# Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility

David Gray Widder[1] and Dawn Nafus[2]

**Abstract**

Responsible AI guidelines often ask engineers to consider how their systems might harm. However, contemporary AI systems are built by composing many preexisting software modules that pass through many hands before becoming a finished product or service. How does this shape responsible AI practice? In interviews with 27 AI engineers across industry, open source, and academia, our participants often did not see the questions posed in responsible AI guidelines to be within their agency, capability, or responsibility to address. We use Lucy Suchman's notion of located accountability to show how responsible AI labor is currently organized, and to explore how it could be done differently. We identify cross-cutting social logics, like modularizability, scale, reputation, and customer orientation, that organize which responsible AI actions do take place, and which are relegated to low status staff or believed to be the work of the next or previous person in the chain. We argue that current responsible AI interventions, like ethics checklists and guidelines that assume panoptical knowledge and control over systems, could improve by taking a located accountability approach, where relations and obligations intertwine and incrementally add value in the process. This would constitute a shift from "supply chain" thinking to "value chain" thinking.

## Introduction and Background

Many big technology companies now have responsible AI[*] programs (Jobin et al. 2019), but those tasked with "owning" these programs are limited in their ability to create change (Metcalf et al. 2019). Even those without designated ethics roles, for whom this work is likely a smaller part of their job, are called to follow responsible AI guidelines (Jobin et al. 2019), checklists (Madaio et al. 2020), and other processes (Sirur et al. 2018). Outside of the biggest vertically integrated companies that build and deploy their own user-facing systems in-house, many engineers operate at arm's length from their firm's immediate customer, who might themselves be multiple steps from an actual deployment involving data subjects or end users. How is responsibility and agency socially organized for AI practitioners in these distanced, distributed "supply chain"-like arrangements? What can be done in situations where responsibility is framed as work, whether as checklists to be filled in or as efforts to account for the needs and interests of non-customer stakeholders, and where this work risks falling through the cracks between actors?

We investigate how AI practitioners scope their agency and responsibility to address possible AI harms. In 27 interviews, our participants relay how they are increasingly asked to account for harms their systems may enable, despite seeing these questions as beyond their agency, capability, or responsibility to address. We were struck by the deeply dislocated sense of accountability, where acknowledgement of harms was consistent but nevertheless another person's job to address, almost always at another location in the broader system of production, outside one's immediate team. Here we suggest that commitments to modularity as an ideal form of technical practice, and the divisions of labor that practice entails, re-inscribe beliefs about software production as a kind of supply chain, where developers recognize their dependence on others' code as a kind of inert object, much like a shipment of goods. They might be necessary supplies, but not exactly where a deep collaborative relationship might develop. These developers had a harder time recognizing how multiple parties adding incremental value has combinatory effects that can lead to better or worse social impact. Where those recognitions did happen, it was typically through social locations that cross-cut, or were separate from, the imagined supply chain of AI software production, like reputations and empathy for the end user. Those same locations can also be used to rebuild responsible AI practices in ways that both recognize the limitations that developers currently experience, and build the inter-firm networks of relationships that can build societal and commercial value in ways analogous to "value chains," which orchestrate activities, whether internal to a company or through outsourcing, that combine to create a shared competitive advantage  Feller et al. (2006).

Other work has similarly shown how engineers do not consider hard-to-modularize aspects of business relations within their scope of work or ability to consider (Orr and Davis 2020). Greene et al. (2019) showed that big tech's responsible AI programs tend to focus scrutiny on AI system design instead of more threatening consideration of the business purposes these systems enable, and

[1] School of Computer Science, Carnegie Mellon University
[2] Intel Labs, Intel Corporation

**Corresponding author:**
David Gray Widder, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.
Email: dwidder@cmu.edu

[*] We generally use "responsible AI" to denote our commitment to feminist theories of technology (Haraway 1991), where ethics cannot be removed from the question of "to whom?" does one owe a response. Occasionally we use "ethical AI" as a synonym where the context makes it appropriate.

Widder et al. (2022) showed that interventions to make systems fair, accountable and transparent are not remedies for harms stemming from how they are used, in that case, nonconsenusal pornography. Current responsible AI interventions, like checklists, model cards, or data sheets ask practitioners to map the technology to their end use. They attempt to put "out of scope" harms back in scope, but here we show how contemporary software production practice works against that attempt, leaving developers in a bind between countervailing cultural forces.

We use "modularity" to refer both to a specific technical practice, and a broader set of cultural beliefs, epistemologies, and organizational arrangements that modularized software mediates and reinforces. The two are inseparable. Modularity has been a staple of software development since the 1970s, where large software systems are decomposed into smaller, self-contained parts. The purpose is to control parts of a system without needing to address the myriad details of the other parts (Shaw 2011). This "information hiding" (Parnas 1972) buries "the complexity of each part behind an abstraction" (Baldwin et al. 2000, p. 64). From a work practice perspective, this facilitates a division of labor and the matching of individual skills to specific tasks (Shaw 2011) by separating concerns of different workers (Tarr et al. 1999; Dijkstra 1982). The goal is to minimize friction as the code passes through many hands, and therefore speed development. Support for benefits of modularity in practice is mixed (Kemerer 1995). Modular software may be less likely to need repair (Vessey and Weber 1983; Martini et al. 2016) and be more easily repaired when errors are found (Korson and Vaishnavi 1986), and insufficiently independent modules may require more changes later (Troy and Zweben 1981). However, too many smaller modules can also cause problems (Basili and Perricone 1984; Shen et al. 1985), some blaming this on error-prone inter-module interfaces (Kemerer 1995), or difficulty in changing these interfaces due to information hiding (Kästner et al. 2011). Nonetheless, large open source projects purposefully strive for modularity in hopes of making their code more understandable (MacCormack et al. 2006), and surveys of professional software engineers see "improved modularity" as a benefit of refactoring their code (Kim et al. 2014).

Many have noted that this divided labor, inscribed in code itself, has enormous cultural and social implications, including the inability to recognize the consequences of technologies by refusing a relationship between "us" developers and "them" technology users or citizens (McPherson 2018; Suchman 2002). Modularity constitutes an epistemic culture (Cetina 1999) that cultivates a capacity to "bracket off", as Malazita and Resetar note (2019) even when that which is bracketed off is not another piece of code, but a human being who might be harmed. Modularity is an everyday, practical form of the modernist fallacy of the separability of society from technology (Latour 1993), and, by extension, separability of code from the harms it enables. It also can be seen as an example of the broader social organization of ignorance (Proctor and Schiebinger 2008), where ignorance is not a mere property of attention or intellectual capacity, but the sanctioning and legitimation of one thing as a proper arena for action (the workings of a single portion of code) at the expense of another (the activities of other developers and users). When ignorance is socially organized it is not total, but situational. In this study developers showed themselves to be perfectly capable of being aware of harm once they step outside their role as a software engineer. While there are many social dynamics that feed into dislocated accountability, including the crude facts of profit incentives, modularity is a key touchstone at the center of technical practice that serves as a lens through which these other matters are framed.

As an analytical concept, the AI supply chain has been used to question responsible AI interventions, like checklists and disclosures, that gloss over context (Gansky and McDonald 2022). We rely extensively on Lucy Suchman's concept of located accountability (Suchman 2002) to reframe responsibility away from a matter of every participant having to anticipate every conceivable outcome by diligently following elaborate checklists and disclosing every possible misuse (what Gansky and McDonald call "metadata maximalism" 2022), toward a set of practices that build on people's partial, situated knowledges (Haraway 1991). As Suchman put it, this requires a shift "from a view of design as the creation of discrete devices, or even networks of devices, to a view of systems development as entry into the networks of working relations" (Suchman 2002, p. 92). Which networks of working relations developers believe they have, and which ones could they have, is a central matter of concern. Seeing matters in terms of a chain of relations enables a view from somewhere–a specific social location (Suchman 2002; Haraway 1991)– that opens the possibility of seeing where action can take place, situating even relatively "general purpose" AI libraries or frameworks in the context of the downstream uses and harms they potentiate or constrain, even if use cannot be fully anticipated or controlled (Lally 2021). While current software supply chains hardly facilitate visibility into different actors' decisions, thinking of software production as a chain nonetheless draws our attention to the tenuous connections and relationships to be found between actors, which can be strengthened and leveraged as sites where ethical debt (i.e. Fiesler and Garrett 2020) either accrues or can be reduced.

The AI *supply* chain is an emic social reality. Developers see their work as "near" or "far" to the end user or general public, believing that more "downstream" one's work is, the more visible and pressing ethical consequences become, as code is more visibly built around a specific end use (see Figure 1). Obligations and dependencies also look different depending on whether one is looking upstream or down, and it is crucial to recognize these social locations when creating deeper expectations of responsibility. We show below that this social reality has created conditions where interventions fall through the cracks between actors, and has defined other chains of relations (business, personal reputations, user experience, etc.) as secondary or out of scope. These latter types of chains, or assemblages of people, practices, and materials, are most definitely "in scope" from an STS perspective. For example, following Latour's 1999 "chains of translation," Carolan (2020) examines data chains that tie together the distributed precision agriculture industry in non-linear, recursive and contested ways. A developer's nightmare to be sure! Nevertheless, when we say "supply chain," we are referring to developers' current social organization into upstream and down, and when we speak more broadly of "chains" or occasionally "value chains," we are using it in the way Carolan does to denote heterogenous, cross-cutting, not always linear social interactions and relations that occupy multiple social locations and cultural logics at the same time. For responsibility to be socially located or meaningfully grounded, there needs to be a way of talking about the development of technological systems as "a boundary-crossing activity, taking place through the deliberate creation of situations that allow for the meeting of different partial knowledges" (Suchman 2002). Thinking through "chains," without taking at face value the linear "supply" framing, is one way to do that.

To conduct this study, we recruited using public emails and existing contacts alongside paid services and snowball sampling. Our 27 participants spanned North America, Asia, Europe, and Africa, included employees from eight companies from small startups to multinationals with more than 100k workers, four researchers from three universities, and seven open source contributors across six projects. Many had ML-related graduate degrees; all had AI expertise in a variety of guises: Machine Learning Engineers, Research Scientists, Developer Experience Researchers, System Integrators, and Project Managers. All identified as men except one woman, reflecting disparities in the AI workforce (Zhang et al. 2021). Each were invited to a semi-structured recorded teleconference interview, which was then professionally transcribed, except for one participant who preferred that we take notes. Most interviews lasted an hour, but were as short as 30 minutes or as long as two hours. Data were analyzed iteratively including a writing a descriptive memo after interviews, a running analytic memo as a reflexive account of the first author's understanding of emergent themes (Strauss and Corbin 1997), and weekly discussion of these themes between the authors. Our analysis is inflected by our positionalities, with one of us (omitted for blind review) a responsible AI practitioner at a large company, and another (omitted for blind review) an early career software engineering scholar.

First, we illustrate how a distributed AI supply chain limits three developers' sense of agency and responsibility. We then show how divisions of labor and a reverence for scale produce disconnections and separations along the AI supply chain, and how reputational concerns and corporate maxims for "passion for the user" create countervailing points of connection. We then show how the current configuration of connection and disconnection shapes the ethics work that is and is not done. Finally, we present three potential interventions, depending on one's view about whether modularity is an ideal to be preserved or a problem to be overcome.

## Views from Up and Down the AI Supply Chain

Outside of the largest technology companies, complex inter-organizational relationships are at the heart of building AI (Thomas 2019). For example, computer vision used in a power plant's surveillance system to detect a person at its perimeter might begin its metaphorical life published as academic research, further developed and made freely accessible in an open source library as a pretrained model, later requiring *in situ* training when deployed to work with the plant's existing hardware and software by a systems integrator. It might be further adapted if the plant has the requisite expertise. Thomas (2019) observes that by 2018, computer vision professionals expected to not need to build systems from scratch, with open source tooling and pretrained algorithms available to "kick start their work," and find a role somewhere in the chain.

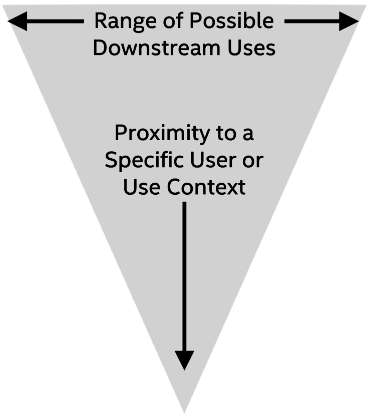
Range of Possible Downstream Uses
Proximity to a Specific User or Use Context

**Figure 1.** Participants perceive that when they work closer to a specific end use context, the range of possible downstream uses for their software module narrows.

Within these chains, participants believe control over their system's impacts increases as possible uses of the system narrows, as it is adapted to fit a particular end use. Higher in the AI supply chain are supposedly general-purpose research outputs or tools, such as an academic ML researcher relaying his enjoyment in *"discovering generalized infrastructure components that are missing from people's workflows."* Aiming to regularize ML model accuracy, his team *"proved out here was a procedure [...] a new way to optimize your machine learning model and depending on the data set you use, the application domain you pick can be potentially endless,"* later saying *"this is a general procedure [...] it's not prescriptive"* in how it is used. Separability between the optimization procedure and what is optimized characterizes a view from the top of the supply chain. The celebration of endless possibilities for downstream use makes harms appear to be a general, unconnected matter. Asked if there are ways his project could be used that would concern him, he answered: *"nothing that would concern me [except] general ways in which you can abuse machine learning. [...] You can set up stuff in a way that is screwed up [...] I don't think it does anything that can be abused relative to what you could do normally with any machine learning algorithm."* This leaves invisible his direct contribution to the "optimization" of harm by enabling it to occur in a more technically optimal manner. Indeed, people working high in the supply chain were particularly prone to employ discourses of technological neutrality (i.e. Winner 1980; Latour 1994; Ihde 1990; Davis 2020; Orr and Davis 2020), referring to what they make as not even comparable to the gun that proverbially can used for both harm or good, but the machinery which makes multi-use parts: *"I make a piece* *of equipment that makes pipe, somebody bought my pipe making equipment, and made the barrel of guns. I don't know how I stop [harm], because I didn't make the gun."*

In the middle of the supply chain lie partial systems like performance benchmarks or pre-trained models, designed to show off accuracy, speed, or ease of use, to entice others to use them as "kick starts" (Thomas 2019) for future finished deployments. These contexts make upstream dependencies and downstream responsibilities more visible. For example, another engineer used *"a composition of already existing components"* from an open source framework and models to develop machine translation*"benchmarks," "showcase[s],"* and *"demo[s],"* which he also made available as open source. Because he did not build the framework, he stated *"it's a part of open source project so [...] we are not taking the full responsibility for the framework itself,"* downplaying whether he had any choice whether to use, it or vet it for problems. Looking downstream he stated: *"there is a very little interest in the actual – maybe not quality of translation, but the form, the meaning of translation, but rather [more interest in] the performance numbers,"* such as translation speed or accuracy. Because the output is not considered a final, end-user-facing matter with real consequences, he does not consider it his job to address biases: *"I don't believe that anyone will try to prove that, hey, the output is biased."* While he was somewhat concerned that his company's logo would be attached, he expected the next person in the chain to know to address it: *"if he's (sic) an engineer working [in the] machine translation area, he or she is aware of what the bias means."* This expertise is what rendered it yet another module, and harms a "general" problem: *"there is always a risk that the translation can be biased."* In our interviews, participants frequently used passive voice to describe decisions they could have made otherwise, for example, this person saying, *"the data input* ***was taken*** *from official available sources,"* and *"existing components,* ***which are*** *packed and prepared."* He points to the least "general" actor in the chain as the site of responsibility: *"I believe that the final responsibility lies at the client's side who is finally deploying the actual service."* Notably, these felt like statements of fact, not attempts to be exculpatory. The participant began the interview apologetically, explaining that his *"very simple"* project provided little to share that might be helpful to research on ethics.

Lower in the AI supply chain, an AI model is integrated into "live" software. Here, harms are closer and more visible, but it is still considered a virtue for software engineers to be able to focus on their technical work, without interacting with those using their software. For example, a tech lead at a company building VR services for defense industry clients explained things that could affect his company's reputation: *"It's a concern to me because there could be flaws in the code, security risks, quality risks, and effectively, if anything goes wrong, it looks bad on us."* Nevertheless, he talks about the separation of engineers from colleagues that handle customer interaction as a kind of relief, that he *"kind of get[s] to turn a blind eye to certain social aspects"* because *"we have program managers that tend to be the buffer [between us and the user]."* He says sometimes he gets pulled in to customer conversations but they are improving the process to make sure *"I'm not involved, because frankly, I shouldn't be."* If software engineers building the software might have an issue with their work being used to train military drone pilots, this separation insulates them from intimate knowledge of this use.

At this level in the supply chain, the design affordances of use are more readily acknowledged. In this example, the participant was confident that his *"app isn't so open ended that it can just be used [...] by accident in a different way,"* but reusable components could open the door to nefarious use: *"you could reverse engineer what we do and then repurpose it [to] insert things that maybe would encourage someone to commit suicide."* The difficulty of doing that gives him some comfort.

Looking upstream, he is uncomfortable that his product depends on Facebook's Oculus: *"we're kind of putting our foundation on sand"* because *"we distribute them via this platform, but the platform is not owned by us. The platform is owned by Facebook and Facebook just very recently had a pretty bad day. So frankly, we don't trust them, and we distribute our applications using them and we use their hardware."* In essence, he is aware that he is vulnerable to the real possibility of having to pay his supplier's ethical debt (Fiesler and Garrett 2020). Looking downstream, he is also aware of the care that needs to be taken with respect to which customers he does business with. He states, *"We're not going to have a random Tom, Dick or Harry come in, buy our products and begin using it. There's always going to be some level of let's say customer qualification,"* which demonstrates how working lower in the supply chain allows for tighter control over which paying customers "qualify" to use his service – an opportunity not felt higher in the supply chain.

Modularity imagines a blissful ignorance of customers and suppliers, a kind of nowhere in which the speaker's module sits. This "view from nowhere" (Haraway 1991), then, is a special case of a view from somewhere. It is the dominant view taken by developers to legitimize, in the name of good software development practice, avoiding entangling themselves in relationships outside of the module. As Strathern (2002) reminds us, claims that technologies need to be set in some context already tell us about the context they are, in fact, in: one believed to be freed of social relations. Here, modularity creates the numerous ways that responsibility is not to be found "here" regardless of where "here" is. Context is perennially displaced to elsewhere. Yet a modularized supply chain is clearly not the only thing going on in these examples. Choices about customers are made, ethical debts are accrued and paid, which are two seemingly impossible things if people really were strictly acting in a fully modularized manner.

A simple account of what happens upstream versus down, then, will not do, because responsibility is still displaced at the most downstream site in which it is otherwise acknowledged. Instead, to consider which interventions could be effective, we ask what social dynamics and sensibilities contribute to the disconnections from one module to the next, and what strengthens or weakens connections between points in these chains.

## Producing Disconnection

Suchman's classic (Suchman 2002) is a powerful critique of the ways that technology companies keep people separate from one another. That critique largely still stands in an AI context, where divisions of labor combine with notions of scale to socially organize ignorance of downstream uses.

While it is reasonable that everyone doing everything constitutes poor project management, it is remarkable that relationships themselves — acknowledging and making right the effects that one person has on another — is seen here as an act of labor that can be divided between people and handed off. It is important to recognize that this is neither a natural nor obviously normal state of affairs, as in other contexts the very notion of it would be utterly rejected (see Liboiron 2021). In this context, however, to divide labor is so naturalized that those at the very end of a rather long chain of causality are seen as distant. Thus even drawing out the connection takes work. "Whose work" then becomes the question.

One participant explained that no one tasked him with doing ethics work, so he doesn't do it: *"I don't have time allocated during my normal week to think about [...] responsible AI. This is not part of the work, at least not the part that someone would tell me from the top to worry about."* At a different company, a user experience researcher stated that ethics assessments are often filled out by software engineers, and that *"it was not my role"* to do it. This posed a problem to him, because there *"might be value in somebody who talks to customers i.e. me, filling it out versus, let's say, an engineer."*, in line with other work showing that it is difficult to retain strict "separation of concerns" between UX and AI-expert roles (Subramonyam et al. 2022) and other interdiciplinary communication challenges in AI development (Rakova et al. 2021; Hong et al. 2021).

Disconnections between developers and subject matter experts also created ethics concerns, such as when a student building cancer detection software did not have the situated knowledge to understand limitations in his dataset: *"like the amount of light that came in through the X-ray machine. Looking at those kinds of things I don't have the technical [knowledge] to know how the information is gathered. I don't know what the X-ray machine looks like."*, this presents a reverse situation to other work examining how AI experts make up for lack of AI expertise in the subject matter experts they collaborate with (Piorkowski et al. 2021). Dataset documentation improvement does not overcome his distanced location from the subject, where he was poorly positioned to interpret any documentation.

Status inflects this division of labor, relegating ethics work to mere details. One participant filled out a privacy questionnaire for his team to use an existing dataset to build a speech recognition benchmark, and felt the questionnaire asked for a lot of seemingly immaterial details his team was unconcerned with: *"It wasn't that easy to get through all the sections [of the ethics assessment] there were some questions about how the storage is secured [...] basically a team member of a research team or engineering team is not aware of, it's – it depends on IT support and configuration."* This worryingly suggests that relationships are so severed as to require dedicated IT support to accomplish their team's ethics work.

Conversely, another university-based participant emphasized that he was encouraged to focus on results:*"It's not like when we're presenting [our research at a conference] they ask you [...] what ethical steps did you take and things like that...Usually they just want to see your result."*

These divisions made the authority to decide questions of ethics ambiguous. One participant spoke of how he learned through an announcement from executives that his company's leadership canceled two of their team's large contracts over ethics concerns. While agreeing with this decision, he lamented that *"I wasn't actually involved. And so, I wouldn't say to tell them to make everybody involved. But make it possible to be involved."* When there is silence followed by an announcement it is unsurprising that people feel confusion and apprehension about making claims about ethics in ethics assessments. Another participant building body scanning technology explained: *"several of the questions [on the ethics assessment] are focused specifically on a machine learning AI statistical model, where many of the other questions are more around the broader product and business. So that was confusing,"* because making those assertions felt like an overstep of his own authority. Open source contexts confuse matters further, as open source licensing invokes ideological frames that reject the idea that developers should exercise any control at all over harmful use (Widder et al. 2022):*"the whole point is you can't control that - can't control what people do."*

Pressure to "scale" to ever more data, users, and customers deepens the sense that others in the chain are unknowable. For example, one participant assigned by his tech company employer to do UX work on an open source machine learning framework stated: *"So right now, I know the clients. And we don't have clients [who do harmful things]. But in the future, once we go public you won't be even able to control that [...] If you have, I don't know, 10,000 clients, I don't know how many clients we'll get when we go public [...] It can be difficult to track every single organization and [...] what they do with system."* His careful knowledge and consideration of his clients from his UX work, the metaphorical glue between "modules" of the supply chain, is the very thing he sees as being dismantled in his (and his company's) ideal future of broad adoption of the framework, in line with findings by Madaio et al. (2022) that AI practitioners encounter difficulty in engaging with downstream marginalized groups in large scale deployments.

This was echoed by the VR service tech lead, while most of his current customers have been *"physically met by one of our team at this point, that doesn't scale"* as they build a service company. Another participant discussed a deep collaboration with a customer to build an AI system on the customer's site, but was unable to know what the customer later did with that system after the initial prototype phase, as follow up work does not scale.

While these participants are using scale to refer a desirable state that creates a regrettable limitation on attention, others framed scale as something that legitimated not doing ethics work at all: *"I think our company is so focused on growing and scaling with users that ethical AI is not really– it's not really a big concern at this point."* Folks who saw things in these terms were concerned that any activities to do with ethics would create friction and lead to lost customers: *"If you bring [ethical AI discussions] for every other use case and every other customer, there is already a lot of customers that we are losing [...] I don't want this to create a bottleneck for our customers,"* and even a limitation on technological progress itself: *"there is going to be hundreds of thousands of industrial uses of AI [...] It's going to make industries advance [...] with the technology. But if we start limiting ourselves from doing so because of ethical concern then it stops progress of so many developments. So, we have to be really prudent on what is actually concerning and what is not."*

Abstraction, in software terms, is a layer that one can add in order to not have to deal with the specifics of individual code modules. As Gray (2021) notes, abstraction also does more cultural work, enabling ideals of scale within technology production that lead to failures to deliver socially beneficial technology. Abstraction at scale, Gray (2021) argues, is the very thing that makes it possible to not see harm. The examples above corroborate this, but it is important to acknowledge that this technology is being developed in a neoliberal economic context where *not* having relations or obligations is a dominant model of appropriate economic behavior  (Grant 1991) 1998). Unlike gift economies or other economic forms that constitute staples of economic anthropology (Plattner 1989), the dominant narrative of economic exchange here is that there are no social ties after the exchange takes place. The parties are *quits*, with no further obligations to one another. Our participants' invocation of "scale" as a way of describing the removal of personal relations, as if it were impossible to know the motivations and desires of one's customers beyond individual personal connection, makes clear they are tapping into this dominant narrative, which uncoincidentally is utterly compatible with the modularized world freed of obligations beyond one's immediate task.

## Connecting Links in the Supply Chain

The previous section showed the many ways that responsibility for harm appeared as outside of any given developer's control. Nevertheless, at every point in the supply chain, ties are not nearly as severed as this dominant discourse suggests. Participants were also located in a series of further cultural logics that work to produce connections that cannot be captured in the imagined triangle of Figure 1. From crudely co-opting the language of responsibility for competitive advantage, to aligning incentives towards customer-centered thinking that considers harms, to pushing for responsible AI through tactics of *soft resistance* (Wong 2021; Nafus and Sherman 2014), engineers step out of their engineering context into a different one as business people, news consumers, workers, and citizens with kinship ties.

"Customer-centric" was an explicit corporate value in many of our participants' workplaces that did require them to understand how customers interact with their software in order to increase product satisfaction and success. User experience design plays a key role here. One participant led his team in a user experience brainstorming session for their product to allow users to scan and monitor their body composition over time: *"We would do structured exercises to really brainstorm and hash out how might someone have a negative experience [...] I think we all share kind of that passion for the user, for the customer."* To this end, they made design modifications in response to feedback from pilot studies with users, framing this as putting the customer's needs first: *"We recognize [health and body composition as] a very sensitive thing [...] The core team has always been very focused on solving problems for the customer."* While paying customers are often the privileged "humans" in "human"-centered design to the exclusion of other affected parties (Pasanen 2019), specific anticipated users and use cases creates a connection point between commercial incentives and better or worse societal impacts, perhaps forming a safe starting point for imagining wider others touched by downstream uses.

User-centered design connects designer and engineer to user, but mechanisms like licensing apply further upstream to connect customer and supplier. One participant's company released its machine learning framework both as freely available open source and as a download available only after signing up with an email address. Of these very different relationships, the participant preferred the second method because *"we can be far more in touch with our customers. We know who they are, we can email them, we can make that more of a community."* Being "in touch" clearly has economic value, but also holds potential to surface awareness of things that can go wrong downstream.

Marketing is another exchange point between actors. "Ethical" or "responsible AI" was seen as a marketing advantage, with one participant suggesting that an *"ethical approach to AI is a very, very good influencing tool and it is [...] channel of growth for [our product] if [our company] is presenting this, trying to push for ethical AI, responsibly-created AI. Then users might choose [our company] over the competition."* Another believed that responsible AI can be used to win sales: *"the first thing that comes to mind is how to sell it, how to earn as much as possible, right? It's [...] how to get the first sale, how to have a right foot in the door. [...] this Ethical and Responsible AI, I think that we are living right now in the world that using these terms could only help you, right? [...] To build trust, right, with our customers."* Whether fortuitous alignment or crass co-opting, participants believed responsible AI efforts serve as a market differentiator, where companies can win business by helping their customers avoid ethical debt and the reputational costs it potentiates.

Similarly, engineers stepped out of the modules they build when thinking about how companies' ethical mishaps affect their own and their company's public reputation and profit. One participant relayed that his company had canceled a contract with a customer company which was using his team's software framework in a widely-reported unethical way, and suggested why this happened: a *"public perception of your moral compass and what you are involved in has a direct impact on your bottom line and that's what makes company owners stand up and do something different,"* arguing that the impact on profit required his company to consider and control downstream uses. Participants more directly associated with potentially harmful projects also feared personal reputational costs: *"Some things can have* *uses that you don't intend, and that you don't want [...] to come back to you."* Attention to reputational concerns seems the most direct acknowledgement of the impossibility of fully disconnected, modularized work. Developers know the impact of their creations will follow themselves or their companies when others believe it was their job to control the problem, even when they don't believe they had this control as developers. The ability to control was sometimes set in the context of a company's resources: *"I always imagine big companies with lots of resources. They do research and they create something. And that sometimes has an unintended effect on something. Then they must be held responsible and they should be aware [...] of what they are doing."* In this way, while researchers frame research as fundamentally abstract and maximally disconnected from eventual application, others see a company's research capacity as an indication it has enough resources to anticipate and remediate unintended (mis)uses.

Reputation and customer value are not new frameworks for legitimating ethics work (Metcalf et al. 2019). Some participants believed a focus on reputation obscures conflicted interests. For example, one participant says he hears the term "ethical AI" from *"C-suite kinds of people,"* but questioned *"are you using that as your buzzword to say that you're doing it? [...] Sometimes it's not clear that there's something actually happening."* While he believed his company doesn't want to *"be a party to any inhumane usages of AI technologies"* by downstream customers, he said they also want to *"make money. And sometimes those are cross purposes."* Similarly, another participant framed Google's treatment of Timnit Gebru as evidence of the limited scope of ethical action within the private sector: *"Google, I think, fired their VP of ethics or something like that. It was a really big deal. So they fired her, and that [...] communicates that they care about ethics to a certain point. But once it impacts their business, or once they deem it not necessary anymore, they can just forget that point."*

There were also instances where corporate rationales were not what motivated ethical action. The participant working on the body scanning project, for instance, emphasized that his team's positive group dynamics, and not anything the company did, was what made it possible to talk about ethics concerns. It got to the point where they were able to study each other as pilot users, having their own bodies scanned, and sharing their intensely personal reactions. For this participant, ethics discussions were an exercise in vulnerability, and responsible design meant a powerful obligation of duty to one's colleagues and friends in the position of "user" in lieu of hollow onstage rhetoric (i.e. Goffman 1959) of "passion for the customer" or "human-centered design", albeit with possible limits on these role-play pilots to create true empathy in system designers for the needs of the "other" Bennett and Rosner (2019).

Furthermore, ethics issues are not so easily disavowed when asked about work by friends and family: *"It sometimes gets hard when other people ask me. [...] 'What do you do?' And at the time, I'd be like, 'Oh, I kind of - I work in the AI workspace?' 'Oh, so you're getting people killed and assassinated through - with drones [...]' and it's like well, how much am I involved in that? [...] You can't say it's not true because it is true. [AI] is used - it has been used for that."* To this participant, work on a "general purpose" framework did not allow him to unsee harms when called to account in social contexts. Others talked about wanting more from their employer. One person noted that they could not necessarily say whether their framework was being used by the US Army, and this not knowing was itself a kind of harm they experienced as workers: *"It was Google or some other company that had engineers that even didn't know where their work is used. And so that's one thing I would really like to be informed, when my software is used. Where? For what purpose?"*

We also heard of developers exerting a soft form of agency and resistance when their moral compass made them uncomfortable with assigned work (Wong 2021). While her team commonly accepted jobs from companies seeking to use computer vision to automate quality control, one potential client asked them to track the actions of garment workers. Reflecting on the time she spent inspecting the training data the client provided, she stated, *"It was a little sad looking at videos. They work from 6:00AM in the morning to 9:00PM at night."* She said that even though the client called the project "object tracking," she was concerned that it would amount to algorithmic management (i.e. Lee et al. 2015): *"the algorithm that we're using is basically looking at people's motions to figure out what exactly they are doing. So, sometimes [...] they're just taking a break. You're just telling the system that this person's not doing anything."* She described how her team deprioritized the project until the client pulled away: *"[it was] not a project that any of us really wanted to work on. Thankfully it didn't go anywhere."* This is softly subversive: subversive in that it ended a project that her employer was asking them to do, but soft in that it resulted from deliberate yet individual *inaction* rather than collective active action.

This exposes limits on communication through the supply chain: some elite and well-resourced AI developers nonetheless still feel they must resort to weapons of the weak (i.e., Scott 1985). It is possible that she was the only one to actually inspect the data, and see the personal data showing rest during 15 hour shifts, making harms more intimately appreciable to her than to others due to division of labor. Through divided labor, less overt action might feel like the only option. Even in the more positive examples of making oneself accountable to the coworkers or social network one cares about, there is still a quality of off-stage norm-making that is not encapsulated in official rhetoric and responsible AI transparency interventions.

## Ethics Encapsulated

In practice, these crosscutting impulses to divide and connect lead to particular ways of handling responsibility and particular areas of priority. What does get attended to are matters shown to be areas of widespread public concern that can be encapuslated into a module of work, and thus do not introduce friction into the development process. For matters that do not fit that intersection, the hierarchies of prestige and value create notions of "real work" versus what is a checklist item to be assigned to lower status or outsourced staff, echoing past work on status in AI labor between that on the model versus data (Sambasivan et al. 2021), and in programming generally (Coleman 2012).

Narrow conceptions of bias and privacy issues fit this intersection. They have entered developers' milieux (see also Fjeld et al. 2020), through high-profile ethical lapses including racial and gender disparities in computer vision (Buolamwini and Gebru 2018) and regulatory action such as the European General Data Protection Regulation. Both provide a shared social location, from outside the supply chain, which developers can and do draw on in work discussions. Nearly half of our participants brought up privacy laws and similar structures in their industry or company, which serve to codify language and processes, as well as providing a concrete way hold companies externally accountable to these processes, as Yeung et al. (2019) argue is necessary to end "ethics washing".

Those touchstones provide a referent for recognizing some harms, but not others. For example, one participant doing AI research for the military was concerned about the mathematically-identifiable biases within the weaponry, saying, *"I think the whole issue of bias and its societal and ethical implications is terribly interesting and we don't have as much conversation, particularly with cyber weapons, as we should."* He was simultaneously less concerned with any bias at work in choices external to his "module" that his customers make about who to point weapons at.

Where ethics issues are not encapsulatable, work on them was frequently left undone or cast as low status work, offloaded to contractors or, in one example to a woman who felt pigeon-holed into doing "*release steps*," that is, administrative labor no one else wanted to do. Assigning this task to contractors was common, because, "*every task can be a trap*," meaning it can take a surprising amount of time. Therefore, as one person put it, "*We have a [contractor] who did the first model card for the [ethics assessment].*" Similarly, a contractor on a different team was frustrated that he was asked to do undone ethics tasks for his temporarily assigned team, an exercise in paying other's ethical debt (Fiesler and Garrett 2020).

Participants often justified their disinterest in doing "release steps" through customer orientation: *"[I'm] excited when I'm working on things that kind of have a clear pathway to a customer-facing feature,"* making clear that they did not see model cards as a customer-facing feature, running counter to their intent to drive transparency between developers.

Others have studied the effectiveness of AI fairness checklists to empower individual advocates and formalize the ad-hoc (Madaio et al. 2020). When tools like model cards do not face customers but are situated as outsourceable tasks required to secure internal approval for code release, there is no meaningful location from which to look downstream at the next or final context of use. This social configuration leaves us with an odd bimodality of responsible AI. On the one hand, prominent dramas about social harms embroil the careers of executives in congressional hearings, while on the other, contractors are asked to do "the paperwork." In the hollow middle, the stars must tightly align between outside the module and in, for action to take place. They sometimes do, but rarely.

## Where to go from here?

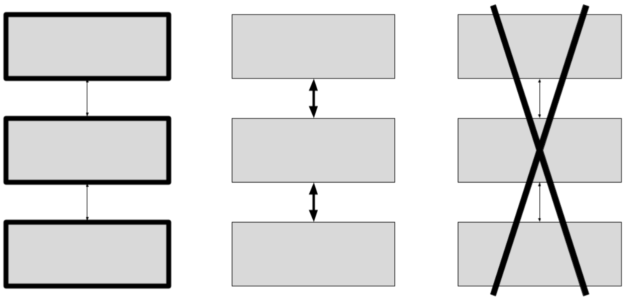

**Figure 2.** Three possible futures: a) Acting within the modules; b) Strengthening the interfaces; and c) Rejecting modularity.

Many efforts at supporting responsible AI make assumptions about the capacity for panopticality that this work demonstrates do not hold. AI fairness checklists (Madaio et al. 2020; Holstein et al. 2019) or the Privacy, Ethical and Social Impact Assessment (Memic 2018) operationalize ethics principles by providing an inventory of common pathways where matters go awry, as if everyone is equally situated to follow those paths in the same way as an auditor seeks a birds' eye view into a supply chain. Other ethics toolkits, such as Vallor et al. (2018), acknowledge the interstitial nature of ethics failures by asking "what combination or cascade of causes led to the ethical failure?" When teams neither have visibility nor control over those cascades, and often do not believe they *should* have this control, the success of these inventory-like approaches is likely to be limited.

Model cards (Mitchell et al. 2019) and datasheets for datasets (Gebru et al. 2021) hold promise for communication between people across modules and organizations. It is possible to treat them as a kind of "nutrition label" (Chmielinski et al. 2022), which Gansky and McDonald (2022) call "metadata maximalism," where facts are announced to an unspecified audience. While our own interviewees struggled to identify who had authority to claim these once-and-for-all facts, our study also suggests model cards might hold appeal because after the declaration is made, the matter is settled, and modularity is safely preserved. There is a different way of thinking about model cards, which is to take the partiality of knowledge as a given, and use it as a boundary object. For example, a model card might declare one set of biases, but in a given deployment, another type of bias might be a more relevant concern, and make its way onto the model card after the fact. In this second approach, a model card is not a one-and done affair, but a place where partiality comes together and relations occur, making the relationship less like a supply chain and more like a value chain where collaborators are working together to collectively address the problem without resorting to views from nowhere.

Suchman's work anticipated the dangers of "metadata maximalism" and argues that partial translations provide the route to a more robust form of accountability. So, what does accepting, and indeed leveraging, that partiality look like in responsible AI? If modularity is dominant but not uncontested, then the answer depends on one's relationship to modularity itself. Here we suggest three approaches are possible: 1) accepting the ethic of modularity and seeking increased developer agency within it, 2) strengthening the interstitial connections that already do exist to enable engineers to step outside of modular thinking, and 3) fundamentally reject the ethic of modularity in its entirety, and in its place, build a new set of expectations about what software development entails. These are summarized in Table1.

### *Acting Within the Modules*

If we fully accept that the dominance of modularity is unlikely to change soon, we would seek to act within it. If a model card were to accompany the technical artifact through multiple handoffs, perhaps there is an opportunity for participants to *append* their partial understanding of the flaws, limitations, divergent provenances, and contexts of use of this documentation. This might require, ironically, doubling down on the division of labor, by clearly delineating what is rightly in a developer's "module," and what is a matter for those in user experience, sales, and legal roles. This would relieve the developer of the need to consider anything beyond the tasks as pre-defined for them. Defining who does which parts of the model card would then have to be done by someone with the authority to step outside a given module, likely someone with a managerial or leadership role. As models change hands through different companies, the appending process might start again, in effect creating a trail of breadcrumbs between peers at different companies. The trail might be shared in contextually appropriate ways, with appropriate technological control over what entity sees what level of detail, for example with federated evaluation and learning (i.e. Wang et al. 2019). Alternatively, Javadi et al. (2021) propose ways (*e.g.* automated analysis of logs) to audit AI services for downstream misuse given that their "ease of integration and use, limited provider involvement during engagement and generalisability" may make misuse more likely.

These approaches preserve the orientation towards scale, and leave out matters that flow through channels outside the supply chain, from journalism, academia, and non-profit organizations. Unless there is a creative way to modularize participation from impacted groups, which itself could be offensive to some of those groups, this approach re-inscribes their exclusion. In that sense, it is likely to systematize that which is already sayable in an onstage, public way, and therefore not fundamentally change the relations of ignorance when more problematic situations do arise. It is then left to regulators, journalists, and academics to force conversation and action about that which is considered unsayable from within the chain.

### *Strengthening the Interfaces*

Another approach would strengthen connections between companies beyond those allowed by the "developer hat." This approach would embrace some of the previous techniques, but then buttress the communication that happens in the process of exchange and outside of development. Model cards might be reinforced by contractual obligations and meaningful customer knowledge and communication, which would involve increased contribution from non-developers. Those playing customer roles in the supply chain might routinize asking suppliers for model cards, if the data it was trained on was properly consented, if crowd workers labeling the data were paid an appropriate wage, etc., which is commonplace in supply chains for physical goods. Even the simple act of asking the question helps the person in the supplier role acknowledge their part in the accumulation of ethical debt (Fiesler and Garrett 2020), and reframe ethics work as an act of delivering customer value. While Gansky and McDonnald (Gansky and McDonald 2022) argue this is unrealistic for data brokerage markets, this does seem more feasible in, say, computer vision supply chains. Still, not every company is equally positioned to share knowledge. One study showed that AI entrepreneurs effectively have to conceal the ethics work they *were* doing from their venture capital funders, who sought to hide admissions of limitations (Winecoff and Watkins 2022).

This study of early-stage startups corroborates our findings that offstage talk matters in guiding what actions people end up taking. Creating fluidity between onstage and offstage communication could have the effect of allowing people to better integrate the multiple social locations they inhabit. Companies could make sure that developers on staff hear about relevant incidents beyond the current touchstone cases, such as those compiled in PAI's Incident Database, to broaden the scope of what is sayable onstage. Developers might use their value as skilled laborers to articulate their preference for working for companies that do meaningfully engage with communities who could be harmed, and make clear they are not prepared to pay personal reputational costs, either onstage or off. In turn, journalists and academics could also place more emphasis on the multi-actor cascades, as journalist Karen Hao has called for (Isaac and Hao 2022).

Thinking interstitially means moving away from the binary question of "do I have full control, or not?" and reasoning in probabilities and frictions. What does the technology make easier or harder, faster or slower, and what do contractual obligations or marketing messages make easier or harder? For example, one participant relayed how his team had exerted some control to make downstream misuse harder, by removing parts of the tooling he provides that would make it too easy to take face detection models (*i.e.*, is this a face?) and turn them into a facial recognition system (*i.e.*, whose face is this?). The participant acknowledged that anyone could still train models themselves to do facial recognition, but interventions like this made it harder to do so. Similarly, the Ethical Source movement uses licenses to introduce legal friction for harmful uses in software supply chains, acknowledging this control is not total[†].

The advantage to this approach is that by enhancing actors' capacities to work across social locations, stronger norms can form. This bears some relationship to values-sensitive design (Friedman 1996), except that a located accountability approach recognizes that values alignment is full of friction, and inflected by numerous positionalities. There might not *be* perfect values alignment, or even the possibility of explicitly articulating values at all (Le Dantec

[†] See: `https://ethicalsource.dev`, `https://firstdonoharm.dev/learn/`

et al. 2009). Instead, there might be, as in the startup study, “working misunderstandings” (Ferguson 1994) where parties misrecognize the actions of one another by necessity. Ignorance can never go away, only be rearranged. Still, inclusion here is not assured and, depending on policy conditions, this approach could risk setting up a path dependence where ethics issues can be better acknowledged and acted upon, but remain a second order, lagging concern.

### *Rejecting Modularity*

What if modularity were eschewed entirely, both in terms of code and the broad social arrangements it mediates? Instead of building discrete modules packaged for broad use at any scale, actors who object to the modularity ethos in the first place might radically collapse the AI supply chain, and prioritize building good relations with people to build technologies with, not for. Echoing criticism of endless AI scale (Bender et al. 2021), Gebru and Hanna propose a new model of AI development, where the goal is not to produce “AI for the value of AI itself”, but to instead be “sensitive to other forms of knowledge” that developers do not have in order to examine and curate datasets for particular end uses, even if this is slower or more expensive (Strickland 2022).

This approach might seem foreign to the software engineers in our study, especially those building general purpose frameworks or scaleable software-as-a-service architectures. Look just outside those norms, however, and there are plenty of examples to be found. Reflecting on her work with North Carolina community healthcare workers building vaccine equity for Black and Latinx communities, Gray (2021) employed design justice principles from Costanza-Chock (2020) to argue that “we must prioritize a deep, methodical connection with subject matter and domain expertise in lieu of an unexamined rush to scale or to shield ourselves from the realities of a social world.” This took the form of 19 months of weekly meetings and six months of biweekly meetings listening to community health care workersto create a software-based patient intake form. Gray recognizes that this intensive process rather than using an off-the-shelf form introduced “friction, or working against scale, [which] is considered a bad thing in [Computer Science]. It is considered inefficient, a waste of engineering time”, but reflects Arendt’s 1963 insight that the “greatest violence comes not from individual malicious actors, but rather it is the product of remoteness from reality”; a remoteness that modularity and scale create. Gray’s approach treats social relations as first order work that cannot be bracketed off as a mere input or requirements capture. Here the relationship is the objective, not the lines of code that may or may not result.

This approach begins to dismantle meaningful distinctions between producer and user. Software experts will have the opportunity to see and avert possible downstream harms they find in the user’s intended context, and those most acquainted with its context of use may see possible harms from how it is built, such as questioning the data, and its applicability to the context of the use they desire. Both improve the underlying value of the product for that context. While the previous approach strengthened norms in a broad but inconsistent way, this does it in a more focused but deep way. Such focus has a long history outside an AI context (see Costanza-Chock 2020, for an overview), where matters go beyond simple harm mitigation or user-centered design and attempt to rectify epistemic injustice itself by reorganizing who gets to make a claim about what is and is not worth knowing and building (Ottinger 2022).

Rejecting modularity in a modularized world raises interesting questions for upstream tools. If, as the saying goes, the master’s tools will never dismantle the master’s house (Lorde 2003), how should teams make choices about upstream tools, like assembly language or compilers? How would they relate to companies known for ethics breaches that also supply otherwise useful libraries and other code? It might turn out that “generic” tools are not in fact generic at all, but generic only to those who are currently included in and well served by the current supply chain. This approach also raises questions for public policy. Given the resource inequalities between community groups and companies that seek to scale, and that those same groups are meeting social needs that arguably benefit a country as a whole, what would an appropriate science and technology policy do to support these efforts?

## Conclusion

Thinking about ethics and responsibility as chains of relations surfaces specific locations in which ethical decision-making can take place. Those locations might be upstream or down, and they might be within the cultural logic of modularity or outside it. The combinations of these locations shape what is considered sayable and what is off-stage talk. They shape what is prestige-garnering work, what is paperwork, and what is high stakes public drama. Ironically, the recognition of the interdependence of modern software development shows that what participants experience as a series of handoffs through a supply chain in many respects resembles a value chain, where actions have a combinatory effect. Ethical debts accrue, and harms occur or are avoided. We have also shown that a realistic notion of responsible AI work takes into account AI developers’ beliefs about responsibility and agency as constrained by modularity, and makes deliberate choices about how strong a role current software production ideals should play in future responsible AI development.

**Table 1.** A table summarizing our three possible ways forward.

| | Working Within Modularity | Strengthening the Interfaces | Rejecting Modularity |
|---|---|---|---|
| What it means | Work within existing dominant culture of modular software | Amplify intersections where engineering meets business and societal touchpoints | Supply chain is collapsed as much as possible and grounded in specific context of use, and relationships are first-order matters |
| Hypothetical Interventions | • Model cards/data sheets become interoperable & append-able, within & between organizations<br>• Extend predefined division of Responsible AI labor to better include Sales/UX/marketing<br>• Information flows from upstream to down | • Appendable model cards/data sheets within and between organizations<br>• Exchange points supported by contractual obligations, norm-setting in marketing, understanding of customer needs<br>• Information flows both ways, customers ask tougher questions of model cards<br>• Normalize onstage discussions of incidents & harms | • All those who software will touch (not just “users”, “customers”) are contributors to creating software, who are recognized and paid for labor<br>• Advocating for policies and organizations that support community-driven technology development<br>• Building AI might or might not be the outcome |
| Real-world example(s) | • Checklist style approaches to responsible AI (Madaio et al. 2020)<br>• Methods to monitor for downstream misuse of AI services (Javadi et al. 2021) | • Ethical Source movement’s use of licensing to create friction against harmful use (i.e. Widder et al. 2022) | • Gray and team’s healthcare software using Design Justice principles (Gray 2021; Costanza-Chock 2020)<br>• DAIR as envisioned (Strickland 2022) |
| Advantages | • Easily implemented, because does not challenge dominant beliefs about scale/division of labor | • Implementation eased by building on existing practices beyond modularity<br>• Provides stage for broader norm setting, while recognizing that points of misalignment are likely<br>• Allows more but not necessarily all kinds of discourse to be onstage<br>• Facilitates direct interaction between different disciplines within and between organizations | • Harms less likely to occur as process designed to prevent disavowal of responsibility<br>• Eventual use of the technology more likely for the context in which it was built |
| Disadvantages | • Offstage discussions have no path to move onstage<br>• Relies on policymaking for any serious hard-to-modularize harms | • Introduces friction but does not fully prevent downstream harmful use<br>• No built-in process to ensure voices of those harmed are heard<br>• Efficacy for harm prevention could still rely on policy context | • Uncertainty, lack of standardized process<br>• Software creation process is slower<br>• Some, such as venture capitalists, will critique lack of scalability<br>• Uncertain reliance on standardized upstream dependencies and tools |

## References

Hannah Arendt. 1963. *Eichmann in Jerusalem: A report on the banality of evil*. Viking Press.

Carliss Young Baldwin, Kim B Clark, Kim B Clark, et al. 2000. *Design rules: The power of modularity*. Vol. 1. MIT press.

Victor R Basili and Barry T Perricone. 1984. Software errors and complexity: an empirical investigation0. *Commun. ACM* 27, 1 (1984), 42–52.

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*. 610–623.

Cynthia L. Bennett and Daniela K. Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the "Other". In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. `https://doi.org/10.1145/3290605.3300528`

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

Michel Callon. 1998. Introduction: the embeddedness of economic markets in economics. *The sociological review* 46, 1_suppl (1998), 1–57.

Michael Carolan. 2020. Acting like an algorithm: digital farming platforms and the trajectories they (need not) lock-in. *Agriculture and Human Values* 37, 4 (Dec. 2020), 1041–1053. `https://doi.org/10.1007/s10460-020-10032-w`

Karin Knorr Cetina. 1999. *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954* (2022).

E Gabriella Coleman. 2012. *Coding freedom: The ethics and aesthetics of hacking*. Princeton University Press.

Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Jenny L Davis. 2020. *How artifacts afford: The power and politics of everyday things*. MIT Press.

Edsger W Dijkstra. 1982. On the role of scientific thought. In *Selected writings on computing: a personal perspective*. Springer, 60–66.

Andrew Feller, Dan Shunk, and Tom Callarman. 2006. Value chains versus supply chains. *BP trends* 1 (2006), 1–7.

James Ferguson. 1994. *The anti-politics machine:" development," depoliticization, and bureaucratic power in Lesotho*. U of Minnesota Press.

Casey Fiesler and Natalie Garrett. 2020. Ethical Tech Starts with Addressing Ethical Debt. (2020).

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT undermines its organizing principles. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1982–1992.

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 (2021), 86–92.

Erving Goffman. 1959. *The presentation of self in everyday life*. Anchor.

Colin Grant. 1991. Friedman fallacies. *Journal of Business Ethics* 10, 12 (1991), 907–914.

Mary Gray. 2021. The Banality of Scale: A Theory on the Limits of Modeling Bias and Fairness Frameworks for Social Justice. (2021). Conference on Neural Information Processing Systems (NeurIPS).

Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Donna J Haraway. 1991. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Simians, cyborgs, and women: The reinvention of nature* (1991), 183–201.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for Natural Language Failures with the AI Playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–11. `https://doi.org/10.1145/3411764.3445735`

Don Ihde. 1990. Technology and the lifeworld: From garden to earth. (1990).

William Isaac and Karen Hao. 2022. Keynote Interview: Karen Hao in Conversation with William Isaac. `https://www.youtube.com/watch?v=9u-62Ijtb1I`.

Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. 2021. Monitoring AI Services for Misuse. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 597–607.

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

Christian Kästner, Sven Apel, and Klaus Ostermann. 2011. The road to feature modularity?. In *Proceedings of the 15th International Software Product Line Conference, Volume 2*. 1–8.

Chris F Kemerer. 1995. Software complexity and software maintenance: A survey of empirical research. *Annals of Software Engineering* 1, 1 (1995), 1–22.

Miryung Kim, Thomas Zimmermann, and Nachiappan Nagappan. 2014. An empirical study of refactoringchallenges and benefits at microsoft. *IEEE Transactions on Software Engineering* 40, 7 (2014), 633–649.

Timothy D Korson and Vijay K Vaishnavi. 1986. Modularity on Program Modifiability. In *Empirical Studies of Programmers: Papers Presented at the First Workshop on Empirical Studies of Programmers, June 5-6, 1986, Washington, DC*, Vol. 1. Intellect L & DEFAE, 168.

Nick Lally. 2021. “It makes almost no difference which algorithm you use”: on the modularity of predictive policing. *Urban Geography* (2021), 1–19.

Bruno Latour. 1993. *We have never been modern*. Harvard university press.

Bruno Latour. 1994. On technical mediation. (1994).

Bruno Latour et al. 1999. *Pandora’s hope: essays on the reality of science studies*. Harvard university press.

Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1141–1150.

Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1603–1612.

Max Liboiron. 2021. Pollution is colonialism. In *Pollution Is Colonialism*. Duke University Press.

Audre Lorde. 2003. The master’s tools will never dismantle the master’s house. *Feminist postcolonial theory: A reader* 25 (2003), 27.

Alan MacCormack, John Rusnak, and Carliss Y. Baldwin. 2006. Exploring the Structure of Complex Software Designs: An Empirical Study of Open Source and Proprietary Code. *Management Science* 52, 7 (July 2006), 1015–1030. https://doi.org/10.1287/mnsc.1060.0552

Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners’ Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

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*. 1–14.

James W Malazita and Korryn Resetar. 2019. Infrastructures of abstraction: how computer science education produces anti-political subjects. *Digital Creativity* 30, 4 (2019), 300–312.

Antonio Martini, Erik Sikander, and Niel Medlani. 2016. Estimating and quantifying the benefits of refactoring to improve a component modularity: a case study. In *2016 42th Euromicro conference on software engineering and advanced applications (SEAA)*. IEEE, 92–99.

Tara McPherson. 2018. *Feminist in a Software Lab: Difference+ Design*. Vol. 6. Harvard University Press.

Inda Memic. 2018. What’s the PESIA Framework? https://blogit.itu.dk/virteuproject/2018/10/30/whats-the-pesia-framework/

Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

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.

Dawn Nafus and Jamie Sherman. 2014. Big data, big questions—this one does not go up to 11: the quantified self movement as an alternative big data practice. *International journal of communication* 8 (2014), 11.

Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735.

Gwen Ottinger. 2022. Responsible epistemic innovation: How combatting epistemic injustice advances responsible innovation (and vice versa). *Journal of Responsible Innovation* (2022), 1–19.

David L Parnas. 1972. On the criteria to be used in decomposing systems into modules. In *Pioneers and Their Contributions to Software Engineering*. Springer, 479–498.

Jussi Pasanen. 2019. Human-Centred Design Considered Harmful. https://www.jussipasanen.com/human-centred-design-considered-harmful/. (2019). Accessed: 2022-05-2.

David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

Stuart Plattner. 1989. *Economic anthropology*. Stanford University Press.

Robert N Proctor and Londa Schiebinger. 2008. Agnotology: The making and unmaking of ignorance. (2008).

Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–23. https://doi.org/10.1145/3449081

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

James C Scott. 1985. *Weapons of the weak: Everyday forms of peasant resistance*. Yale University Press.

Mary Shaw. 2011. Modularity for the modern world: summary of invited keynote. In *Proceedings of the tenth international conference on Aspect-oriented software development*. 1–6.

Vincent Yun Shen, Tze-jie Yu, Stephen M. Thebaut, and Lorri R. Paulsen. 1985. Identifying error-prone software—an empirical study. *IEEE Transactions on Software Engineering* 4 (1985), 317–324.

Sean Sirur, Jason RC Nurse, and Helena Webb. 2018. Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR). In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. 88–95.

Marilyn Strathern. 2002. An anthropological comment. *Virtual society?: Technology, cyberbole, reality* (2002), 302.

Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.

Eliza Strickland. 2022. Timnit Gebru Is Building a Slow AI Movement. https://spectrum.ieee.org/timnit-gebru-dair-ai-ethics. *IEEE Spectrum* (31 March 2022). Accessed: 2022-07-7.

Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

Lucy Suchman. 2002. Located accountabilities in technology production. *Scandinavian journal of information systems* 14, 2 (2002), 7.

Peri Tarr, Harold Ossher, William Harrison, and Stanley M Sutton. 1999. N degrees of separation: Multi-dimensional separation of concerns. In *Proceedings of the 1999 International Conference on Software Engineering (IEEE Cat. No. 99CB37002)*. IEEE, 107–119.

Suzanne L Thomas. 2019. Migration versus management: the global distribution of computer vision engineering work. In *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)*. IEEE, 12–17.

Douglas A Troy and Stuart H Zweben. 1981. Measuring the quality of structured designs. *Journal of Systems and Software* 2, 2 (1981), 113–120.

Shannon Vallor, Brian Green, and Irina Raicu. 2018. Ethics in technology practice. *The Markkula Center for Applied Ethics at Santa Clara University. https://www. scu. edu/ethics* (2018).

Iris Vessey and Ron Weber. 1983. Some factors affecting program repair maintenance: an empirical study. *Commun. ACM* 26, 2 (1983), 128–134.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. 2019. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252* (2019).

David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *Proceedings of the conference on fairness, accountability, and transparency*.

Amy A Winecoff and Elizabeth A Watkins. 2022. Artificial Concepts of Artificial Intelligence: Institutional Compliance and Resistance in AI Startups. *arXiv preprint arXiv:2203.01157* (2022).

Langdon Winner. 1980. Do artifacts have politics? *Daedalus* (1980), 121–136.

Richmond Y Wong. 2021. Tactics of Soft Resistance in User Experience Professionals' Values Work. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

Karen Yeung, Andrew Howes, and Ganna Pogrebna. 2019. AI governance by human rights-centred design, deliberation and oversight: An end to ethics washing. *The Oxford Handbook of AI Ethics, Oxford University Press (2019)* (2019).

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. 2021. The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312* (2021).