

Strategies for Increasing Corporate Responsible AI Prioritization

Angelina Wang¹, Teresa Datta², John P. Dickerson^{2, 3}

¹Stanford University

²Arthur

³University of Maryland

angelina.wang@stanford.edu

Abstract

Responsible artificial intelligence (RAI) is increasingly recognized as a critical concern. However, the level of corporate RAI prioritization has not kept pace. In this work, we conduct 16 semi-structured interviews with practitioners to investigate what has historically motivated companies to increase the prioritization of RAI. What emerges is a complex story of conflicting and varied factors, but we bring structure to the narrative by highlighting the different strategies available to employ, and point to the actors with access to each. While there are no guaranteed steps for increasing RAI prioritization, we paint the current landscape of motivators so that practitioners can learn from each other, and put forth our own selection of promising directions forward.

Introduction

There is increasingly research on how to implement fair and responsible artificial intelligence and algorithmic bias techniques. However, relatively little of this research has translated into industry practice (Holstein et al. 2019), and recent research has often found that a key obstacle is institutional barriers which hamper adoption (Madaio et al. 2022, 2020). In an ideal world we might have both better-scoped and higher quantities of regulation in place, and while there are certainly movements in this direction (e.g., EU AI Act, GDPR), policymaking can be slow and imprecise. In the meantime it is important to understand the motivations that companies have historically had for incorporating responsible AI concerns so that we can learn from what has worked in the past in order to enact greater impact in the present.

A major obstacle facing Responsible AI (RAI)¹ in practice is the lack of corporate investment of resources. Individual RAI practitioners develop their own strategies within their respective companies (Ali et al. 2023), but there is not yet a consolidated analysis of these separate techniques

which could be shared with each other. The research question we seek to answer in this work is: *How can we motivate companies to increase the resources they spend on RAI?* We foreground the word *we*, and underline throughout the work who has the power to pull on each of these levers that we unveil. To discover the strategies, we conduct 16 semi-structured interviews with participants who a) currently or in the past worked at companies that used algorithms with RAI-related concerns, and b) had insight into or control over the decision-making process that set RAI priorities. We provide a summary of our findings in Fig. 1, ultimately discovering six high-level reasons that motivate companies to prioritize RAI: *external cues* like publicity, *regulatory pressures*, *organizational macro-motivators* like the funding type of a company, relevance to *company success*, unique circumstances due to *company culture and individuals*, and the *effort and ease* of implementation. It boils down to prioritization: companies are rarely against RAI, but when there are many competing interests, RAI is often neglected.

However, the story is not as simple as “increase regulation” or “publicize RAI failures,” strategies which prior work has covered as well (Ali et al. 2023). Rather, there is nuance to each of these strategies, and the *details* are the strength of our work. We provide a novel perspective by revealing the complexity behind each strategy as well as the actors most able to exercise them. For example, while public RAI failures are known to motivate prioritization, they can also scare companies from engaging in RAI at all, fearing backlash for an attempt that doesn’t meet a journalist’s standards. Additionally, while we know that companies are scared of not complying with legislation for lawsuits, our interviews provide the novel insight that companies are not just scared of legitimate lawsuits, but rather any publicity that could lead to a frivolous lawsuit because of what might be surfaced during the legal discovery phase. Or while we might have an idea of how RAI can serve as a company differentiator, the critical impact of generative AI on this differentiation has not been previously examined. Our interviews unveil that with the rise of generative AI, differentiating via RAI provides an arena for companies that cannot afford to compete on compute-heavy tasks.

Ultimately, our interviews detail a tangled web of ambiguity: there are no guaranteed steps forward for increasing RAI prioritization (not even regulation, which can be heavy-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Throughout this work, we use the term RAI to describe the broad category of themes related to fair machine learning, AI ethics, and algorithmic bias. While it can be reductive to collapse all different notions of bias and fairness under the same terminology (Blodgett et al. 2020), in our case we find this to be an appropriate level of abstraction, and will bring up specific instances where this abstraction hides relevant details.

handed, misdirected, and hard to get right). Yet, by painting the landscape of strategies and showing the challenges within each, we bring clarity and structure to encouraging ways forward. As an example, a practitioner who is having trouble getting buy-in for an RAI-related concern learn that they can leverage the argument that it will help their company differentiate from a competitor. Additionally, by naming actors with power to enact change, we offer a call to action to previously overlooked groups, such as venture capitalists and consultants. We conclude by discussing four concrete directions that emerge to us as the most promising. By looking to the past we scope ourselves to the efforts that have already been tried. However, we believe that through this consolidation of separately-performed techniques, we can illuminate the landscape of approaches for RAI prioritization and help to inspire novel methods going forward. As anonymized participant #4 (P4) puts it, “There’s a lot of really great ways that people in the field can meet each other to navigate some of the challenges that we all share.”

Related Work

AI Ethics in Industry

There are many difficulties to implementing RAI in practice. These are due to factors like organizational barriers (Deng et al. 2023; Rakova et al. 2021; Madaio et al. 2020), principles which are hard to operationalize for reasons such as being too vague (Holstein et al. 2019; Schiff et al. 2020, 2021; Mittelstadt 2019; Munn 2023; Green 2020), research outputs which do not match practitioner needs (Madaio et al. 2020, 2022), and the structure of labor and workplaces (Shilton 2012; Widder and Nafus 2023). Of these obstacles, our work focuses on the lack of resources invested by companies.

Many important works have explored the dynamics between RAI and corporations under capitalism. Ali et al. (2023) conduct a thorough analysis of RAI implementation in technology companies. They find three key barriers: the overriding authority of software product launches, the difficulty of quantifying ethics, and the frequent team reorganization that can cause loss of knowledge. Rather than focusing on the barriers RAI practitioners have, in our work we examine RAI successes to understand what can be amplified to increase RAI buy-in. Both of our works share the goal of suggesting steps forward to increase RAI investment, but due to the different framing, settle on different ways forward that are collectively important. Metcalf, Moss, and danah boyd (2019) capture how the inherent culture of Silicon Valley can be adversarial to substantive ethics due to differing incentive structures and sometimes-conflicting values, and Phan et al. (2022) describe the conceptualization of ethics knowledge as simply another form of capital that is exchanged. On the other hand, recent work describes a shift in perspective that a company’s responsibility encompasses social impact in addition to just profit—though potentially only larger companies have the resources to do this, compared to smaller ones (De-Arteaga, Feuerriegel, and Saar-Tsechansky 2022). Other works look specifically to smaller startups and the unique pressures they face in this domain (Wincoff and Watkins 2022; Hopkins and

Booth 2021; Vakkuri et al. 2020).

Ultimately while ways of increasing RAI prioritization have surfaced in prior work, our singular focus on this question brings out new details and nuances, and can help to inform *who* can do *what* in increasing RAI prioritization.

Organizational Theory

In our work, we focus on inductively discovered findings rather than deductive, theory-driven themes, but provide discussion of relevant broader theories here. Central is the economic system of capitalism allowing private actors to operate with a central motivation of maximizing profit (Gilpin 2000). In theory, this motive for profit is said to ensure the efficient allocations of resources (Smith 2002). However, this market fundamentalism can play out in many ways: scholarship has examined how organizational policies are often decoupled from daily practices (Meyer and Rowan 1977; Bromley and Powell 2012), and our work examines similar ongoing tensions like between valuing ethics in the abstract and prioritizing profits in practice.

Neo-institutional theory is one of the key frameworks for understanding organizational behavior and change (Powell and DiMaggio 2012), and posits that sources of organizational change can be endogenous or exogenous, i.e., internally or externally-triggered. Actors in an organization can endogenously construct change, and our work sheds lights on the specifics of this in the RAI space. Through the lens of organizational field theory, organizations can internally forge new paths through the interactions of the inhabitants of these institutions (Schneiberg 2007; Hallett and Ventresca 2006). Additionally, corporations can be reordered by specific highly-connected individuals who take strides to disrupt the old institutional structures (Weber, Roth, and Wittich 1978; Suchman 1995; DiMaggio 1988). Meanwhile, exogenous sources include environmental “jolts” (e.g., sudden societal events), force organizations to adapt in unique ways. These “transitional blips” allow for the strengthening of cultural priorities and introduction of unrelated changes (Meyer 1982). We discuss a variety of external triggers for corporations: from publicity to social movements to regulations. Similarly, social movement literature discusses how corporate protests and large-scale media coverage affecting corporate stock prices can be particularly effective at inciting change (King and Soule 2007; Davis and Thompson 1994). There are parallels to draw from domains like privacy and ESG (environmental, social impact, and corporate governance) issues to RAI’s place within capitalism. In all of these spaces, a value besides profit comes into tension (Hartzmark and Shue 2023; Cohen, Gurun, and Nguyen 2020).

Methods

To answer our research question: *Based on past successes, how can we motivate companies to prioritize RAI?*, we conducted semi-structured interviews with 16 participants from 13 companies.

Participant Recruiting

We sought participants who a) currently or in the past worked at companies that used algorithms with RAI-related

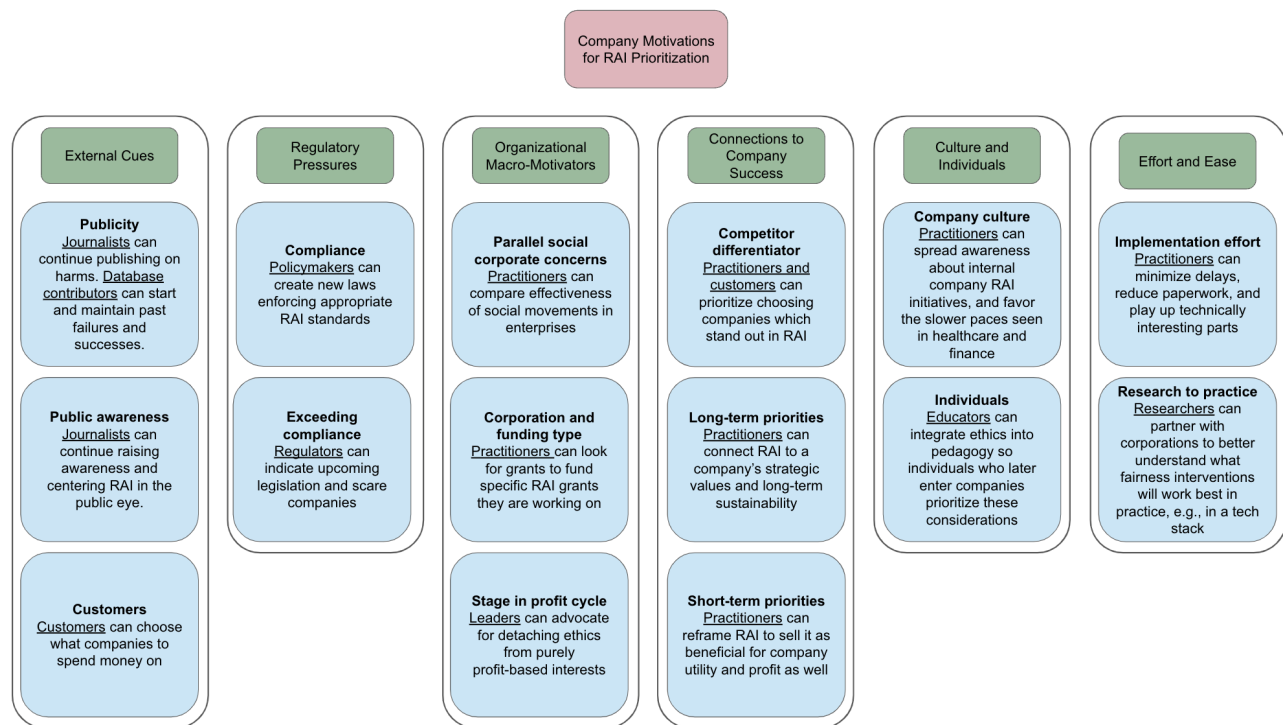


Figure 1: Six themes surfaced from our interviews as directions for increasing the motivation that companies have for prioritizing responsible AI. We summarize each of these six themes, and underline examples of actors that are able to act on each.

concerns and b) had insight into or control over the decision-making process that set RAI priorities. We recruited these participants by reaching out to our personal networks, posting on Twitter, LinkedIn, Slack groups, and snowball sampling by asking participants to refer us to others who might be useful to talk to. In total, we interviewed 16 participants from 13 companies between July - September 2023, reaching thematic saturation by our last two interviews (Green and Green 2004). Given the specialized background of our participants, the pool of eligible participants was relatively small. Each participant was compensated with a \$25 gift card. The industry and company size of each participant is listed in Tbl. 1. Their roles are: scientist/engineer (10), product (2), manager/executive (4). Our participants span a wide range of domains: Software as a Service (SaaS), healthcare, finance, manufacturing, and more.

Interview

The first author conducted all of the interviews, which were 30-60 minutes and conducted on an online video platform. Each interview began with explaining the purpose of our study and asking whether participants were comfortable with the level of anonymity we would provide. We then asked participants if we could record the video and automatically generate a transcript; if not, the interviewer took extensive notes. The interview instrument of guiding questions is included in the Appendix. Generally, they were customized for each interview and depended on the participant's role at the company. We sent all participants a draft before submis-

sion to ensure comfort with their level of anonymity given the sensitivity of the subject.

Analysis

We conducted an inductive, data-driven thematic analysis to identify the key themes in our 16 interviews (Braun and Clarke 2006; Thomas 2006). The first author coded all of the interviews, and picked out the six most theme-dense interviews which the second author coded as well. Then, the union of all codes were consolidated, resulting in 967 unique codes. These codes were relatively detailed (e.g., "Trade-offs are never explicit numbers of value, but things like how long can a team work on this," "Team started out as engineers using canonical notions of fairness from papers, then as evolved got product and legal involved and asking more substantive questions"). The authors iteratively clustered the codes through five rounds of thematic groupings until consensus was reached on six high-level themes (section headers) composed of 15 sub-themes (subsection headers).

Results

We surface six high-level themes about what motivates companies to prioritize RAI. *External cues* and *regulatory pressures* are frequently cited as the most powerful reasons. Four more arise as important and may be overlooked: *organizational macro-motivators* like how a company is funded; relevance to *company success*, *company culture and individuals*, and *effort and ease* of implementation. Topics in each category are not mutually exclusive, but we group them to bring

Ptcpt.	Industry	# Empl.	Ptcpt.	Industry	# Empl.	Ptcpt.	Industry	# Empl.
P1	Healthcare	10-100	P7	Hiring	10-100	P12	Social Media	1k-50k
P2	Entertainment	1k-50k	P8	Entertainment	50k+	P13	Finance	50k+
P3	E-commerce	100-1k	P9	Finance	50k+	P14	SaaS	10-100
P4	Finance	50k+	P10	Electronics	50k+	P15	Electronics	50k+
P5	Manufacturing	50k+	P11	[redacted]	50k+	P16	SaaS	10-100
P6	Entertainment	100-1k						

Table 1: Each participant (Ptcpt.) with their company’s industry and employee number (# Empl.). The roles are (data) scientist/engineer: 10, product: 2, manager/executive: 4.

structure to the strategies we have for affecting the prioritization that companies place on RAI. Overlaid on these themes are the different actors that currently have power (Widder et al. 2023) and access to each lever (e.g., journalists are relevant for impacting *external cues*, but so are consumers).² We underline the named actors throughout, and in Appendix Fig. ?? summarize these group-specific strategies by difficulty of implementation. Of course, these are not the only people able to pull these levers, just some of the closest.

Ultimately we find a complex web of factors, and acting on one or even a set of strategies is not guaranteed to increase RAI prioritization. However, while we are careful to warn against a confinement to these options which would foreclose additional strategies, our analysis of previously successful motivators helps to inspire a promising path forward and equip practitioners with a toolkit of strategies.

External Cues

One of the largest motivators voiced by participants to prioritize RAI was external cues. This refers to axes like media publicity, general public awareness, and customer demands.

Publicity Publicity, and specifically the fear of bad press, was brought up by every participant as a major motivating factor. As P14 mentions, the “impetus for change usually comes when something bad happens, right?” P[4, 5, 8, 9, 11, 15] all specifically bring up the utility of having a ready list of public failures of other companies, and P15 explains, “the most effective thing was pointing to failures of other companies.” P4 echoes that they are “a wealth of knowledge of cautionary tales that help me make those conversations and get that buy-in,” since examples convey how even when you think you have control, things can always go wrong. The Partnership on AI’s AI Incident Database³ was brought up by a few participants as being useful for this.

However, as great of a motivator as bad publicity is, there are also downsides. P12 mentions that even though their company was working on RAI, they were too scared to publicly talk about it or implement anything that would make those efforts visible, lest that lead to bad publicity. Their company preferred to remain silent on the topic rather than risk being in the public eye for claiming to be doing RAI

²We do not touch on collective action or activism (Belfield 2021) as none of our participants brought it up, but it is a critical component of power (Widder et al. 2023).

³<https://partnershiponai.org/workstream/ai-incidents-database/>

work that could be criticized. Given how powerful of a motivator bad press is, it seems important to maintain it, but useful to bring nuance so that companies have the space to pursue RAI without facing quick and reductionist headlines.

In general, good press did not seem to be a very strong motivator. The only notable instance was a public ethics award given to one participant’s organization, which they cited as a motivator for the company to live up to the award. Even though rewarding behavior that should be expected is not ideal, creating reward incentives for companies could help communicate it as something the public values. This is analogous to the LEED rating system or Fair Trade labels, though all have been shown to be susceptible to different forms of ethics-washing (Bowen and Aragon-Correa 2014; Doan et al. 2017; Low and Davenport 2005).

Overall, journalists have immense power in carefully exposing irresponsible AI practices at companies, and maintainers and contributors of incident databases give practitioners an arsenal of arguments to draw from. However, concerningly, there is a decline in journalism funding (Bauder 2022).

Public awareness Downstream of journalism is the growing public awareness that AI is not objective and can cause disparate harm. P[1, 3, 4, 5, 7, 8, 11, 12, 14, 15] all bring up how RAI has entered the public consciousness, and that this collective zeitgeist has facilitated getting company buy-in on RAI. P14 notes, however, that they believe RAI is becoming less trendy than it once was. In the same way that social justice movements (e.g., #MeToo, #BlackLivesMatter) can lose momentum, we need to collectively ensure these topics are not just a passing trend. Again, journalists will play a critical role in maintaining relevance in public awareness.

This public awareness can also have other positive effects for RAI: P8 brings up the externality of “even if those obvious problems don’t have a legal hook yet, because the more things are in the press, the more likely someone will sue. Even if those lawsuits might not have great standing or grounds, just for the optics of it or if those lawsuits get to a discovery phase... it could lead to embarrassing things being uncovered.” Public awareness can also affect the recommendations that consultants and other influential external advisors provide. P4 shares anecdotally that they were told “it’s a lot easier as an external advisory function or a provider or a partner to be able to push... to get some things done versus internally.” P10 also describes how external consultants provide recommendations based on what they project to be future trends, and for such a speculative enquiry, matters

that are high on the public radar can become a self-fulfilling prophecy. Many participants also brought up large language models and the increase in generative AI as leading to an increase in awareness around the capabilities and harms of these models (P[4-13, 15]). However, in many cases the excitement to develop and deploy these generative models outweighed the increase in concern. As P11 describes, “It mostly pushed the innovation impulse... So it did increase the number of people who saw the salience [of RAI], but it didn’t necessarily give it on balance more power because it got less momentum than the business.”

At this point, our conflation in wording by using “RAI” as a catch-all term runs into collisions. P5 brings this up by differentiating between those components of RAI that are more known and thus prioritized, compared to lesser-understood components which are left behind. They explain “things like fairness, privacy, those are already very high awareness so we get lots of great traction around conducting privacy impact assessments and around doing bias testing and performance testing. But when it comes to things like accountability and transparency and labor impacts, that gets a little murkier.” Thus, now that we have established the first level of public awareness, internal advocates can help move towards more nuanced notions of RAI that encompass components like labor impact.

Customers The third aspect of *external cues* is the power of the customer⁴ (P[1-3, 5-14]). This is relevant both for retaining current customers as well as attracting new ones. However, it did seem to be differently prioritized by companies across domains. For example, healthcare and finance are two domains that stand out as caring more about customer perception of RAI: P9 says “as a bank, if you damage your brand, people will not trust you as much as they did before.” Another relevant axis for how much customer demand can drive RAI is the type of customers desired. P7 notes that “if you don’t build trust among underserved communities, and among small companies who basically cannot afford to have bias in their systems, we will not make profits, we won’t get customers.” This is similar to how prior work has found that startups need to work harder than larger companies in establishing trust and legitimacy (Winecoff and Watkins 2022).

One way customers can express demand is through their purchasing behavior and technology usage. P8 mentions boycotts and grassroots campaigns that have shifted company prioritization, and P14 mentions successful examples of other companies in the marketplace. In the technology space, companies built on the customer demand of privacy include DuckDuckGo and Signal. Customers can exercise signaling behaviors with their capital in order to communicate to companies their desire for RAI. This serves as motivation to companies because, as P4 puts it, “Why are you leaving customers on the table? Because of your unfair practices, right? You could catch more business.”

⁴We use the term “customer” to include both consumers in customer-facing apps as well as clients in business-to-business organizations.

Regulatory Pressures

Beyond *external cues*, the other most frequently cited motivator for prioritization of RAI were *regulatory pressures*.

Compliance First and foremost, nearly all of our participants bring up the relevance of legal compliance as a reason to prioritize RAI (P[0-5, 7-15]). More specifically, it is the fine: “If you do something and your company is liable, it’s billions [of dollars], so actually you need to take care of these things” (P10).

However, P[1, 8, 11, 12, 15] also bring up situations where the presence of regulatory pressures, whether real or imagined, actually hamper RAI adoption. One such well-studied obstacle is that of regulation prohibiting the collection and/or usage of sensitive attributes, which is often necessary in RAI work to even measure the amount of unfairness that is present (Ho and Xiang 2020; Kumar, Hines, and Dickerson 2022; Bogen, Rieke, and Ahmed 2020; Andrus et al. 2021). P15 mentions that they find legal work-arounds, for example collecting pronouns instead of gender. P12 also mentions that in many cases it is very unclear what the regulation is asking for and prohibiting, and so trying to adhere to it is itself a challenge.

Ultimately there is a somewhat mixed bag with respect to whether more regulation is desired or not. P8 notes two reasons more regulation might actually be harmful to RAI goals: “uncertainty over whether the act of engaging in the work itself might be creating risk” and “if something is required in all cases in a blanket manner where that requirement is actually not well-suited for a certain circumstance, but people have to do it anyway, that leads to a lot of confusion.” Echoing the second point, P12 adds that in lower risk domains, more legal regulation could be harmful by adding extra overhead with small impact. Policymakers can spend more time with experts to understand the specific domains where more regulation can help with RAI prioritization, and how to scope it.

Exceeding Compliance In addition to merely adhering to regulation, we find two reasons that regulatory pressures may motivate companies to do more than the law explicitly asks. The first is to prepare for future regulation that may be more strict. For example, P13 points out that “we think ‘hey, there might be change in the future,’ so we already have systems in place where we already are following a much stricter regulation than the one that already exists.” Conversely, P4 mentions that because regulation might change in ways you cannot anticipate, there is a negative incentive to go above and beyond in case what you do conflicts with future regulation. The second reason is to demonstrate to regulators that self-regulation is sufficient, and to fend off stricter impositions. By demonstrating responsibility, companies hope to have a seat at the table when regulation is discussed. For example, P8 says “failure of private actors to act of their own volition is likely to lead to more strict requirements... those requirements might end up being more difficult to achieve than what the voluntary work would have done.”

There are a number of ways to interpret these statements, and one takeaway is that even signaling done by regulators about future regulation can help motivate action. However,

given the ambiguity that regulation can take, it is important to communicate it in ways that encourage benevolent action, and are accompanied by substantive follow-through.

Organizational Macro-Motivators

Participants also described broader organizational factors for pursuing RAI which range from taking advantage of parallel social movements like environmental sustainability to external market forces and company profitability.

Parallel social corporate concerns Participants cite being able to leverage growing enthusiasm around other social movements, namely environmental sustainability and data privacy to boost RAI. P[4, 5, 8, 10] describe the benefits of leveraging the paths paved by environmental movements in industry, such as carbon footprint reporting, environmental impact assessments, and ESG investing. These structures have helped define the role of expertise for environmental activists, providing them with not only a voice but also authority to weigh in on trade-offs and decisions. P4 described leveraging executive commitments to environmental efforts, “If you can connect it to those sustainability conversations, I think that has the most traction right now.” Practitioners can understand what tactics worked best in parallel spaces to better define the roles and authority of expertise for RAI.

Corporation and funding type Structural financial aspects of an organization also affect the ability to implement RAI practices. Both P7 and P14 cite non-profit compared to for-profit organizations as more conducive places for prioritizing RAI principles. One participant’s company is registered as a Public Benefit Corporation, indicating that their company mandate is to specifically benefit the public in some way.⁵ Outside of the corporation’s filing status, P1 and P7 also discuss the additional flexibility that receiving funding from external grants allows. For these participants, grants were used as a funding source for certain RAI-related projects which allowed for longer explorations into aspects of fairness and equity. Founders can exercise choice in signaling their mission through their corporation type and filing status, and RAI advocates can look for grants to fund specific projects they are working on.

Stage in profit cycle There are differing financial pressures depending on the maturity of an organization. P9 and P15 both described the benefits of larger, more established companies having the space to focus on maintenance and impact in a way that smaller companies do not. P1 also describes, “it really depends on where the company and product is, and it’s lifecycle. [That helps decide] how much you invest [on RAI].” Meanwhile, for startups, their stage in the company life cycle matters. Startups that are not currently raising capital may have more room to pursue principles of impact, as P7 describes, they “don’t have to appease investors or expectations of profit” as heavily as they might have had to when seeking to raise a new round of funds.

⁵Public Benefit Corporations are different from a “B Corp,” or a Certified B Corporation, which is a third-party certification of social and environmental standards operated by the non-profit B Lab (Koehn 2016).

Relatedly, P[5, 7, 12] specifically discussed recent downward economic forces resulting in company-wide hiring pauses and “an organizational environment in which people just have more on their plate and it’s easier for things to fall through the cracks because there is incentive to focus on different things” (P5). This makes it difficult for RAI efforts to grow or even continue. Conversely, when the market is good, P[1, 2, 5, 10, 12] all cited greater access to resources “to focus on impact rather than revenue” (P1). RAI efforts often rise when the economy is good and ebb when the economy is bad. Venture capitalists can be variance reducers here, supporting RAI in times of both wealth and austerity.

Company Success

Individuals frequently spend effort connecting RAI to company success to increase buy-in within the scope of already-existing corporate incentives (Ali et al. 2023; Metcalf, Moss, and danah boyd 2019; Phan et al. 2022; Rakova et al. 2021). P[4, 5, 8, 15] brought up looking at their company history to see what worked best in the past (e.g., “finding examples of where we’ve seen positive changes in our operations, and really do audits of how did they get there?” - P4). While some strategies will be company-specific, part of the motivation of this work is to consolidate narratives so RAI practitioners can learn from each other. All of the strategies in this section are ways for internal practitioners to frame RAI for increased prioritization.

Competitor Differentiator One powerful motivator for companies to prioritize RAI is as a differentiator from competitors (Widder and Nafus 2023). Many participants specifically brought up publicity around companies that have mistreated or gotten rid of their RAI teams as a motivator for companies to set themselves apart (P[2, 4, 5, 7, 10, 12, 15]). P5 describes that their company is trying to “establish [themselves] in a position of thought leadership. But also distinguish ourselves against the types of FAANG companies that have made very visible pivots away from investing in these kinds of efforts.” P10 adds that differentiating based on RAI offerings rather than aspects like generative AI can be more appealing because it is easier to compete with larger, more well-sourced companies on this front.

However, the incentive to use RAI as a peer differentiator is not guaranteed to lead to more responsible AI. As P8 comments, “all the companies are now trying to compete on demonstrating responsibility for more advanced forms of AI. But that’s very undefined, and so, you don’t have to do much to try to demonstrate. There’s no one who can call you out on saying that’s not enough because no one agrees on what’s enough.” This indicates an urgent need for journalists to partner with experts to weigh in for the public on which company’s RAI efforts are substantive. We see the incentive to ethics-wash throughout our work, and will elaborate on this more during our recommendations.

Long-term priorities Another route practitioners take to tie RAI into company success is through alignment with long-term priorities. For example, through saving time down the line by incorporating RAI now. P8 motivates that “scrambling to do something later on would be much more

difficult than proactively adjusting course to do something that would be reasonable, and probably closer to the practices that the company was already doing.” P10 invokes the idea of “ethical debt” (Cunningham 1992; Petrozzino 2021), where even though incorporating ethics may take more time now, you will save time in the long run.

Another way is linking RAI to a company’s strategic goals or values. P11 gives the example that “I think the concept and the term is popular, and I think people wanted to be seen as engaging with it often for pro-social reasons... So if the CEO sends out a letter that says... we believe in rigor and quality. And then you say, this aligns with our company’s goals of rigor and quality.” P5 reports similarly that it was “usually about reiterating some form of some value everybody has or that everybody should have or that everybody hopefully has, and then giving them recommendations that can help them be more in line with that goal.” P6 points out how a specific company-level change helped with this, “we went through and re-did our values at a company level, which is part of the reason why I think that I’m able now to talk about it because user-centricity is now one of our values when it wasn’t explicit before.”

Short-term priorities On the other hand, when a company decides how to prioritize large numbers of demands, short-term priorities can often beat out long-term ones, and these tend to, but do not always, deprioritize fairness. P[1-3, 5, 6, 8, 10, 12, 14, 15] all bring up that ultimately RAI is secondary to utility of the product and financial success. P6 straightforwardly explains “you need revenue to pay for the people who are actually going to solve the [RAI] problem,” and P13 laments “I don’t think any company at all is going to care that much about fairness or responsible AI until it starts hurting their wallets.” For the most part, however, it is not that people do not care about RAI, but rather the relative urgency of RAI compared to other priorities feels smaller. The themes we discuss can help to create this urgency.

Thus, it often takes a shift in perspective or reframing of RAI under a utility perspective to elevate its priority (P[0-11, 13-15]). P5 gives an example where they told a team: “you guys who said that you need to collect as much data as possible... If we make things more explainable and transparent that will get people to talk more and that means more data for us.” P4 argues for a bigger shift, and that when a perceived tension between RAI and utility comes up, “why is this a compromise? Again, I think it’s a failure of imagination... Are we going to hurt people in the process? I think there needs to be a shift in how we think about value.” In terms of how personalized these reframings need to be, P11 describes that there was a “standard battery of these concepts that we would draw on, but when you’re talking to someone in model risk versus a data scientist, their needs and goals are very different,” and P8 echoes there are “people who truly deeply care about AI safety and could be convinced to care about fairness, but it’s not through a lens of compliance. It’s through a lens of AI safety and long term benefit to people.” Even though this reframing needs to be catered per role and personal priority, techniques for these roles and value sets can be shared amongst RAI practitioners.

One concrete strategy to “sell RAI” that practitioners found successful is tying together both numbers and a story (P[6, 11, 14]). As P6 explains, “a truly complete story has both the numbers element and the emotional, look this is what a user actually experiences.” Whereas our participants often find the story easier to craft, many express a need for better measurement techniques to formulate the number (P[4, 6, 8, 11, 12, 13, 14]). Fairness is notoriously hard to measure (Jacobs and Wallach 2021; Wang et al. 2022; Katzman et al. 2023), but there is an urgent need for interpretable, quantitative measures, both for decision-makers as well as customers (“a lot of the fairness issues are super technical and may or may not relate to the ability to explain to consumers whether they’re being treated fairly or not.” - P8). RAI measurements are important in communicating and explaining progress to non-technical stakeholders, who are often the decision-makers choosing what to prioritize. P8 describes companies “not really preferring to dive into really ambiguous spaces that could both alleviate and create risk without clarity.” And yet, given that the ambiguity around RAI interventions will likely never go away entirely, we need to educate decision-makers about this inherent ambiguity and convey that it does not detract from RAI importance but does require more patience. As with many cases where it is contextual who is well-positioned to act on this, it could be an external consultant, executive, or coworker who is best situated to communicate this.

Culture and Individuals

Despite the common themes we’ve drawn from our interviews, our findings also reveal how subjective and context-specific the reasons for prioritizing RAI can be. Here, we discuss the role of company culture and individuals.

Company culture Starting at the very high-level, country-specific cultural norms are one component that influence RAI prioritization. P10 works for a company based outside of the USA and mentions “People don’t change jobs in [country]... So practically their personality does get connected [with the company].” They explain this means that an individual’s sense of morality becomes intertwined with that of the company’s, and they thus prioritize ethics.

Industry domain and company size are also relevant to company culture. Similar to how customers tend to be more valued in domains like healthcare and finance, P13 from finance notes that “we’re not like the other tech companies where we are constantly trying to be at the cutting edge” and explains that this slowness allows better RAI integration. In terms of company size, P9 shares that compared to startups that may try to hire people who already have expertise, their larger company hopes to retain employees for a long time, and can afford to offer training sessions to teach already existing employees about RAI topics.

Overall, many participants indicate that company culture feels ingrained and inherent, indicating it may be hard to change. In fact, P[1, 7, 10, 14] all specifically used the phrase “DNA” to describe company culture around RAI. Due to the often separate RAI teams and complicated organizational structures, participants also brought up the need to spread in-

ternal awareness for within-company RAI initiatives (P[5, 9-15]). Awareness of within-company RAI initiatives can help shift the culture, but often involves extra labor on the part of internal practitioners to spread the word.

Individuals We find that often it is specific circumstances and individual experiences which cause people to champion RAI. We separate the impact of executives in more managerial roles compared to individual contributors.

Participants share that individuals higher up in the organization have some, but not full, power in increasing RAI prioritization. The cases where a single individual's desire has been enough is when they have sufficient power (P10), the company is hierarchical (P10), or the company is small enough and leadership is unified in their mission (P7). The circumstances leading to these individual efforts can be highly specific. For example, one participant mentioned their CEO attended the FAccT conference and was inspired. P16 also mentioned the relevance of personal risk: "for executives, because it's not only company risk or losing your job, but it's personal brand risk." The other side of how personal these choices are is that individual circumstances can also cause RAI deprioritization. P8 points out two possible reasons: executives don't want to set a precedent for themselves that they will be held to in the future, and if previous RAI attempts did not pay off they may now have resistance. The unique circumstances that bring powerful individuals to RAI suggest it may be hard to implement any uniform strategy, but that it is still critical to create as many of these "serendipitous" situations as possible, e.g., by engaging more stakeholders at conferences in the Responsible AI space, e.g., FAccT, AIES. One key intervention point is education, as that is a formative time where many will encounter a set of pedagogy.

RAI work frequently falls to people of marginalized identities (P[1, 2, 5-7, 11, 16]) doing what is often volunteer work that does not help career advancement (Deng et al. 2023; Ali et al. 2023). P[4, 7, 11] mention the prevalent burnout amongst people working in this space, and P11 notes a unique challenge with respect to human resource problems. They explain that the same interpersonal problems are present in RAI career ladders as in other engineering career ladders, but unlike more generalized careers where someone can switch teams, there is often only one RAI team within a company, and thus no way to work under different management. This can lead to high turnover and losses in institutional knowledge (P5) (Ali et al. 2023).

Another part of the problem is no one thinks RAI is their responsibility (Widder and Nafus 2023; Lancaster et al. 2023). P5 explains that "engineers are very much in line with legal going 'we just make stuff, I'm not responsible for what people do.'" Individuals seem to believe that letting the default persist is not making a choice, when it is in fact a value-laden choice of its own.

As with many kinds of collective efforts, much of the success of RAI efforts boils down to interpersonal dynamics. P5 observes that "there are some individuals internally... who have spent a lot of social capital to make [RAI] happen" and P8 adds that there is a "lot of internal maneuvering to

get things through that are not built into a natural incentive structure in any organization." P11 elaborates that an important relevant factor is general pleasantness of working with someone even in terms of following up on emails. They elaborate that "you can demonstrate being thoughtful, having good reasons, being a good collaborator over a period of time. Then even more resistant people, they'd sort of warm up... So, you're building out that your credibility, your reputation." P12 similarly mentions that their team focuses on being proactive with RAI metrics ready so any team that comes to them will not have to wait long. It is unfair that RAI workers are held to a higher standard, which can contribute to the prevalent burnout. P[5, 11, 13] mention how RAI teams are overburdened and under-resourced, leading to additional strain.

Related to company "DNA," individuals also sometimes think of being ethical as personally intrinsic. When asked about RAI adoption, many simply bring up the relevance of individuals being "good" (P[1, 6-12, 14]), and doing this work because they think it is the right thing to do or want to be seen as "pro-social." P11 finds that in convincing people to adopt RAI, appealing to their morals often worked. However, intention without expertise can also have downsides, and believing that inherent "goodness" is sufficient for RAI can be a serious harm. P8 flags that "sometimes people just start [RAI efforts] because they're interested and that can actually be problematic if they measure something that is using a definition a company's already decided is not the preferred one or has detriments." Thus, we need better education that emphasizes how difficult RAI work is, and that it requires collaborating with experts.

Effort and Ease

Finally, we consider how the effort and ease of RAI implementation in practice contribute to adoption likelihood.

Implementation effort Though perhaps obvious, many participants explicitly bring up how easier RAI efforts are far more likely to be adopted (P[4-6, 8, 10, 11, 13-15]). However, navigating ethical dilemmas is rarely easy, and necessarily requires time to both contemplate and implement. In practice, this often requires RAI teams to "direct towards the more ethical path of least resistance" (P4). P5 says that compromises are necessary in order to not be ignored: "We're not in the business of saying 'stop.' We try not to be a gate. We try to be a steering wheel." These responses can be demoralizing and feel like accepting the status quo rather than systemic change, and P5 laments that "you're never going to reach the radical thing that you want to reach," but explains this to be better than the counterfactual. There is a difficult balance in implementing (even incremental) improvements, even if they do not always solve the source problem (Green 2021). Sometimes though, speed and ease of an RAI intervention can be more of a rhetorical tool, and emphasized in framing rather than actual practice, by comparing it to what would take longer. P15 gives an example from trying to create an internal review system and "trying to frame it as it's such a quick and easy process. All you have to do is [...]. [It] doesn't take that long and we really pared it down com-

pared to what the traditional university does.” Some other ways to make RAI intervention easier include tying into already existing pipelines. P[1, 2, 4, 10, 13] all voiced how it was beneficial to piggyback fairness issues onto other internal changes (Deng et al. 2023), e.g., P1 remembers “taking advantage of the fact that we’re integrating to new hardware as an opportunity to make this [RAI] change” and P13 mentions “if you’re already spending a bunch of money to try and curate the data... then it makes sense to have the same team look at the data and make sure that data is also fair.”

We also find three key components of implementation difficulty that can influence individual support or resistance. Individuals do not like delays to deployment, detest paperwork, and do not find RAI interesting to work on. For delays in deployment, by the time a team encounters RAI concerns they often already feel their product is ready to be deployed, and see any delays as burdensome. However, the seemingly simple solution of incorporating RAI from the start is hard in practice. Beyond just being difficult to change the status quo (P4), P13 paints the landscape: “at any given point of time, there’s just like thousands of teams working on thousands of different models. So it’s basically impossible for [RAI folks] to be involved in the process of model creation right from the get-go. Because 80% of the models that we work on don’t ever see the light of day because they’re all proof of concepts” (Vakkuri et al. 2020). On the other hand, delaying can also be a compromise. P12 describes using “time-boxes” where they convince teams to delay deployment by some predetermined number of weeks during which the team must demonstrate effort on implementing RAI interventions. They find this compromise useful, because teams have an end date assurance. However, it can be hard to enforce such effort-based approaches, and after deployment, effort is not a sufficient metric to end-users who suffer the consequences. Another key point of tension is paperwork. P5 shares that they use paperwork as both a carrot and a stick, offering to reduce the bureaucratic compliance paperwork for teams that can demonstrate they are incorporating RAI concerns: “You can just kind of lean on people and annoy people with paperwork.” Finally, P5 notes that some engineers just want to work on what they find technically interesting, to the point of sometimes even prioritizing that over what customers explicitly say that they want. P[1, 5, 6, 9-11, 13] all say that people see RAI as grunt work, and if it were intellectually interesting, it might be more prioritized. RAI presents genuinely challenging technical problems, so often this just needs to be made clear to practitioners. Through both framing and changes, if RAI interventions were presented by practitioners in ways that seem to include justifiable and reasonable delays, do not increase seemingly useless paperwork, and are technically interesting, much of this individual resistance, which can be significant, would go away. And much of this can be in the framing—practitioners often do not feel paperwork is actually effective, so if it were streamlined to be more efficient and those filling it out were genuinely convinced about the importance (P11), these obstacles would be reduced.

Research to practice In translating RAI research into practice, a number of challenges are brought up by practitioners that impede RAI prioritization. P12 discusses that academic research often doesn’t consider things like the significant barriers of the engineering tech stack that make in-processing techniques far harder to implement compared to post-processing techniques, as the latter are easier to scale since they don’t require coordination with large numbers of teams. In terms of the more principles-based approaches, P8 finds it’s “easy for the arguments to get theoretical and philosophical,” and P4 says “that’s where I think the rubber meets the road, how do we actually get things done?” Overall, prior work covers the numerous research-to-practice gaps in RAI (Holstein et al. 2019), and we bring this up as a research area that can ease the adoption of RAI, and thus the willingness of companies to implement.

Recommendations Forward

Overall, we learn from our interview studies a wide range of ways to motivate companies to increase RAI prioritization. We present a summary of shared strategies in the Appendix (Fig. ??), organized by the actor who is most able to enact each. From these, a few directions stand out to us as the more promising ways forward. Whereas we have endeavoured to remain mostly descriptive thus far, this section will be more prescriptive.

Accountability for ethics-washing. A recurring theme is that in many cases, ethics-washing is more incentivized than substantive ethics work. To shift the incentives away from the former to the latter, we suggest greater collaboration between RAI experts and journalists to hold companies accountable and make sense of cheaply made ethics statements. P8 and P10 both remark that performativity can actually lead to substantive action, e.g., “The performative-ness of it actually does open space for more performance of it” (P8). This is only true if companies are not rewarded for empty ethics statements, but also not unduly punished for publicizing genuine yet unsuccessful efforts. Another way to communicate substantive RAI progress is by creating measurements which are more resistant to ethics-washing. These include measurements that incorporate human feedback and connect to more concrete harms (Blodgett et al. 2020; Wang et al. 2022, 2023).

Ethics modules. Incorporating RAI requires two things on the part of an individual: intention and expertise. Both can be taught through the integration of ethics modules during the education process that instill this as a priority, which may conflict with market pressures, early on (Smith et al. 2023; Fiesler, Garrett, and Beard 2020; Raji, Scheuerman, and Amironesei 2021; Grosz et al. 2019). However, ethics modules in their current state may be overly focused on educating future executives and engineers on how they might identify and resolve ethics problems. Instead, the focus should be on communicating the difficulty and ongoing nature of ethics so that students have a genuine understanding and appreciation of the difficulty of the problem, and know how to collaborate with ethics experts rather than feeling falsely empowered to resolve these problems on their

own. Like our finding that people sometimes feel that being a “good person” is a sufficient qualification to do RAI work, it’s important to communicate that RAI and ethics work is a field that requires collaborating with those with expertise (McLennan et al. 2022).

Shifting structural incentives. We have seen alternatives to more traditional corporation structures such as non-profits and public benefits corporations (PBCs), as well as indicators like ESG ratings. In addition, there are new initiatives to support founders and investors prioritizing RAI.⁶ These have the potential to allow a company to be judged by shareholders for more than just their immediate profits, creating room for values like product quality and public perception (Dorff, Hicks, and Solomon 2020). There is a parallel to value-based healthcare (Porter and Teisberg 2006; Teisberg, Wallace, and O’Hara 2020), which is a framework for healthcare that shifts the incentive structure to focus on patient health outcomes rather than cost reduction. Of course, any of these could be used for ethics-washing, and prior work has postulated on why the supposed benefits of new forms of incorporation—signaling brand quality, limiting legal liability, and attracting capital—may not always materialize (Koehn 2016). However, these kinds of broader structural shifts can create an incentive structure that more highly prioritizes issues like RAI. They would require making the requirements for such legal qualifications and ratings strict enough to filter out disingenuous efforts, as well as potentially incorporating more financial benefits like tax incentives into PBCs and ESG ratings, instead of relying only on improved public perception.

Research informed by needs. Participants shared that RAI research outputs often don’t match their needs (Holstein et al. 2019). Two specific ways to close this gap emerge from our findings. One is that researchers can pursue integrated interventions that fit well into the typical software engineering stack. While this differs across companies, there are shared characteristics which have consistently made post-hoc interventions easier, and thus more likely, to be implemented (P12). This is in comparison to other interventions that require touching many parts of the tech stack. New research can pursue whether there are methods besides post-hoc that are also easier to fit into the stack. At the same time, this will require companies being more open about their internal processes and constraints. Another direction is in the interdisciplinary work happening between technology and the law. As participants expressed, it is often unclear which interventions are permitted by the law. P12 expressed that they were able to have more productive conversations with company lawyers after engaging with research that clarified the connection between law and concrete RAI approaches (Ho and Xiang 2020; Xiang 2020; Kumar, Hines, and Dickerson 2022). Companies should also share ways of implementing RAI that is legally permissible with each other. For example, P15 mentions sometimes using pronouns instead of gender in RAI implementation, because while the latter is legally impermissible, the former is allowed. In this way, practitioners can share effort in navi-

gating the complexity of regulation while pursuing substantive RAI efforts.

Discussion and Limitations

Overall, in our work we unveil a set of strategies for affecting RAI prioritization. A limitation is that the recommendations may err on the side of being too accommodating to the world as it is rather than speaking to an ideal world—we do not wholly endorse all strategies so much as provide a slate of options from which to choose from. For example, is it right to say we should make fairness problems more technically interesting and “fun” in order to incentivize technical developers to work on them? Perhaps not, but in our endeavor of presenting all discovered methods of increasing RAI prioritization, it is certainly one possibility. As Audre Lorde famously put it “For the master’s tools will never dismantle the master’s house. They may allow us temporarily to beat him at his own game, but they will never enable us to bring about genuine change” (Lorde 1984). Ultimately by relying on the strategies that have historically worked within the confines of capitalism, our suggestions may be more of a patch than a treatment for the source problem. However, given the time it can take for more radical changes to take place, there is still benefit to working on increasing RAI prioritization in companies today, so long as it does not detract from such greater changes. RAI practitioners individually struggle towards this in their separate companies, and our work consolidates their techniques and efforts. Even though it seems like RAI will not be prioritized over profit, in our work we unveil that this is not clear-cut! There are successful strategies for stakeholders to elevate RAI even then, and the details of our work can be used as a toolkit for enacting this RAI prioritization. In the meantime, we should not lose hope for more radical changes and continue to push on that front as well (Hampton 2021; Davis, Williams, and Yang 2021; Green 2020; Gorz 1964).

Conclusion

In this work, we look to the corporate landscape for examples of what has worked in prioritizing RAI in order to unveil the strategies at our disposal and pressure points to lean on in order to better motivate corporations under the landscape of capitalism. We find a complex picture of numerous actors able to exert pressure in numerous points, with no guaranteed outcomes. And yet, by drawing lessons from these past successes and applying enough force, we can collectively increase RAI prioritization. However, if we believe that the current state of RAI adoption and prioritization is not sufficient, then it is clear that just relying on these strategies are not enough. While we believe this will be a useful toolkit for those looking to increase the prioritization of RAI, these strategies we unveil are just one set of the total possible interventions we should pursue.⁷

⁶<https://www.rilabs.org/>

⁷Appendix is located at: <https://arxiv.org/pdf/2405.03855>.

Acknowledgments

We thank Namrata Mukhija and Daniel Nissani for helping with participant recruitment, Victoria Vassileva for initial conversations on these topics, and Amy Winecoff for feedback on the draft. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship to AW, and was work initiated during AW's internship at Arthur.

Research Ethics and Social Impact

Ethical considerations. In this work, the biggest ethical consideration was the treatment of our participants and ensuring their comfort with their level of anonymity. This can be a sensitive topic that bears on someone's workplace environment, and we worked hard to ensure each participant felt comfortable with the level of anonymity they would be provided. To achieve this, we sought consent before either recording the interview or taking written notes. We also started each interview by asking each participant how they would prefer their background to be identified (e.g., job role, company industry, company size). Upon writing our draft and in response to some participants' concerns, we did not associate the role with each participant, as we found that our results were not specific to each person's role. Finally, we also sent the completed draft to each participant to ensure that they were ultimately comfortable with the level of anonymity they were represented with, since we understand that this can be a loaded space. We compensated each participant \$25 for an interview which ranged from 30-60 minutes.

All of the data was stored on Google Drive, which is secure and encrypted. We will delete all interview records upon final publication.

Positionality. One of the most relevant aspects of positionality for this piece is that two of the authors work in industry for a company in the RAI space, and one author is in academia. These positions affect the perspectives on corporations. In positions of relative socioeconomic privilege, none of the authors are personally susceptible to some of the more insidious effects of AI and so may find it easier to write about incremental, non-revolutionary improvements, compared to others.

Adverse impact. As discussed in Sec. , by having scoped our work to past efforts, we hope to not foreclose considerations of more radical change. However, we realize that papers of this sort that propose changes within existing capitalist structures could be seen as legitimizing the current structure, which is not our goal. We hope that our work can shine a light on directions forward, in parallel with other movements like greater instances of collective action or abolition in the cases that call for it.

References

Ali, S. J.; Christin, A.; Smart, A.; and Katila, R. 2023. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Andrus, M.; Spitzer, E.; Brown, J.; and Xiang, A. 2021. "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Bauder, D. 2022. US newspapers continuing to die at rate of 2 each week. *AP News*.

Belfield, H. 2021. Activism by the AI Community: Analysing Recent Achievements and Future Prospects. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Blodgett, S. L.; Barocas, S.; III, H. D.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Association for Computational Linguistics (ACL)*.

Bogen, M.; Rieke, A.; and Ahmed, S. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Bowen, F.; and Aragon-Correa, J. A. 2014. Greenwashing in Corporate Environmentalism Research and Practice: The Importance of What We Say and Do. *Organization & Environment*.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*.

Bromley, P.; and Powell, W. W. 2012. From Smoke and Mirrors to Walking the Talk: Decoupling in the Contemporary World. *Academy of Management Annals*.

Cohen, L.; Gurun, U. G.; and Nguyen, Q. H. 2020. The ESG-innovation disconnect: Evidence from green patenting. Technical report, National Bureau of Economic Research.

Cunningham, W. 1992. The WyCash Portfolio Management System. *ACM SIGPLAN Special Interest Group on Programming Languages*.

Davis, G. F.; and Thompson, T. A. 1994. A social movement perspective on corporate control. *Administrative science quarterly*, 141–173.

Davis, J. L.; Williams, A.; and Yang, M. W. 2021. Algorithmic reparation. *Big Data & Society*.

De-Arteaga, M.; Feuerriegel, S.; and Saar-Tsechansky, M. 2022. Algorithmic Fairness in Business Analytics: Directions for Research and Practice. *Production and Operations Management*.

Deng, W. H.; Yildirim, N.; Chang, M.; Eslami, M.; Holstein, K.; and Madaio, M. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Dimaggio, P. 1988. *Interest and agency in institutional theory*. Ballinger Publishing Co.

Doan, D. T.; Ghaffarianhoseini, A.; Naismith, N.; Zhang, T.; Ghaffarianhoseini, A.; and Tookey, J. 2017. A critical comparison of green building rating systems. *Building and Environment*.

Dorff, M. B.; Hicks, J.; and Solomon, S. D. 2020. The Future or Fancy? An Empirical Study of Public Benefit Corporations. *Harvard Business Law Review*.

- Fiesler, C.; Garrett, N.; and Beard, N. 2020. What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. *Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE)*.
- Gilpin, R. 2000. *The challenge of global capitalism: The world economy in the 21st century*. Princeton University Press.
- Gorz, A. 1964. *Strategy For Labor: A Radical Proposal*. Beacon Press.
- Green, B. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Green, B. 2021. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*.
- Green, J.; and Green, J. 2004. *Qualitative Methods for Health Research*. Sage Publications.
- Grosz, B. J.; Grant, D. G.; Vredenburg, K.; Behrends, J.; Hu, L.; Simmons, A.; and Waldo, J. 2019. Embedded EthiCS: Integrating Ethics Across CS Education. *Communications of the ACM*.
- Hallett, T.; and Ventresca, M. J. 2006. Inhabited institutions: Social interactions and organizational forms in Gouldner's Patterns of Industrial Bureaucracy. *Theory and society*, 35: 213–236.
- Hampton, L. M. 2021. Black Feminist Musings on Algorithmic Oppression. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Hartzmark, S. M.; and Shue, K. 2023. Counterproductive sustainable investing: The impact elasticity of brown and green firms. In *Counterproductive Sustainable Investing: The Impact Elasticity of Brown and Green Firms: Hartzmark, Samuel M.—uShue, Kelly*. [SI]: SSRN.
- Ho, D. E.; and Xiang, A. 2020. Affirmative Algorithms: The Legal Grounds for Fairness as Awareness. *University of Chicago Law Review Online*.
- Holstein, K.; Vaughan, J. W.; III, H. D.; Dudík, M.; and Wallach, H. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Conference on Human Factors in Computing Systems (CHI)*.
- Hopkins, A.; and Booth, S. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Katzman, J.; Wang, A.; Scheuerman, M.; Blodgett, S. L.; Laird, K.; Wallach, H.; and Barocas, S. 2023. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. *AAAI Conference on Artificial Intelligence*.
- King, B. G.; and Soule, S. A. 2007. Social movements as extra-institutional entrepreneurs: The effect of protests on stock price returns. *Administrative Science Quarterly*, 52(3): 413–442.
- Koehn, D. 2016. Why the New Benefit Corporations May Not Prove to Be Truly Socially Beneficial. *Business & Professional Ethics Journal*, 35(1): 17–50.
- Kumar, I. E.; Hines, K. E.; and Dickerson, J. P. 2022. Equalizing Credit Opportunity in Algorithms: Aligning Algorithmic Fairness Research with U.S. Fair Lending Regulation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Lancaster, C. M.; Schulenberg, K.; Flathmann, C.; McNeese, N. J.; and Freeman, G. 2023. “It’s Everybody’s Role to Speak Up... But Not Everyone Will”’: Understanding AI Professionals’ Perceptions of Accountability for AI Bias Mitigation. *Journal of Responsible Computing*.
- Lorde, A. 1984. The Master’s Tools Will Never Dismantle the Master’s House. *Sister Outsider: Essays and Speeches*.
- Low, W.; and Davenport, E. 2005. Has the medium (roast) become the message? The ethics of marketing fair trade in the mainstream. *International Marketing Review*.
- Madaio, M.; Egede, L.; Subramonyam, H.; Vaughan, J. W.; and Wallach, H. 2022. Assessing the Fairness of AI Systems: AI Practitioners’ Processes, Challenges, and Needs for Support. *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*.
- Madaio, M.; Stark, L.; Vaughan, J. W.; and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Conference on Human Factors in Computing Systems (CHI)*.
- McLennan, S.; Fiske, A.; Tigard, D.; Müller, R.; Haddadin, S.; and Buyx, A. 2022. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*.
- Metcalfe, J.; Moss, E.; and danah boyd. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*.
- Meyer, A. D. 1982. Adapting to Environmental Jolts. *Administrative Science Quarterly*, 27(4): 515.
- Meyer, J. W.; and Rowan, B. 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American journal of sociology*, 83(2): 340–363.
- Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*.
- Munn, L. 2023. The uselessness of AI ethics. *AI and Ethics*.
- Petrozzino, C. 2021. Who pays for ethical debt in AI? *AI and Ethics*.
- Phan, T.; Goldenfein, J.; Mann, M.; and Kuch, D. 2022. Economies of Virtue: The Circulation of ‘Ethics’ in Big Tech. *Science as Culture*.
- Porter, M. E.; and Teisberg, E. O. 2006. Redefining Health Care: Creating Value-Based Competition on Results. *Harvard Business Review Press*.
- Powell, W. W.; and DiMaggio, P. J. 2012. *The new institutionalism in organizational analysis*. University of Chicago press.
- Raji, I. D.; Scheuerman, M. K.; and Amironesei, R. 2021. You Can’t Sit With Us: Exclusionary Pedagogy in AI Ethics

- Education. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.
- Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; and Lennon, M. 2020. Principles to Practices for Responsible AI: Closing the Gap. *European Conference on AI Workshop on Advancing Towards the SDGs*.
- Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; and Lennon, M. 2021. Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine*.
- Schneiberg, M. 2007. What's on the path? Path dependence, organizational diversity and the problem of institutional change in the US economy, 1900–1950. *socio-Economic review*, 5(1): 47–80.
- Shilton, K. 2012. Values Levers: Building Ethics into Design. *Science, Technology, & Human Values*.
- Smith, A. 2002. An Inquiry into the Nature and Causes of the Wealth of Nations. *Readings in economic sociology*, 6–17.
- Smith, J. J.; Payne, B. H.; Klassen, S.; Doyle, D. T.; and Fiesler, C. 2023. Incorporating Ethics in Computing Courses: Barriers, Support, and Perspectives from Educators. *Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE)*.
- Suchman, M. C. 1995. Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review*, 20(3): 571–610.
- Teisberg, E.; Wallace, S.; and O'Hara, S. 2020. Defining and Implementing Value-Based Health Care: A Strategic Framework. *Academic Medicine*.
- Thomas, D. R. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*.
- Vakkuri, V.; Kemell, K.-K.; Jantunen, M.; and Abrahamson, P. 2020. “This is Just a Prototype”: How Ethics Are Ignored in Software Startup-Like Environments. *Agile Processes in Software Engineering and Extreme Programming*.
- Wang, A.; Bai, X.; Barocas, S.; and Blodgett, S. L. 2023. Measuring stereotype harm from machine learning errors requires understanding who is being harmed by which errors in what ways. *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*.
- Wang, A.; Barocas, S.; Laird, K.; and Wallach, H. 2022. Measuring Representational Harms in Image Captioning. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Weber, M.; Roth, G.; and Wittich, C. 1978. *Economy and Society: An Outline of Interpretive Sociology*. Economy and Society: An Outline of Interpretive Sociology. University of California Press. ISBN 9780520035003.
- Widder, D. G.; and Nafus, D. 2023. Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society*.
- Widder, D. G.; Zhen, D.; Dabbish, L. A.; and Herbsleb, J. 2023. It’s about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them? *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Winecoff, A. A.; and Watkins, E. A. 2022. Artificial Concepts of Artificial Intelligence: Institutional Compliance and Resistance in AI Startups. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.
- Xiang, A. 2020. Reconciling Legal and Technical Approaches to Algorithmic Bias. *Tennessee Law Review*.