Making Transparency Advocates: An Educational Approach Towards Better Algorithmic Transparency in Practice

Andrew Bell, Julia Stoyanovich

New York University, New York, NY, USA
alb9742@nyu.edu, stoyanovich@nyu.edu

Abstract

Concerns about the risks and harms posed by artificial intelligence (AI) have resulted in significant study into algorithmic transparency, giving rise to a sub-field known as Explainable AI (XAI). Unfortunately, despite a decade of development in XAI, an existential challenge remains: progress in research has not been fully translated into the actual implementation of algorithmic transparency by organizations. In this work, we test an approach for addressing the challenge by creating transparency advocates, or motivated individuals within organizations who drive a ground-up cultural shift towards improved algorithmic transparency.

Over several years, we created an open-source educational workshop on algorithmic transparency and advocacy. We delivered the workshop to professionals across two separate domains to improve their algorithmic transparency literacy and willingness to advocate for change. In the weeks following the workshop, participants applied what they learned, such as speaking up for algorithmic transparency at an organization-wide AI strategy meeting. We also make two broader observations: first, advocacy is not a monolith and can be broken down into different levels. Second, individuals' willingness for advocacy is affected by their professional field. For example, news and media professionals may be more likely to advocate for algorithmic transparency than those working at technology start-ups.

Introduction

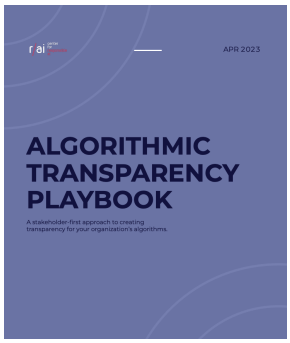
There are widespread concerns about the significant risks posed by artificial intelligence (AI) systems in both the public and private sectors, particularly for marginalized or historically disadvantaged groups (Hu and Rangwala 2020; Sapiezynski, Kassarnig, and Wilson 2017; Obermeyer et al. 2019). One major risk factor, compounded by the release of Large Language Models, is the lack of transparency in AI systems that make high-stakes decisions (Rudin 2019; Kirilenko et al. 2017). These concerns have led to the emergence of *Explainable Artificial Intelligence* (XAI), a subfield focused on studying how well AI systems can be understood by humans (Bell, Nov, and Stoyanovich 2023). While significant progress has been made in developing and evaluating methods for explaining complex AI systems—through multi-disciplinary approaches combining machine learning and human-computer interaction (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016; Datta, Sen, and Zick 2016; Covert, Lundberg, and Lee 2020; Abdul et al. 2020; Yang et al. 2019; Holzinger, Carrington, and Müller 2020; Bell et al. 2022)—evidence suggests that companies and organizations using AI often undervalue or remain unaware of these methods (Dastin 2022; Hill 2022). As a result, XAI faces an existential challenge: how can we move beyond the research setting to *ensure the real-world implementation of transparent AI systems* (Beattie, Taber, and Cramer 2022)?

While government regulation seems like the natural solution to this challenge, the rapid development of AI technologies has greatly outpaced public oversight, resulting in an incomplete patchwork of laws and regulations (Jobin, Ienca, and Vayena 2019). To date, over 70 nations and intergovernmental organizations have published over 1,000 AI strategies, actions plans, policy papers, or directives (OECD.AI 2021). Unfortunately, many of these efforts face a significant limitation: they remain uncertain about how to meaningfully implement transparency (Jobin, Ienca, and Vayena 2019; Loi and Spielkamp 2021; Gasser and Almeida 2017). For example, in the United States, the Biden Administration has issued broad AI guidance under the *Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (The White House 2023), but the U.S. Congress has taken little action to strengthen regulations or enact specific laws governing AI transparency practices.

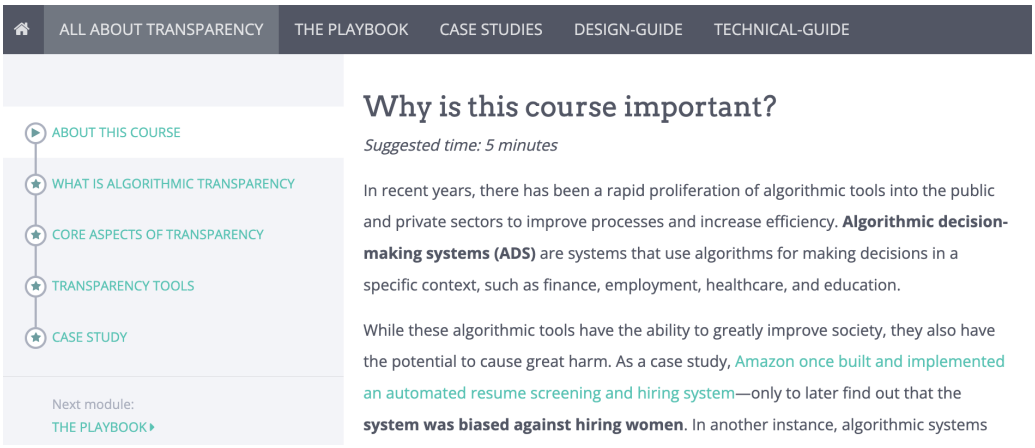
Meanwhile, the private sector demonstrates inconsistent interest in algorithmic transparency and responsible AI practices. As an example, during substantial layoffs in May 2023, Microsoft disbanded its entire AI ethics team.[1] In light of these challenges, this work explores a complementary pathway to ensuring safe, transparent AI: educating and empowering *transparency advocates* within organizations.

We define transparency advocates as a subset of what Meyerson (2003) called "tempered radicals," or committed employees who drive institutional change over time (sometimes clandestinely), with a focus on algorithmic transparency. Tempered radicals can be very effective: in one example, over a 30-year period, a Black senior executive at a

[1] https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs



(a) Cover of the *Algorithmic Transparency Playbook*, available for download at https://r-ai.co/algorithmic-transparency-playbook.



(b) The free open-source online course is available at https://r-ai.co/transparency-playbook-course.

Figure 1: Preview of online materials.

large West Coast bank covertly hired 3,500 women and minority employees to improve company diversity. We hypothesize that transparency advocates can similarly drive significant, bottom-up organizational change toward improved algorithmic transparency.

Study approach. Over several years, we created a workshop on algorithmic transparency that provides an overview, introduces best practices and tools for implementing transparency, and outlines strategies for advocating it. This workshop is part of the education and training mission of the Center for Responsible AI at New York University (NYU R/AI) [2]. We conducted the workshop twice: first with professionals in news and media, and then with professionals at technology startups. Through one-on-one interviews and pre- and post-workshop surveys, we explored two research questions: (1) How effective is the workshop in increasing participants' algorithmic transparency literacy? and (2) Can the workshop increase participants' willingness to advocate for algorithmic transparency in their professional lives?

Summary of findings. In total, 27 professionals (15 from news and media and 12 from technology startups) participated in the workshops. We divide our results into two categories: workshop-specific findings and broader findings.

Workshop findings. Interviews with participants demonstrate that the workshops were effective in both teaching algorithmic transparency and increasing participants' willingness to advocate for it. With respect to the former, participants expressed that the workshop was particularly helpful in uncovering knowledge gaps in algorithmic transparency. Three participants noted that it made them realize "they didn't know what they didn't know [about transparency]."

In terms of advocacy, four participants reported taking advocacy actions in the days following the workshop. Most significantly, one participant attended an organization-wide strategy meeting on AI and spoke up on behalf of transparency, citing the workshop as a major motivator and directly applying its lessons.

Our qualitative results are also supported by pre- and post-workshop surveys, which suggest that the workshop increased participants' general understanding of algorithmic transparency as well as their willingness to advocate for it.

Broader findings. First, we found that advocacy is not a monolith and can occur at three levels. The first is conversational, where individuals raise awareness by speaking to their colleagues about the importance of algorithmic transparency. The second is implementational, where AI engineers directly implement tools to improve transparency (e.g., create data sheets (Gebru et al. 2021), model cards (Mitchell et al. 2019), or nutritional labels (Stoyanovich and Howe 2019)). The third is influential, where individuals attempt to steer the overall direction of their organization's overall direction towards transparency.

Second, we found that transparency advocacy depends on the domain of use. For example, professionals in news and media are more likely to advocate for transparency but may lack the tools to act on it. In contrast, those in technology startups are more likely to have the tools and technical knowledge but lack the resources to prioritize it. Understanding these domain-specific barriers will be critical for achieving meaningful algorithmic transparency in practice.

Related Work

Organizational barriers to transparency. Organizations often forgo transparent, responsible AI practices due to misaligned incentives. This is especially true in for-profit organizations, where such practices may be perceived as barriers to increasing revenue (Metcalf, Moss et al. 2019). When companies do pursue responsible AI practices, it's often in response to external pressures rather than proactive, value-driven decisions by leadership (Metcalf, Moss et al. 2019). Employees at one large technology company reported that their day-to-day work prioritized profit-motivated tasks, such as launching products and increas-

[2] https://r-ai.co/education

ing user engagement, over ethical considerations (Metcalf, Moss et al. 2019; Madaio et al. 2020; Rakova et al. 2021). In fact, the priorities of companies can, at times, be in *direct tension* with responsible AI. For example, optimizing user engagement—a common profit-driven objective—can lead to irresponsible outcomes, such as creating online radicalization pipelines (Phadke, Samory, and Mitra 2022).

Another organizational barrier to transparency is that practitioners are at times unable to identify their companies' *specific goals* with respect to broad terms like AI ethics (Metcalf, Moss et al. 2019; Raji et al. 2020). Additionally, there are human "blind spots,"—individuals in different, disconnected teams who are unaware of responsible AI practices (Holstein et al. 2019). As a result, the responsibility for AI ethics often falls to motivated individuals, often referred to as "ethics owners" (Metcalf, Moss et al. 2019).

Regulation. At present moment, AI regulation is insufficient for ensuring organizations are transparent about their use of algorithms. Regulation is also not a silver bullet—even among the few positive examples, omissions or loopholes exist that can be exploited. For example, the EU AI Act establishes transparency obligations for AI systems, differentiating the required level of transparency based on predefined AI risk categories. This approach has been criticized as *all AI systems* have the potential to pose high risk.

The majority of existing AI directives and strategies lack specificity and means of enforcement (UNICRI 2020; Munn 2023). As an example, consider the enacted European Union's General Data Protection Regulation (GDPR), which includes text to guarantee individuals a "right-to-explanation," or a right to be given an explanation for an output of an algorithm that impacts them. However, despite being one of the most expansive and robust data protection laws to date, GDPR's right-to-explanation has yet to deliver any meaningful benefits for citizens (Selbst and Powles 2018; Doshi-Velez et al. 2017; de Laat 2022).

Tempered radicals. Myerson coined the term "tempered radicals" to describe individuals who influence change within organizations slowly but steadily over time (Meyerson 2003). Tempered radicals prefer to make bottom-up change, rather than relying on company leadership or government regulation. There are numerous successful examples of tempered radicalism, especially in ethically motivated practices. These individuals have advanced minority representation, inclusion, and sustainability across various contexts, including companies, universities, and religious organizations (Walton and Kirkwood 2013; Griffiths, Pio, and McGhee 2022; Meyerson and Tompkins 2007; Ngunjiri, Gramby-Sobukwe, and Williams-Gegner 2012; Kirton, Greene, and Dean 2007).

Tempered radicals offer a natural approach to advancing responsible AI practices. Interestingly, ground-level employees already seem to bear this responsibility: interviews with researchers revealed that employees at a large tech company often feel it is *their* job to represent ethical technology values (Rakova et al. 2021).

AI education. In recent years, there has been a growing number of initiatives teaching *AI literacy*, helping citizens better understand AI at a conceptual level, including its opportunities and risks (Domínguez Figaredo and Stoyanovich 2023). While these initiatives have been primarily focused on K-12 students and emphasized the technical aspects of AI (i.e., computer programming) (Domínguez Figaredo and Stoyanovich 2023; Williams 2021), several promising courses have emerged that teach responsible AI to the general public (Lewis and Stoyanovich 2021; Bell, Nov, and Stoyanovich 2023). Best practices for AI education are still evolving and require a multi-disciplinary effort to incorporate the social sciences, pedagogy, and data science (Lewis and Stoyanovich 2021). This work intends to build up this knowledge base.

Methods

The Algorithmic Transparency Workshop

Development process. We designed a 2-hour workshop on algorithmic transparency, consisting of 5 modules that provide an overview of transparency, describe best practices and tools for its implementation, and outline strategies for advocating for transparency. The workshop also includes a role-playing activity where participants act out practical barriers to implementing transparency. Workshop materials—including the content of the modules, the full *Transparency Playbook*, and a slide-deck version of the course—are free to use and can be found on the workshop website.

Prior to conducting the workshops, we published peer-reviewed work on a stakeholder-first approach to implementing algorithmic transparency and created a practitioner-focused playbook on the topic (Bell, Nov, and Stoyanovich 2023). This work was informed by interviews with professionals across a variety of domains and backgrounds, including large technology companies, algorithmic safety audit firms, government organizations, and early-stage startups.

We added two topics to the workshop based on practitioner input: transparency for procured tools (common in government organizations) and balancing transparency with intellectual property considerations (common in industry). We also conducted multiple "trial runs" of the workshop, refining the content and the presentation based on participant feedback. For example, the *Transparency Tools* module, which contains five real-world case studies of algorithmic transparency tools, emerged in response to feedback from trial run participants who suggested to add examples of how transparency is used in practice.

Structure and design. Workshop modules are summarized in Table 1. Each module includes a lecture component, with 2-3 interactive elements, such as questions, reflections, and short discussions. For example, in the *Transparency Tools* module, participants explore technical tools associated with real-world algorithmic systems, such as model cards (Mitchell et al. 2019) and explainer dashboards[3]. The module features a live demo of an explainer dashboard, fol-

[3]https://titanicexplainer.herokuapp.com/multiclass

Module	Topics	Time (*mins*)
All About Transparency	Defining algorithmic transparency, types of transparency, stakeholders and their goals	20
Transparency Tools	Transparency labels, model cards, feature importance, Shapley values, dashboards	10
The Transparency Playbook	How to disclose the use of AI, transparency for algorithms protected by IP or procured from vendors, the gold standard approach to transparency	15
Breakout Activity	Role-playing game where participants take on either the role of pro-transparency or anti-transparency managers at a fictional news and media company	15
Becoming a Transparency Advocate	Common objections to transparency (i.e., "transparency means more costs," "transparency means sacrificing privacy") and how to rebut them	10

Table 1: Modules covered in the workshop.

lowed by a discussion with the audience about the types of transparency it offers and which stakeholders it benefits.

Breakout activity. The moderated breakout activity aims to increase participant engagement and deepen their connection to the content, and improve participants' ability to advocate for transparency by demonstrating the tensions that emerge when organizations consider "disclosing" their algorithm use. For example, we aimed to highlight that some managers may object to algorithmic transparency to protect intellectual property, and to equip participants to counter these arguments.

Participants are asked to imagine themselves as managers at a fictional social news startup (i.e., *HackerNews*) that had recently implemented an AI content moderation tool. Half of the participants are asked to role-play *skeptical managers* (i.e., the "Devil's Advocate" position), opposing disclosure of the AI tool, while the other half are asked to role-play *pro-transparency* managers. Participants then make and record their arguments for and against transparency for different stakeholders at their organization (e.g., affected users, developers, managers, etc., as discussed in the workshop content) according to their role. An example of a completed activity can be seen in the Appendix of the extended version of the paper (Bell and Stoyanovich 2024).

Recruitment, Participation, and Domains of Study

We conducted the workshop twice, each time for a different audience. The first workshop, held virtually, was attended by 15 news and media professionals. The second, conducted in person, was attended by 12 professionals working at or with technology startups. At both workshops, we administered a pre- and post-workshop surveys and conducted semi-structured follow-up interviews.

News and media organizations and technology startups are deeply affecting (and affected by) emerging AI technologies. The release of generative AI tools like ChatGPT has significantly disrupted workflows in news and media companies, prompting existential conversations about adaptation. AI has had a similar impact on the startup landscape: roughly a quarter of all venture capital funding went to AI-based startups in 2023, as compared to only 11% in 2018.[4]

We ran the workshops through the NYU Center for Responsible AI, in partnership with other entities at our university that work within the respective domains: AI & Local News at the NYC Media Lab [5] co-hosted the virtual workshop and helped recruit participants from news and media, and the NYU Tandon Future Labs [6] co-hosted the in-person workshop and helped recruit participants from their startup community. In total, 27 domain professionals attended the workshops: 15 in the news and media workshop and 12 in the startup workshop. Nearly all participants work with AI technologies and algorithmic tools. Their job titles include Chief Digital Officer, Data Journalist, Newsroom Developer, UX Designer, Startup Co-founder, and Product Strategist.

Content customization. To improve relevance and practical applicability, we customized the content for the audience's domain. This customization manifested in two ways. First, we tailored case studies and examples to tools and systems used in news and media or startups, respectively. For example, the news and media workshop included a discussion about the media company *CNET*'s recent use of AI to generate articles on its site, many of which contained errors.[7] As part of the workshop, we discussed what went wrong and how *CNET* could have benefited from a transparent AI strategy. Second, we designed the breakout activity to address a practical challenge specific to the participants' domain.

Data Collection and Analysis

Interviews. In the days following each workshop, we conducted semi-structured interviews with 7 participants, whose domains and expertise are detailed in Table 2. The full interview protocol is included in the Appendix of the extended version of the paper (Bell and Stoyanovich 2024). To encourage participants to speak candidly about their workplace experiences and their employers, we chose not to record the interviews. Instead, we took detailed notes throughout the sessions, capturing quotes relevant to our research.

Pre- and post-workshop surveys. We administered an 8-question pre-workshop survey to assess participants' base-

[4]https://news.crunchbase.com/ai-robotics/us-startup-funding-doubled-openai-anthropic-2023/

[5]https://engineering.nyu.edu/research-innovation/centers/nyc-media-lab/projects/ai-local-news

[6]https://futurelabs.nyc/

[7]https://www.theverge.com/2023/1/25/23571082/cnet-ai-written-stories-errors-corrections-red-ventures

line knowledge of algorithmic transparency and their willingness to advocate for it. Following the workshop, we administered an 18-question post-workshop survey to evaluate its impact. The survey was adapted from previous work by Lewis and Stoyanovich (2021), who used a similar study design for a technical course on responsible data science. In total, 6 constructs were measured (reported in Table 3), and our choice of scale for measuring each construct was consistent with Lewis and Stoyanovich (2021). Full survey details are included in the Appendix of the extended version of the paper (Bell and Stoyanovich 2024).

Analysis. To analyze our qualitative data—which included both interview notes and answers to free-response questions on the post-workshop survey—we followed the six stage approach to thematic analysis described by Braun and Clarke (2006). First, we familiarized ourselves with the data by re-writing and organizing interview notes and noting initial recurrent ideas (e.g., "frequent use of AI"). Second, we generated 33 initial codes by highlighting salient interview quotes and ascribing them a code (e.g., "thinking about user needs," "arbitrary thresholds for disclosure"). Interview coding was done manually using a word processor. Third and fourth, we collated the 33 codes into 6 separate themes, and evaluated their robustness over two separate working sessions. Fifth, we definitively named the themes, and, sixth, we analyzed them through the lens of our two research questions: (1) How effective is an educational workshop in increasing participants' algorithmic transparency literacy? and (2) Can the workshop increase participants' willingness to advocate for algorithmic transparency in their professional lives, becoming *transparency advocates*?

Regarding quantitative data, only 15 participants completed both the pre- and post-workshop survey (7 from news and media and 8 from technology startups). Due to the relatively small sample size and the greater substantive value of our qualitative findings, we report survey results as descriptive statistics and forgo statistical analysis.

Results

Thematic Analysis Findings

Frequent use of internally developed and procured algorithmic tools. All participants reported frequent or almost daily contact with AI in their jobs, utilizing a wide range of algorithmic tools, including generative AI, recommender systems, computer vision, and tools for carrying out A/B testing. Participants from news and media mentioned the use of both third-party and proprietary AI tools. In contrast, participants from startups relied almost exclusively on proprietary, internally developed tools.

Uncovering knowledge gaps. Participants generally found the workshop useful, with each identifying different aspects as most impactful. These included learning about the different levels of transparency (P1 and P6), existing toolkits (P2 and P6), stakeholder identification (P3), and transparency tensions (P6 and P7). Several participants highlighted that the workshop's greatest strength was uncovering knowledge gaps. P2, P3, and P4 said that it helped them realize "they didn't know what they didn't know." P3 reflected that, after the workshop, they realized their organization "probably doesn't do enough disclosure and transparency." Interestingly, P7 offered a different perspective, stating, "[The workshop] showed me I know a lot more than I thought I knew."

Taking action. Participants P2, P3, and P4 reported taking transparency advocacy action in the days following the workshop. P2 said they had "already used the [course material]" in conversations with colleagues, and P4 noted "I've probably had five conversations about AI transparency compared to close to zero [before the workshop]." P3 stated that they had already begun implementing elements from the workshop, such as stakeholder identification, into their workflow. P7 said, "I like the concept of being a transparency influencer—it shows that we can make [impacts] no matter where we are in the loop."

Notably, in the days immediately following the workshop, P4 stepped into the role of a"transparency advocate" by speaking for algorithmic transparency at an organization-wide AI strategy meeting. They described the experience: "I was just in a TV workshop and [I asked if] we need to be disclosing and transparent [about AI], and then it got really quiet." But they optimistically added, "it's definitely on the agenda now."

Participants' advocacy actions appeared to be motivated, at least in part, by the workshop. Many noted that it provided valuable resources for advocacy. As P2 explained, "I always would've advocated for transparency anyway ... " but the workshop improved their potential for transparency advocacy by making them aware of different types of resources related to transparency."

P3, P4, and P5 also commented on their future plans for transparency advocacy based on the resources introduced in the workshop. P3 said "If we ever go down the road of building a model, it feels like [model cards] are something we should probably do." Similarly, a participant from the startup workshop, who was not interviewed but completed the post-workshop survey, wrote, "As a co-founder at a health-tech company, I will definitely advocate for algorithmic transparency due to the benefits not only in terms of business acumen, but as a responsibility."

Organizational challenges: resisting change P1 and P4 reported that their organizations recently held internal meetings to discuss AI strategies and create a "Code of Conduct" for its use—an indication that news and media organizations are responding to the rapid proliferation of AI tools. However, both participants pointed out that this transition has not been smooth. P1 stated that discussions around the use of generative AI for creating story headlines has "ruffled a lot of feathers," dividing the organization into two schools of thought: those who are *pro* new AI tools, and those who are against their use and are resistant to change. Similarly, P7, who works in local government, described a comparable organizational culture where not everyone is interested in understanding AI. Reflecting on an internal AI workshop, P7 remarked, "[The room] was full, but relative to the amount of people in the building, it was not that full." They added,

Alias	Domain	Expertise
P1	Newspaper	Works for a print media company with an online presence; experience in journalism
P2	Researcher	Holds a doctorate in human-computer interaction; expertise in transparency
P3	Newsroom	Manages team of developers who are also journalists at popular online media company
P4	Local TV news	Works on development at syndicated local TV news network; has journalism experience
P5	Startup Co-founder	CMO at an early-stage, consumer-facing, AI-based company
P6	Product Manager	Worked in UI/UX design for an early-stage startup; experience in the defense industry
P7	Project Manager	Works in local government using open data to improve K-12 curriculum design

Table 2: Domain and expertise of interview participants.

"Some people are interested [in AI] but not everyone."

Organizational challenges: market fundamentalism Participants working in startups presented a different organizational challenge: while people are interested in AI, the primarily focus remains on generating revenue. When asked how often algorithmic transparency (or responsible AI) comes up as a topic of conversation at their AI-based startup, P5 said "Not once... not with investors, not with attorneys, not with users. User's don't care. [Users only ask], 'is it fast? Is it cool?' That's it." P6 shared a similar reflection from their startup experience, saying: "When you are in a small, bootstrapped startup, resources are tight. As a product manager, it was my job to ruthlessly prioritize [what we work on]." They added that during rapid development cycles, "transparency might not make the cut." P5 poignantly summarized this theme, stating: "The race has begun [in AI]... [Anything not related to winning] is just not a concern... [AI start-ups] don't care. Part of that is capitalism, part of that is behavioral."

When is transparency necessary? A surprising theme that emerged from interviews was that each participant seemed to have a personal barometer for determining *when transparency is necessary.* P1, P3, and P4 agreed that disclosing the use of AI for generating news article headlines did not seem necessary. P6 mentioned that after the workshop, they had begun grappling with the question, "Is more transparency always better?" They pointed out evidence suggesting that excessive transparency about an algorithm may overwhelm end users with information (Bell et al. 2022; Jacoby, Speller, and Kohn 1974). Both P5 and P6 noted that, in the startup domain, online Terms of Service and User Agreements are sometimes used to circumvent the responsibility of transparent AI practices. P6 said that the legal team often thinks, "We can just put it in the disclaimer," to legally cover unethical practices. They added, "[For users,] once they click opt-in, it's game over."

Another finding was the existence of "unwritten rules" for transparency. For example, P1 mentioned the following guideline: "If you are questioning whether or not you need to tell people [about AI], you need to tell people." Similarly, P6, drawing from their experience in the defense sector, reflected, "There is a difference between intentionally hiding something and being intentional about what you show."

Pre- and Post-workshop Surveys

Table 3 shows the pre- and post-workshop survey results for each measured construct. Post-workshop means were higher across all constructs, with low standard deviation, suggesting that the workshop positively impacted participants' understanding of algorithmic transparency and their willingness to advocate for it. The largest improvements were observed in participants' general understanding of transparency and their awareness of transparency stakeholders.

Discussion

Overall, our findings suggest that the workshop had a positive impact both on participants' understanding of algorithmic transparency and their willingness to advocate for it. We hypothesize that this success was driven by two key factors.

First, the workshop features a strong curriculum developed as part of a multi-year, ongoing project by the authors. As mentioned in the Methods section, the workshop was developed with practitioner feedback and has been continually improved. This iterative development underscored the importance of making the workshop freely available online so that others may use and replicate it.

Second, the workshop content was tailored to the participants' respective domains, exemplifying a stakeholder-first approach to responsible AI literacy (Domínguez Figaredo and Stoyanovich 2023). For example, P4—a professional in news and media—mentioned that their organization used AI in a manner nearly identical to the fictional scenario featured in the breakout activity of the news and media workshop.

Levels of Advocacy

Promisingly, we observed several *real-world* actions taken by participants following the workshops. These actions appear to have been motivated, at least in part, by the workshop. We categorize these actions into three categories: *conversational, implementational, and influential*.

Conversational. Three participants said that after the workshop they had more frequent conversations about algorithmic transparency with colleagues and peers. While such conversations may not *directly* affect organizational change, they play a role in increasing awareness about algorithmic transparency, which in and of itself can significantly influence behaviors over time (Jacobsen and Jacobsen 2011).

		Pre- survey		Post- survey	
Construct	**Scale**	μ_0	σ_0	μ_1	σ_1
General Understanding of Algorithmic Transparency	10-point	5.00	1.80	7.79	0.97
Theme: Benefits and Purpose	3-point	2.00	0.78	2.64	0.50
Theme: Stakeholders	3-point	1.64	0.50	2.57	0.51
Theme: Tensions Between Goals	3-point	1.71	0.73	2.50	0.52
Willingness for Advocacy: Professional Life	5-point	4.14	0.77	4.71	0.47
Willingness for Advocacy: Personal Life	5-point	3.71	0.91	4.43	0.65

Table 3: Means (μ) and standard deviations (σ) for pre- versus post-workshop survey responses ($n = 15$)

Implementational. Several individuals implemented algorithmic transparency directly into their work. This type of advocacy can be characterized by narrow, immediate, and *practical* changes rather than broader organizational culture shifts. As a prime example, one participant, a manager of a small team of software developers, reported integrating stakeholder identification material from the workshop directly into their team's workflow. Notably, this action *did not require organizational approval*—as a manager, they had the authority to make proactive workflow changes to promote algorithmic transparency. From our perspective, this type of advocacy is critical for driving bottom-up change within organizations, and may be consistent with the at-times clandestine actions of tempered radicals (Meyerson and Tompkins 2007).

Influential. This type of advocacy is characterized by individuals taking action to affect cultural change towards algorithmic transparency within their organization. The most notable example comes from a participant who, in the days following the workshop, spoke up about algorithmic transparency at an organization-wide AI strategy meeting. They raised concerns about the organization's approach to transparency and disclosure, using arguments learned at the workshop. Interestingly, they encountered some of the same negative responses anticipated during the role-playing activity described in Methods section. Ultimately, the participant left the meeting feeling optimistic and hopeful that their company would start taking steps toward more transparent algorithmic practices. This example highlights the potential ripple effect of the workshop: by inviting one person to think more deeply about algorithmic transparency and providing them with basic tools for advocacy, we may have contributed to a medium-sized U.S.-based media company adopting more responsible AI practices.

The Importance of Domain-of-use

Unexpectedly, we found vast differences in the attitudes towards algorithmic transparency in new and media vs. technology startups. For professionals in news and media, where there is an ethos of being "champions of the truth," transparency and disclosure align naturally with their values. As a result, many in this domain *already care about transparency*, and only need guidance on *how to implement it effectively*. On the other hand, professionals at fast-paced technology startups often *cannot* afford to care about algorithmic transparency, despite possessing the technical knowledge to implement it. Although many of these professionals may care about transparency and responsible AI practices, the circumstances of AI-focused startups may prevent them from finding the time and resources to effectively act on those values.

This finding aligns with prior researcher, which found that employees at large technology companies often express interest in value-driven work but are not given the time or space to pursue it (Metcalf, Moss et al. 2019). We encourage further researcher to explore whether, and how, these barriers can be overcome, noting that domain-of-use must be taken into account in such research.

Conclusion, Lessons and Social Impact

This work outlines a promising approach to using education to affect ground-up change towards responsible AI. With this study, we hope to contribute to the broader effort to translate responsible AI practices from research settings into real-world applications, especially in high-stakes domains.

As with many studies of this nature, some findings are limited by the small sample size, posing questions about the scalability of our approach and the generalizability of our findings. This issue was further exacerbated by participant drop-off in the online workshop, which motivated us to conduct the second workshop in person. Additionally, because participation was optional, there was likely some bias toward individuals already inclined to become transparency advocates. While this may have enhanced engagement, it also highlights scalability and generalizability concerns.

We plan to continue exploring educational approaches to promote values aligned with responsible AI and encourage others to do the same. To support this effort, we have made all workshop materials used in this study publicly available online and free to use.

Course website —
https://r-ai.co/transparency-playbook-course

The algorithmic transparency playbook —
https://r-ai.co/algorithmic-transparency-playbook

Acknowledgements

This research was supported in part by NSF Awards No. 1922658, 2326193, 2312930, and NSF GRFP (DGE-2234660).

References

Abdul, A.; von der Weth, C.; Kankanhalli, M.; and Lim, B. Y. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Beattie, L.; Taber, D.; and Cramer, H. 2022. Challenges in Translating Research to Practice for Evaluating Fairness and Bias in Recommendation Systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 528–530.

Bell, A.; Nov, O.; and Stoyanovich, J. 2023. The Algorithmic Transparency Playbook: A Stakeholder-first Approach to Creating Transparency for Your Organization's Algorithms. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–4.

Bell, A.; Solano-Kamaiko, I.; Nov, O.; and Stoyanovich, J. 2022. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–266.

Bell, A.; and Stoyanovich, J. 2024. Making Transparency Advocates: An Educational Approach Towards Better Algorithmic Transparency in Practice. arXiv:2412.15363.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Covert, I.; Lundberg, S. M.; and Lee, S. 2020. DBLP:journals/corr/abs-2004-00668 Feature Contributions Through Additive Importance Measures. *CoRR*, abs/2004.00668.

Dastin, J. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*, 296–299. Auerbach Publications.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, 598–617. IEEE.

de Laat, P. B. 2022. Algorithmic decision-making employing profiling: will trade secrecy protection render the right to explanation toothless? *Ethics and Information Technology*, 24(2): 17.

Domínguez Figaredo, D.; and Stoyanovich, J. 2023. Responsible AI literacy: A stakeholder-first approach. *Big Data & Society*, 10(2): 20539517231219958.

Doshi-Velez, F.; Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Scott, K.; Schieber, S.; Waldo, J.; Weinberger, D.; et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Gasser, U.; and Almeida, V. A. F. 2017. A Layered Model for AI Governance. *IEEE Internet Comput.*, 21(6): 58–62.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H. M.; III, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Commun. ACM*, 64(12): 86–92.

Griffiths, C.; Pio, E.; and McGhee, P. 2022. Tempered radicals in manufacturing: Invisible champions of inclusion. *Journal of Management & Organization*, 1–22.

Hill, K. 2022. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*, 170–177. Auerbach Publications.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 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.

Holzinger, A.; Carrington, A.; and Müller, H. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz*, 1–6.

Hu, Q.; and Rangwala, H. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students.

Jacobsen, G. D.; and Jacobsen, K. H. 2011. Health awareness campaigns and diagnosis rates: evidence from National Breast Cancer Awareness Month. *Journal of health economics*, 30(1): 55–61.

Jacoby, J.; Speller, D. E.; and Kohn, C. A. 1974. Brand choice behavior as a function of information load. *Journal of marketing research*, 11(1): 63–69.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. Artificial Intelligence: the global landscape of ethics guidelines. *CoRR*, abs/1906.11668.

Kirilenko, A.; Kyle, A. S.; Samadi, M.; and Tuzun, T. 2017. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3): 967–998.

Kirton, G.; Greene, A.-M.; and Dean, D. 2007. British diversity professionals as change agents–radicals, tempered radicals or liberal reformers? *The International Journal of Human Resource Management*, 18(11): 1979–1994.

Lewis, A.; and Stoyanovich, J. 2021. Teaching responsible data science: Charting new pedagogical territory. *International Journal of Artificial Intelligence in Education*, 1–25.

Loi, M.; and Spielkamp, M. 2021. Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 757–766.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4765–4774.

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 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.

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

Meyerson, D. 2003. Tempered Radicals: how everyday leaders inspire change at work Boston. *MA: Harvard Business School Press, xi–xii*, 41: 59–60.

Meyerson, D.; and Tompkins, M. 2007. Tempered radicals as institutional change agents: The case of advancing gender equity at the University of Michigan. *Harv. JL & Gender*, 30: 303.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In danah boyd; and Morgenstern, J. H., eds., *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, 220–229. ACM.

Munn, L. 2023. The uselessness of AI ethics. *AI and Ethics*, 3(3): 869–877.

Ngunjiri, F. W.; Gramby-Sobukwe, S.; and Williams-Gegner, K. 2012. Tempered radicals: Black women's leadership in the church and community. *Journal of Pan African Studies*, 5(2): 84–109.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

OECD.AI. 2021. Database of national AI policies.

Phadke, S.; Samory, M.; and Mitra, T. 2022. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 770–781.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Sapiezynski, P.; Kassarnig, V.; and Wilson, C. 2017. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of FATREC Workshop on Responsible Recommendation at ACM RecSys*.

Selbst, A.; and Powles, J. 2018. "Meaningful Information"' and the Right to Explanation. In Friedler, S. A.; and Wilson, C., eds., *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, 48. PMLR.

Stoyanovich, J.; and Howe, B. 2019. Nutritional Labels for Data and Models. *IEEE Data Eng. Bull.*, 42(3): 13–23.

The White House. 2023. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.

UNICRI. 2020. Towards Responsible Artificial Intelligence Innovation. *European Comission*.

Walton, S.; and Kirkwood, J. 2013. Tempered radicals! Ecopreneurs as change agents for sustainability–an exploratory study. *International Journal of Social Entrepreneurship and Innovation*, 2(5): 461–475.

Williams, R. 2021. How to train your robot: project-based ai and ethics education for middle school classrooms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 1382–1382.

Yang, Y.; Kandogan, E.; Li, Y.; Sen, P.; and Lasecki, W. S. 2019. A study on interaction in human-in-the-loop machine learning for text analytics. In *IUI Workshops*.