusive translates to business value. The Women4AI toolkit highlights that equity and inclusivity in AI development is 'ultimately about helping your organization avoid the risks from biased outcomes and reap the rewards from economies and societies which increasingly expect inclusion as standard'. Likewise, the Haas toolkit underscores that using the tools and 'plays' it collects to mitigate bias in AI is crucial 'to unlock value responsibly and equitably'. If inclusive AI is good for business and if toolkits can help achieve inclusive AI, then it follows that the toolkits are good for business, too. There is a game-like quality to DEI in AI that the toolkits encourage, leveling up as a strategy to unlock rewards for business. The toolkits manage their intended users' expectations, recognizing that not all aspects of the DEI-fulfilling AI challenge are easily and immediately addressable. The creators of the Haas toolkit point out that 'de-biasing' AI fully is 'not achievable', while the Women4AI toolkit acknowledges that developing inclusive AI 'is a journey, not a destination' and the toolkit can only serve as a starting point. Yet, the toolkits tend to strategically highlight what *is* immediately solvable: the Haas toolkit, for instance, suggests that using it will lead to some 'quick wins', as it introduces its users to resources that can have concrete, immediate effects on the AI development pipeline. The toolkits' pedagogy: ensuring that business executives and developers get rewarded on the journey towards inclusive, equitable AI with 'easy wins' and 'low-hanging fruit'. Because if these were not in sight, if there was no promise of eventual satisfaction and fun along the way, the intended users of the toolkits could get discouraged and fail to persevere. Even when the toolkit is a rather dense conference report, its appealing design serves as a promise of both convenience and joy—even if this promise remains unrealized (and, perhaps, unrealizable).

Each of these sets of tensions in the toolkits' framing suggest a desperation wrapped up in a sincere commitment to DEI in AI. Perhaps because DEI is difficult, the authors want to encourage rather than repel the potential toolkit user. Hence promises of comprehensiveness *and* comprehensibility, of simplicity but not being simplistic, of action *and* play, are like treats to lure or even trick the user down the path of DEI; like honey to coat an oddly-shaped, hard-to-swallow pill.

On making AI Ethics Inconvenient

Toolkits are now-ubiquitous material-cultural informational artifacts that organize many of our shared organizational, political, and institutional work. Their ubiquity does not make them benign or mundane, however. In 2021, the Indian government arrested a 22-year-old climate activist and founder of the Indian chapter of Fridays for the Future, Disha Ravi, for assembling an online toolkit for social media action and mobilization to support farmers who had been protesting against the Farm Bills for over a year, camped out on the outskirts of and in the

capital, New Delhi.[30] The Indian government charged Ravi with 'collaborating' to 'spread disaffection against the Indian state' and sedition.[31] Ravi shared the document with Greta Thunberg who tweeted about it, which angered the state even further. In 2011, when former Brazilian president Jair Bolsonaro was a congressperson, he launched a campaign against a school-based education program to combat homophobia, arguing that the distribution of information packages aka toolkits, which he called 'gay kits', in schools might actually 'turn' children gay through exposure.[32] Toolkits can take on many forms, and their potential for political action owes to the speed with which they promise the replication of ideas at scale. These two instances demonstrate how their convenience can be perceived as inconvenient.

We bring that spirit to this critique. Furthermore, at the time of this writing, one of us is designing a toolkit for software developers to fulfill the requirements of the EU AI Act associated with high-risk applications of AI.[33] This analysis of toolkits as convenient media therefore has been developed in parallel with a deep engagement with the affordances of this form, and how its limits might be tested to maintain inconvenience. 'Inconvenience' does not require that toolkit design be user-unfriendly, its messaging pessimistic, or that the form be abandoned altogether. In our practice, we find that it means reconfiguring the system of rewards that the toolkit embodies, ensuring that 'user satisfaction' does not hinge solely on ticking a box or marking a task as complete; it means highlighting that compliance is the bare minimum, a starting point; practically, it means that any task or step that the toolkit incorporates is followed by a 'go further' section—suggesting that there is always more to be done and that the toolkit users can and should do more. A toolkit must inspire an 'ongoing-ness' of work. It also means moving beyond the logic of modularity in where ethics work happens and how it fits within existing workflows; it means facilitating reflection on the complexity of the issues. So, in practical terms, we move away from self-contained 'modules'—sets of tasks to be completed by different teams in a predetermined sequence—and into 'spaces': interconnected areas of concern that different stakeholders must pass through and continue coming back to, throughout software development and deployment. It means highlighting, rather than gliding over, 'inconvenient questions'—for instance, about the end user's meaningful consent and what feminism and decolonial theory can teach us about its elicitation through design.[34]

30 Wikipedia contributors, '2020–2021 Indian farmers' protest', *Wikipedia,* https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest, accessed 22 February 2024.

31 'India activist Disha Ravi arrested over farmers' protest "toolkit"', *BBC News*, 14 February 2021, https://www.bbc.co.uk/news/world-asia-india-56060232.

32 Ed Bracho-Polanco, 'How Jair Bolsonaro used 'fake news' to win power', *The Conversation*, 8 January 2019, https://theconversation.com/how-jair-bolsonaro-used-fake-news-to-win-power-109343.

33 See 'The In-depth EU AI Act Toolkit', *Leverhulme Centre for the Future of Intelligence*, http://lcfi.ac.uk/projects/ai-innovation-praxis/eu-ai-act-toolkit/. The Act classifies AI systems according to different levels of risk they may pose in predefined areas of applications; a system can be classified as a source of high risk if it has potential to adversely impact people's health, safety, or fundamental rights (such as dignity or equality) in predefined areas of use, including biometric identification, law enforcement, and recruitment; producers of such systems are required to implement measures to mitigate the AI system's undesirable societal consequences.

34 Joana Varon and Paz Peña, 'Artificial intelligence and consent: a feminist anti-colonial critique', *Internet Policy Review* 10.4 (2021), https://policyreview.info/articles/analysis/artificial-intelligence-and-consent-feminist-anti-colonial-critique.

Equity, justice, and fairness are rich in friction; they demand a historical, structural, and institutional reckoning; and constant and passionate engagement with actual diversity—of thought, experience, situation, values—with little reassurance of total success. The toolkits we have referred to do acknowledge the complexity of the DEI challenges in AI development, deployment, and governance; as the Aspen toolkit makes clear, there are no 'silver bullets to ongoing challenges'. And yet, despite this recognition highlighted as a premise for DEI work in the AI ecosystem, the messaging of the toolkits for inclusive AI—precisely because they are framed as *toolkits* rather than *reports* or even *guidelines*—points to the paradox that the toolkitficiation of DEI work, and ethics more broadly, is necessarily ridden by. There might be no simple solutions to the problem that structural injustice constitutes, and yet a *toolkit* implies the existence of such ready-made *tools*; a *blueprint* suggests that the boundaries and hierarchies of the AI ecosystem can still be *redrawn*, as long as the teams, companies, and governments adhere to a clear-cut inclusivity template; a *playbook* signals that there are tested tactics and methods, a means of redirecting activity to achieve the desired (and desirable) outcome.

Toolkits might acknowledge complexity and difficulty, but their logic remains that of actionability and convenience. None of the struggle Ahmed talks about—the negotiation with immobilization, culture, language, established practice, human social relations, or organizational workflows—are in evidence here. But meaningful ethicopolitical life is, as Louise Amoore argues in *Cloud Ethics*, precisely about 'irresolvable struggles, intransigence, duress, and opacity, and it must continue to be so for if a future possibility for politics is not to be eclipsed by the output signals of algorithms.'[35] Amoore alerts us that there is more than just the matter of how, and if ,DEI in AI toolkits can deliver on empowering designers and decision-makers to produce ethical technologies: the *expansion* demanded of algorithms to parse, process, and encompass the intricacies and entanglements of human social life, to make accountable, ethical, fair, unbiased, and trustworthy decisions, cannot happen in a vacuum. The algorithms require frequent human intervention to maintain and manage their behavior; this human intervention, in turn, requires its own constant tending-to within shifting and unequal social, cultural, institutional, and organizational arrangements. This is far from smooth, this is hard to automate, but this is the work of our time.

References

'India activist Disha Ravi arrested over farmers' protest "toolkit"', *BBC News*, 14 February 2021, https://www.bbc.co.uk/news/world-asia-india-56060232.

Smith, Genevieve and Rustagi, Ishita. *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook*, Berkeley, CA: Berkeley Haas Center for Equity, Gender and Leadership, 2020, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf.

'The In-depth EU AI Act Toolkit', *Leverhulme Centre for the Future of Intelligence*, http://lcfi.ac.uk/projects/ai-innovation-praxis/eu-ai-act-toolkit/.

Adam, Alison. *Artificial Knowing: Gender and the Thinking Machine*, London: Routledge, 1998.

35 Louise Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*, Durham, NC: Duke University Press, 2020, p. 172.

Ahmed, Sara. *On Being Included: Racism and Diversity in Institutional Life*, Durham, NC: Duke University Press, 2012.

AI Incident Database, https://incidentdatabase.ai.

AIAAIC Repository, https://www.aiaaic.org/aiaaic-repository.

Amoore, Louise. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*, Durham, NC: Duke University Press, 2020.

Aspen Institute Science & Society Program. *A Blueprint for Equitable AI: Building and Distributing Artificial Intelligence for Equitable Outcomes*, Washington, DC: The Aspen Institute, 2023, https://www.aspeninstitute.org/wp-content/uploads/2023/01/Equitable-AI-Aspen-Institute.pdf.

Bastian, Rebekah. 'AI Brings Opportunities And Risks To Workplace DEI Efforts,' *Forbes*, 8 May 2023, https://www.forbes.com/sites/rebekahbastian/2023/05/08/ai-brings-opportunities-and-risks-to-workplace-dei-efforts/?sh=4614ed8b4b2a.

Benjamin, Ruha. *Race after Technology: Abolitionist Tools for the New Jim Code*, Cambridge, UK: Polity Press, 2019.

Bracho-Polanco, Ed. 'How Jair Bolsonaro used 'fake news' to win power', *The Conversation*, 8 January 2019, https://theconversation.com/how-jair-bolsonaro-used-fake-news-to-win-power-109343.

Broussard, Meredith. *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge, MA: The MIT Press, 2018.

Buolamwini, Joy. *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*, New York: Random House, 2023.

Crawford, Kate. 'Artificial Intelligence's White Guy Problem', *The New York Times*, 25 June 2016, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

D'Onfro, Jillian. 'Google Employees Protest 'Anti-LGBTQ' Conservative's Appointment To AI Ethics Council,' *Forbes*, 1 April 2019, https://www.forbes.com/sites/jilliandonfro/2019/04/01/google-employees-protest-anti-lgbtq-conservatives-appointment-to-its-ai-ethics-council/?sh=776ce37413e1.

Global Future Council on Artificial Intelligence for Humanity. *A Blueprint for Equity and Inclusion in Artificial Intelligence*, Geneva: World Economic Forum, 2022, https://www3.weforum.org/docs/WEF_A_Blueprint_for_Equity_and_Inclusion_in_Artificial_Intelligence_2022.pdf.

Googlers Against Transphobia, 'Googlers Against Transphobia and Hate', *Medium*, 1 April 2019, https://medium.com/@against.transphobia/googlers-against-transphobia-and-hate-b1b0a5dbf76.

Hagendorff, Thilo. 'The Ethics of AI Ethics: An Evaluation of Guidelines,' *Minds & Machines* 30 (2020): 99-120.

Hamid, Sarah T. 'Community Defense: Sarah T. Hamid on Abolishing Carceral Technologies,' *Logic(s)* 11: Care (2020), https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/.

Hollanek, Tomasz. 'The Ethico-politics of Design Toolkits: Responsible AI Tools, From Big Tech Guidelines to Feminist Ideation Cards,' forthcoming.

Intersectional AI Toolkit, https://intersectionalai.miraheze.org/wiki/Intersectional_AI_Toolkit.

Lung, Natalie and Ceron, Ella. 'Developer Conference Axed After Fake Female Profiles Outcry,' *Bloomberg*, 27 November 2023, https://www.bloomberg.com/news/articles/2023-11-28/tech-conference-faces-backlash-on-claims-of-fake-women-speakers.

Manovich, Lev. *The Language of New Media*, Cambridge, MA: The MIT Press, 2001.

Neves, Joshua and Steinberg, Marc. 'The Cultural Politics of In/Convenience', *Global Emergent Media:*

In Progress, January 2023, https://www.globalemergentmedia.com/in-progress/the-cultural-politics-of-in%2Fconvenience.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York: New York University Press, 2018.

OECD, 'Catalogue of Tools & Metrics for Trustworthy AI', *OECD.AI Policy Observatory*, https://oecd.ai/en/catalogue/tools.

Peters, Dorian; Loke, Lian; and Ahmadpour, Naseem. 'Toolkits, cards and games – a review of analogue tools for collaborative ideation', *CoDesign* 17:4 (2020): 410-434.

Posner, Miriam. 'Breakpoints and Black Boxes: Information in Global Supply Chains,' *Postmodern Culture* 31:3 (2021), https://www.pomoculture.org/2021/12/01/breakpoints-and-black-boxes-information-in-global-supply-chains/.

Rizvi, Jia. 'How AI Can Be Leveraged For Diversity And Inclusion,' *Forbes*, 19 November 2023, https://www.forbes.com/sites/jiawertz/2023/11/19/how-ai-can-be-leveraged-for-diversity-and-inclusion/?sh=-6565f7af4ee9.

Simonite, Tom. 'What Really Happened When Google Ousted Timnit Gebru,' *Wired*, 8 June 2021, https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/.

Varon, Joana and Peña, Paz. 'Artificial intelligence and consent: a feminist anti-colonial critique', *Internet Policy Review* 10.4 (2021), https://policyreview.info/articles/analysis/artificial-intelligence-and-consent-feminist-anti-colonial-critique.

Widder, David Gray and Nafus, Dawn. 'Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility,' *Big Data & Society* 10:1 (2023).

Wikipedia contributors, '2020–2021 Indian farmers' protest', *Wikipedia,* https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest, accessed 22 February 2024.

Women4AI Daring Circle. *Action Toolkit on Inclusive AI*, Paris: Women's Forum for the Economy & Society, 2021,file:///Users/mi373/Documents/2024/writing https://storageprdv2inwink.blob.core.windows.net/3b7997a1-b9ec-4c99-89ea-f5199663f903/b9c9f204-6bec-41d8-8453-05aaf4328ff5.

Wong, Richmond Y.; Madaio, Michael A.; and Merrill, Nick. 'Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics,' *Proceedings of the ACM on Human-Computer Interaction* 7, Issue CSCW1 (2023).