Making Transparency Influencers: A Case Study of an Educational Approach to Improve Responsible AI Practices in News and Media

Andrew Bell
alb9742@nyu.edu
New York University
New York, New York, USA

Julia Stoyanovich
New York University
New York, New York, USA

ABSTRACT

Concerns about the risks posed by artificial intelligence (AI) have resulted in growing interest in algorithmic transparency. While algorithmic transparency is well-studied, there is evidence that many organizations do not value implementing transparency. In this case study, we test a ground-up approach to ensuring better real-world algorithmic transparency by creating transparency influencers — motivated individuals within organizations who advocate for transparency. We held an interactive online workshop on algorithmic transparency and advocacy for 15 professionals from news, media, and journalism. We reflect on workshop design choices and presents insights from participant interviews. We found positive evidence for our approach: In the days following the workshop, three participants had done pro-transparency advocacy. Notably, one of them advocated for algorithmic transparency at an organization-wide AI strategy meeting. In the words of a participant: "if you are questioning whether or not you need to tell people [about AI], you need to tell people."

CCS CONCEPTS

• **Human-centered computing** → **Field studies**; • **Applied computing** → **Interactive learning environments**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Transparency, explainability, artificial intelligence, machine learning, tempered radicals

ACM Reference Format:

Andrew Bell and Julia Stoyanovich. 2024. Making Transparency Influencers: A Case Study of an Educational Approach to Improve Responsible AI Practices in News and Media. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613905.3637113

1 INTRODUCTION

There are widespread concerns about the significant risks posed by artificial intelligence (AI) systems used in the public and private sectors, particularly for marginalized or historically disadvantaged groups [22, 35, 42]. One major risk factor is the lack of transparency into complex AI systems that make — or assist in making — high-stakes decisions [25, 41]. Concerns over the lack of transparency of algorithmic systems have given way to a sub-field known as *Explainable Artificial Intelligence* (XAI), which is concerned with studying how well an AI system can be understood by humans [5]. While significant progress has been made in developing and evaluating methods for explaining complex AI systems by combining multi-disciplinary approaches from machine learning and human-computer interaction [1, 7, 9, 11, 21, 28, 39, 48], there is evidence that companies and organizations using AI do not value — or even know about — such methods [10, 19]. As a result, XAI is facing an existential challenge: how do we move from the research setting to *ensuring the actual implementation of transparent AI systems in the real world [4]?*

While national governments are natural candidates for addressing this challenge in specific high-stakes domains, the rapid development of AI technologies has greatly outpaced public oversight, creating an incomplete patchwork of laws and regulations [24]. To date, over 50 nations and intergovernmental organizations have published AI strategies, actions plans, policy papers or directives [46]. Unfortunately, all these documents have one major limitation: they are filled with uncertainty on how transparency should actually be implemented in a meaningful way [16, 24, 27].

The United States has chosen to rely (at least in part) on the private sector for helping ensure responsible and transparent AI practices. To this end, in a July 21, 2023 address, President Joseph Biden stated that he had received voluntary commitments to responsible AI from 7 different large tech companies. Skeptics raise the concern that these "voluntary commitements" *should not* replace regulation, given the possibility that large companies may abandon responsible AI values in favor of profit-motives. For example, amid layoffs that occurred in May 2023 that affected 10,000 people at Microsoft, the company felt comfortable laying off their entire AI ethics team.[1]

This case study explores an alternative pathway to ensuring safe, transparent AI: educate and create *transparency influencers*. We define transparency influencers as a subset of what Meyerson [31] called "tempered radicals," or committed employees who create institutional change over time (sometimes clandestinely), who are focused on algorithmic transparency. The approach of tempered radicals can be very effective: in one example, over a 30 year period, a Black senior executive hired an additional 3,500 women and minority members to work at a large West Coast bank as part of a covert effort to improve diversity within the company. We

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3637113

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

hypothesize that transparency influencers can affect similar, significant bottom-up organizational change towards better algorithmic transparency.

Study approach. As a precursor to this work, the authors created the *Algorithmic Transparency Playbook*, which is a stakeholder-first approach for creating transparency for an organization's algorithmic systems [5]. There are two associated artifacts—a 31-page *PDF document* (the full *Playbook*) and an open-sourced online course[2]—which were used to teach a course at the 2023 ACM CHI Conference on Human Factors in Computing Systems [5]. Based on the success of that course, we condensed the material into a single 2-hour workshop that could be evaluated as a case study for exploring two related research questions:

RQ1: How effective is an educational workshop based on *The Algorithmic Transparency Playbook* in increasing participants' algorithmic transparency literacy?

RQ2: Can the workshop increase participants' willingness to advocate for algorithmic transparency in their professional lives —*i.e.,* become *transparency influencers*?

Summary of findings. We held the workshop for 15 professionals in the news, media, and journalism industry. Through interviews with participants, we found that the workshop was effective both in teaching algorithmic transparency and in increasing willingness for advocacy. In the days following the workshop, three participants had taken some form of advocacy action. Most significantly, one participant attended an organization-wide meeting on AI strategy, and spoke-up on behalf of transparency using lessons from the workshop. They became a *transparency influencer.*

Participants also expressed that the workshop helped improve their understanding of algorithmic transparency, particularly by uncovering knowledge gaps. In interviews, three participants said the workshop helped them realize "they didn't know what they didn't know [about transparency]."

Overall, this case study represents a positive example of how an educational approach can be taken to create individual-driven, bottom-up change towards responsible AI practices at public and private intuitions [14]. We also learned several valuable lessons about designing such a course. For example, we believe that the workshop was effective, at least in part, because of our design decision to tailor the content to the learners (*i.e.,* professionals in the media and news industry), creating a deeper connection with the material. We offer this as a design recommendation for other researchers conducting similar workshops.

Advancing the state of practice. This case study exists at the intersection of XAI, HCI, and responsible AI literacy. A recent review found that the majority of work in responsible AI education is focused on teaching people how to program, leaving a gap for educational initiatives aimed at broader audiences [14]. We draw from HCI to design such an initiative teaching journalists about XAI. This study is also revelant to emerging work combining HCI and journalism [3, 13].

[2]https://dataresponsibly.github.io/algorithmic-transparency-playbook/

2 RELATED WORK

Organizational barriers to transparency. Organizations often forgo implementing transparent, responsible AI practices due to misaligned incentives. This is especially true for large, public companies, which are highly motivated by revenue and market fundamentalism [30]. In interviews with researchers, employees at one large technology company revealed that a majority of their day-to-day work was focused on profit-motivated tasks like launching products and increasing user engagement, rather than considering the ethics of those products [29, 30, 38]. Several members of ethics teams at large companies have stated that their roles exist merely as the result of external pressure, rather than due to pro-active, value-based decisions from company leadership [30]. In fact, the priorities of companies can be in *direct tension* with responsible AI. For example, the goal of optimizing user engagement can result in irresponsible outcomes like creating online radicalization pipelines [36].

Organizations face additional challenges beyond making and keeping internal commitments to responsible AI. One problem is that practitioners are at times unable to identify their companies' *specific goals* with respect to broad terms like AI ethics [30, 37]. For instance, while several companies make broad claims about valuing *AI transparency,*[3] individuals within the company may differ on agreeing to what extent transparency matters, who it is ultimately for, and how it should be implemented. A second problem is human "blind spots," or individuals who aren't aware of responsible AI practices, that exist in different parts of disconnected teams within an organization [20]. As a result, the responsibility of AI ethics often falls to single, motivated individuals, the so-called "ethics owners [30]." Overall, there is a need for developing organizational tactics and stakeholder management to ensure responsible AI practices within companies and organizations using AI [38].

Regulation. Unfortunately, despite some positive examples ensuring the implementation of transparent AI, regulation cannot be thought of as a silver-bullet. Existing and emerging laws, rules and directives contain loopholes that can easily be exploited. Further, the majority of AI directives and strategies created around the world lack specificity and means of enforcement [6, 33, 46]. For example, the European Union's General Data Protection Regulation (GDPR), enacted in 2016 and in effect since 2018, 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 GDPR being among the most expansive and the most robust data protection laws in existence, the right-to-explanation has yet to materialize into any meaningful benefit for citizens. The legal meaning and obligation of the text has been debated heavily by scholars, who are unsure under which circumstances it applies, what constitutes an explanation, and how the right is applicable to different AI systems [12, 15, 43].

Bottom-up change. Myerson coined the term "tempered radicals" to describe individuals who influence change from inside an organization slowly but surely over time [31]. Tempered radicals prefer to make bottom-up change, rather than relying on company

[3]One of Meta's five pillars of Responsible AI is "transparency and control." See https://ai.meta.com/responsible-ai/.

leadership or government regulation. There are numerous successful examples of tempered radicalism, especially when it comes to ethically motivated practices. They have been successful at promoting minority representation, inclusion, and sustainability in many different contexts, such as companies, universities, and religious organizations [18, 26, 32, 34, 47].

Tempered radicals are a natural approach for helping push organizations towards responsible AI practices. Interestingly, ground-level employees already seem to bear this responsibility: in interviews with researchers, employees at one large tech company said they often feel like it is *their* job to represent ethical technology values [38].

Importance of algorithmic transparency education. This case study seeks to test an educational approach for improving algorithmic transparency, based on evidence that education can increase pro-social behavior [2, 8, 45]. Our case study focuses on teaching individuals who use AI in their day-to-day work about the meaning and the need for algorithmic transparency. This kind of education is essential for responsible AI, where the goal is to ensure that the responsibility for the design, development, use, and oversight of AI systems can meaningfully rest with people. In this way, education about AI (and specifically about AI transparency) is both critically important and immediately relevant to the human-computer interaction community [40].

3 METHODS

We held a 2-hour educational workshop on algorithmic transparency and its advocacy for domain experts from the news, media, and journalism fields. The workshop was made up of 5 modules, one of which was a role-playing activity. We also administered a pre- and post-workshop survey, and conducted semi-structured interviews.

3.1 Recruitment and participation

We chose to the news, media, and journalism fields the focus of this case study for two reasons. First, the release of generative-AI tools like ChatGPT has made discussions around transparency and disclosure particularly salient for organizations in those spaces. As will be discussed later in this work, many media organizations are having existential conversations on how they will adapt to new AI technologies, and how to do so responsibly.

Second, we were able to partner with the AI & Local News project at the NYC Media Lab[4] to help lead recruitment. In total, 15 professionals in the news, media, and journalism industries attended the workshop. Many of the participants work with AI technologies, and their job titles included Chief Digital Officer, Audience Development Editor, Managing Editor, Data Journalist, and Newsroom Developer. There was also one Postdoctoral Researcher.[5]

3.2 Workshop structure and design

We created the workshop by condensing material from the *Algorithmic Transparency Playbook Course*, a course that was taught by the authors at the 2023 ACM CHI Conference [5], to make it more accessible to a lay audience. The workshop was adapted to include examples that are directly relevant to the news, media, and journalism fields. The workshop was taught by the lead author via *Zoom*, and was made up of 5 modules, summarized in Table 1. The slides used during the workshop can be found in the footnote below.[6] In addition to the modules, we reserved 10 minutes each for the pre- and post-workshop surveys (described in Section 3.3), and an additional 20 minutes of audience Q&A and open discussion.

Design considerations. After conducting the course at the 2023 CHI conference, we spoke informally with participants to gather feedback. Past participants gave two main suggestions: first, that we expand the *All About Transparency* module by deepening the content, and, second, that we add more emphasis on the tools available for transparency. With respect to the former, additional content and time were given to the *All About Transparency* module, and, with respect to the latter, the *Transparency Tools* module was added.

A primary design choice we made for the workshop was to specifically tailor the content to our audience—*i.e.,* to the use of algorithms in news, media, and journalism. This manifested in two ways. First, throughout the course, we added case studies and examples directly drawn from the use of AI in journalism. For example, we added a discussion around recent news that the media company *CNET* had been using AI to generate articles on its site, many of which contained errors.[7] We discussed what went wrong and how *CNET* could have benefited from a transparent AI strategy. Second, the breakout room activity was created to be about a fictional company in the news and media space.

Breakout room activity. To add engagement and deepen participants' connection with the course content, we included a 15-minute activity using the *Zoom* breakout room feature. The purpose of the activity was to improve participant's ability to advocate for transparency by demonstrating the type of tensions that emerge when organizations begin considering transparency. For example, we wanted participants to be aware that many managers object to adding transparency to algorithms to save costs or protect Intellectual Property — and also to be aware how they can rebut those arguments. In total, we had 4 breakout rooms with 3-4 participants and a moderator in each room. The rooms were moderated by the authors and two colleagues.

Participants were asked to imagine that they were managers at a fictional social news company that recently began using an AI content moderation tool. Half of the participants were asked to role-play *skeptical managers*, who were against disclosing the use of the AI tool, while the other half were asked to role-play *pro-transparency* managers. In each breakout room, participants had access to a Google Jamboard that allowed them to make arguments for different stakeholders (*e.g.,* readers, moderators, developers, etc., as discussed in the workshop content) according to their role. An example of a completed activity can be seen in Figure 1.

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

[5]While the workshop was open to the public, participation in this IRB-approved case study was restricted to (1) adults over 18 years old, (2) *not* full-time students, (3) individuals attending the workshop from inside the US.

[6]https://docs.google.com/presentation/d/1M7Xgfp86PBP1g_8KrmlogyIko975IXR2/edit?usp=sharing&ouid=100559290438924098736&rtpof=true&sd=true

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

Table 1: Modules covered in the workshop.

Module	Topics	Time (*mins*)
All About Transparency	Defining algorithmic transparency, types of transparency, stakeholders and their goals	20
Transparency Tools	Transparency labels and model cards, feature importance and Shapley values, explainer 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 Room 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 Influencer	Common objections to transparency (*i.e.,* "transparency means more costs," "transparency means sacrificing privacy") and how to rebut them	10

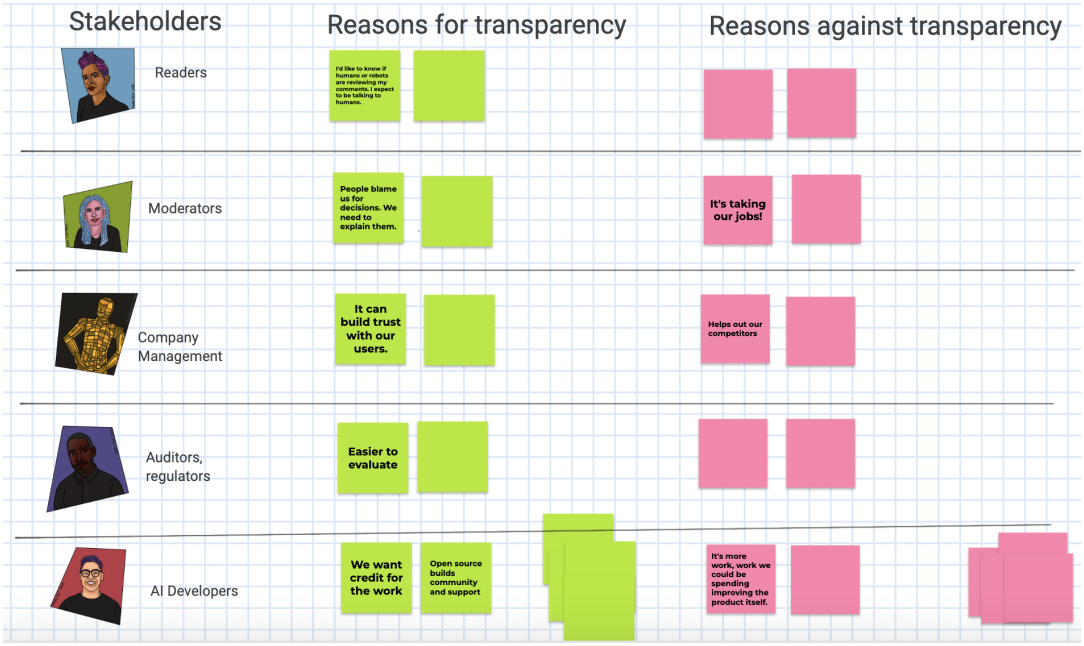

Figure 1: Completed breakout room activity (described in Section 3.2) from one of the four breakout rooms. From top-to-bottom, the green cards (reasons *for* transparency) read: "I'd like to know if humans or robots are reviewing my comments. I expect to be talking to humans.", "People blame us for decisions. We need to explain them.", "It can build trust with our users.", "Easier to evaluate", "We want credit for the work", "Open source builds community and support." From top-to-bottom, the red cards (reasons *against* transparency) read: "It's taking our jobs!", "Helps out our competitors", "It's more work, work we could be spending improving the product itself." Illustrations by Falaah Arif Khaan.

3.3 Data collection and analysis

Pre- and post-workshop surveys. We conducted a pre-workshop survey to assess participants' self-reported baseline knowledge of algorithmic transparency and willingness for advocacy (8 questions), and a post-workshop survey to evaluate the impact of the workshop (18 questions).[8] Themes measured in the survey include general AI sentiment, perceived transparency literacy, and willingness to engage in transparency advocacy. The intended use of the surveys was to quantitatively assess the workshop's impact, as well as collect free-responses from participants on questions like "How likely are you to advocate for algorithmic transparency in your work? Please tell us why." In total, 15 participants completed the pre-workshop survey, and 7 of them also completed the post-workshop survey. Due to the small sample size, we do not draw

[8]The full surveys have been included in the Supplementary Material.

any quantitative conclusions from this data, but report qualitative findings. We discuss participant drop-off in Section 5.

Interviews. In the days following the workshop, we conducted semi-structured interviews[9] with four participants, whose domain and expertise are found in Table 2[10]—note henceforth we also refer these participants as domain experts. This was a particularly interesting group to speak because they represent a broad range of domains within the news, media, and journalism spaces, and most are practitioners with lived knowledge of both journalism and AI. In our interviews, we sought to understand how the participants felt the workshop met its goals in relation to *RQ1* and *RQ2*. Because we were asking participants to speak candidly about experiences at work (and incidentally their employer), we did not record the interviews to make participants more comfortable; instead, we took notes throughout and wrote down quotes relevant to our research work.

Analysis. Our analysis included interview notes and answers from free-response questions on the post-workshop survey from the interviewed domain experts. We began by carefully reading the interview notes and post-workshop survey responses to identify and code the salient and recurring themes. These codes were then grouped into the five major themes we report below, and supported with relevant quotations. Importantly, our major themes were created through the lens of the two research questions: *RQ1* How effective is an educational workshop in increasing participants' algorithmic transparency literacy? and *RQ2* Can the workshop increase participants' willingness to advocate for algorithmic transparency in their professional lives, becoming *transparency influencers*?

4 RESULTS

4.1 Transparency literacy (RQ1)

Frequent use of internally developed and procured algorithmic tools. All participants noted that they had frequent or almost daily contact with AI in their jobs. Overall, they used a wide range of different algorithmic tools. The tools included generative AI for creating story headlines, systems for suggesting news story topics to editors and for recommending news videos to online readers, and tools for A/B testing story headlines. One common theme regarding these tools is that many of them were acquired from third party vendors. Somewhat surprisingly, all participants, even those who work at companies that built the majority of their algorithmic systems in-house, mentioned using at least one procured algorithmic tool.

Uncovering knowledge gaps. The general sentiment of participants was that the workshop was useful, and that they felt that it improved their understanding of algorithmic transparency. Each participant mentioned different aspects of the course that they found as the most salient, including learning about the different levels of transparency (P1), the existing tools for transparency (P2), and the stakeholder identification (P3). Interestingly, several participants found that the major strength of the workshop was uncovering knowledge gaps. In fact, P2, P3, and P4 each said that the workshop helped them realize that "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."

4.2 Transparency advocacy (RQ2)

Taking action. Domain experts P2, P3, and P4 of stated they had taken some kind of transparency advocacy action in the days following the workshop. Two participants mentioned they have begun having more conversations with colleagues and peers about algorithmic transparency. P2 said they "already used the Algorithmic Transparency Playbook" by mentioning it to colleagues, and P4 said "I've probably had 5 conversations of AI transparency compared to close to 0 [before the workshop]." P4 also mentioned "I was [texting] co-workers about steps we can take during the [workshop]. It's important to disclose how our AI moderation works – and any of the other tools and ideas we have in mind." P3 stated how they have already begun implementing some elements of the workshop into their workflow. Specifically, "the breakdown of stakeholders was put into use immediately."

Notably, P4 took part in internal discussions at their organization *directly related to the use of the algorithmic tools* in the days following the workshop, where they fully stepped into the transparency influencer role. They described their experience in the following way: "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." They "brought up a lot of the points [from the workshop]" and made several observations about organizational challenges to transparency (mentioned below), some of which were discussed in the workshop.

The interviews demonstrate that the advocacy was, at least in part, the result of the workshop. A popular sentiment was that the workshop was effective in providing better resources for advocacy. P2 phrased it in this way: "I always would've advocated for transparency anyway ... ", but the workshop improved their potential for transparency advocacy "... from being made aware of different types of resources related to transparency." P3 expressed a similar sentiment, saying the workshop was helpful "if only to raise my knowledge [of algorithmic transparency] and attention [to it]."

P3 and P4 also commented on their future plans for transparency advocacy, which were, again, based on the resources covered in the workshop. The former said "If we ever go down the road of building a model it feels like [model cards] are something we should probably do". They said the same was true for vendors: "If we ever procure a model we should check model cards." The latter expressed a similar sentiment, saying that they would push for model cards for some of the AI tools their organization is currently using, like the video recommendation system.

Organizational challenges. P1 and P4 discussed that their organization recently held internal meetings to discuss AI strategies and create a "Code of Conduct" for its use — a clear indication that the news and media fields are responding to the rapid proliferation of AI tools like recommender systems and generative AI. Notably, both participants pointed out that this is not a smooth transition, particularly for older, legacy employees. P1 stated that discussions

[9]The full interview protocol has been included in the Supplementary Materials.
[10]Job titles and employer names were omitted to protect the anonymity of participants.

Table 2: Participants domain and expertise

#	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

around the use of generative AI for creating story headlines has "ruffled a lot of feathers" and has seemingly divided the organization into two schools of thought: those who are *pro* new AI tools, and those who are against their use. One positive that emerged from internal discussions was that "transparency is key" was unanimously agreed upon as being included in their organization's Code of Conduct regarding the use of AI.

P4, who works primarily with local TV news, referred to a so-called "TV newsroom mindset" that emerged as a barrier to algorithmic transparency during organizational AI strategy discussions. They further explained that "it's not in [TV newsroom peoples'] nature to be very disclosing." Significantly, they noted that in organizational discussions about AI transparency, they heard the same types of objections to transparency that were mentioned in the workshop.

When is transparency necessary? One surprising theme that emerged from interviews was the large variation on when each participant (or members of their organization) felt it was necessary to be transparent about the use of AI. P1, P3, and P4 all agreed that it does not seem necessary to disclose the use AI for generating — or supporting the writing of — article headlines. P3 specifically mentioned, "When headlines are AI-assisted... we don't expose that to anyone." They went on to say that "our users might not actually care." This also included *not* disclosing the use of algorithmic tools used to A/B test headlines for different groups of users. P4 said that when they directly asked one of their colleagues "do we need to disclose [the use of AI]?", their colleague responded with "it's too late for that now isn't it?" This indicates that some believe that since the "genie is already out of the bottle" it isn't worth pursuing transparency.

P1 also defended forgoing disclosing the use of AI to generate news articles that are normally "templated" anyway, like business press releases. They used the analogy of politicians not crediting every sentence of their speeches to different speech writers. Notably, this participant believed it was wrong to use AI to generate an entire news article, and they even questioned whether generating text with AI and editing it was actually more efficient than simply writing the article oneself. A similar view was expressed by P4.

Another interesting finding within this theme was the mentioning of "unwritten rules" for transparency. P1 mentioned that their organization had adopted the phrase, "if a journalist could do it, a journalist should do it." They also gave the following guideline: "if you are questioning whether or not you need to tell people [about AI], you need to tell people."

5 DISCUSSION

Types of advocacy. We found positive evidence that the workshop was successful in improving advocacy for algorithmic transparency among domain experts working in the news, media, and journalism fields. We identified three advocacy approaches taken by domain experts which we categorize as *converse, implement, and influence*.

Regarding *conversational advocacy*, two participants mentioned that after the workshop they had significantly increased the amount of conversations they had about algorithmic transparency in their everyday life. These conversations took place both with colleagues and peers. While conversations about transparency may not *directly* affect organizational change, they play a role in increasing awareness around algorithmic transparency, which in-and-of itself can result in significant change in peoples' behaviors over the long term [23].

We define *implementational advocacy* as a type of advocacy where individuals implement algorithmic transparency directly into their work, without necessarily consulting managers or speaking with colleagues. This type of advocacy is focused on narrow immediate change, rather than on creating broader cultural shifts within an organization. As a prime example of implementational advocacy, one participant (who manages a small team of software developers) said they had already began integrating material learned from the workshop about stakeholder identification into their team's workflow. From the viewpoint of the authors, this type of advocacy is critical for creating bottom-up change within organizations, and inline with the at-times clandestine actions of tempered radicals [32].

Influencing is the type of advocacy that was most explicitly discussed in the workshop, and is defined by individuals taking behaviors to affect cultural change towards algorithmic transparency within their organization. Significantly, P4 immediately began influencing in the days following the workshop, and spoke up about algorithmic transparency in an organization-wide meeting on AI strategy. They brought up arguments for disclosure and transparency learned in the breakout room activity, and were even met with some of the same anticipated negative responses. Overall, P4 walked away from the meeting feeling optimistic and hopeful that their company would start taking steps towards the more transparent use of algorithmic systems. This is perhaps the most direct impact of this case study: by inviting one person to think more deeply about algorithmic transparency and providing them with basic tools for advocacy, we hope to have caused a downstream impact of a medium-sized US media company adopting more responsible AI practices.

Identification of further organizational challenges to transparency. Interviews with participants showed confusion over when it is actually necessary to be transparent about the use of AI. For example, several domain experts felt that using generative AI to create news story headlines or to generate "templated" news articles like business press releases does not necessarily warrant algorithmic transparency. The key takeaway, with implications for the design

of future educational interventions, is that it is important to reinforce the connection between (1) whether and when transparency is warranted and (2) the degree of automation and the potential harms (*i.e.,* with the level of risk to stakeholders due to the use of automation) [17, 44]. In other words, it is important to ground the questions about algorithmic transparency in a deeper understanding of responsible AI and technology ethics.

Best practices for teaching about algorithmic transparency. It is our belief that participants felt positively about the workshop because it was designed specifically for audiences in news and media. For example, P4 mentioned in that their organization was actually using AI in a near-identical way to the fictional scenario created for the breakout room activity. P1, P3, and P4 were also already familiar with the *CNET* news story,[11] which likely helped deepen their connection with the course material. Overall, we plan to replicate this design consideration in future iterations of this workshops given to experts in different domains. It is also the view of the authors that is a generalizable lesson of this case study: we recommend others creating responsible AI workshops and courses aimed at professionals should tailor the content to their audience.

Regarding the breakout room activity, participants had mixed opinions. Coincidentally, we conducted interviews with both participants in the breakout room shown in Figure 1, *i.e.,* P3 and P4. While P3 found the breakout room activity awkward and clunky, P4 felt connected to the activity because it centered around a use of AI familiar to them in their work. We hypothesize that this contradiction was also due to differences in the temperament between P3 and P4, and so we feel its best not to draw conclusions about the utility of the activity without further exploration.

Limitations and other lessons. A challenge we faced in this case-study was participant drop-off. While we started the session with 15 participants, only 7 attended the entire two hour workshop. This means we likely received positively biased feedback, and we may be missing important details on how to improve the workshop design to reach more participants. Additionally, because of the drop-off, we were unable to perform our intended quantitative analyses using the pre- and post-workshop surveys. A majority of the drop-off occurred as we transitioned to the lecture portion to the breakout room activity, which was 45 minutes into the workshop. In the future we should explore *why* participants left the workshop, but we hypothesize it was because participants felt they had gleaned enough value from the lectures and were not interested in participating in the activity. We also plan to run future iterations of this workshop with professionals from various sectors, and we feel it is important to hold future sessions in-person to help mitigate drop-off.

A limitation of this case study is that it is, indeed, a case study, and so it leaves open questions about the generalazibility of our approach and findings. It is unlikely we have reached data saturation, even within the news and media field. While we observed promising and hopeful results for increasing individuals' willingness to advocate for algorithmic transparency, much additional work is needed to explore the many different points-of-view that are likely missing from this work. Further, this work should be adapted and evaluated to different domains (*e.g.,* technology, healthcare) and cultural contexts. For example, one could imagine that in contexts where power structures are ordered in a more hierarchical way, it may be more difficult to realistically affect bottom-up change.

[11] https://www.theverge.com/2023/6/6/23750761/cnet-ai-generated-stories-policy-update

6 CONCLUSION AND SOCIAL IMPACT

Overall, this case study represents a positive example of affecting ground-up change in organizations towards responsible AI by providing the proper education and resources to individuals. It is our hope that this case study will exist as part of a broader body of work helping to ensure responsible AI practices studied in research settings are transferred into the real-world use of technologies, especially in high-stakes domains that impact society at large.

ACKNOWLEDGMENTS

This work was supported in part by NSF awards 1916505, 1922658, 2312930, 2326193, and by the NSF GRFP (DGE-2234660). The authors would like to thank Matt MacVey from AI & Locals News for helping recruit participants, and organize and run the workshop. We would also like to thank Falaah Arif Khan and Awais Hameed Khan (no relation) for helping moderate the breakout room activity.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 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.

[2] Ali Ahmed. 2008. Can education affect pro-social behavior? Cops, economists and humanists in social dilemmas. *International Journal of Social Economics* 35, 4 (2008), 298–307.

[3] Daniel Angus and Skye Doherty. 2015. Journalism meets interaction design: An interdisciplinary undergraduate teaching initiative. *Journalism & Mass Communication Educator* 70, 1 (2015), 44–57.

[4] Lex Beattie, Dan Taber, and Henriette Cramer. 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.

[5] Andrew Bell, Oded Nov, and Julia Stoyanovich. 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.

[6] Andrew Bell, Oded Nov, and Julia Stoyanovich. 2023. Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy* 5 (2023), e12.

[7] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 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.

[8] Cornelia Betsch, Robert Böhm, Lars Korn, and Cindy Holtmann. 2017. On the benefits of explaining herd immunity in vaccine advocacy. *Nature human behaviour* 1, 3 (2017), 0056.

[9] Ian Covert, Scott M. Lundberg, and Su-In Lee. 2020. DBLP:journals/corr/abs-2004-00668 Feature Contributions Through Additive Importance Measures. *CoRR* abs/2004.00668 (2020). arXiv:2004.00668 https://arxiv.org/abs/2004.00668

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

[11] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*. IEEE, 598–617.

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

[13] Skye Doherty, Jane Johnston, and Ben Matthews. 2023. Materialising New Forms of Journalism: A Process Model. *Digital Journalism* 11, 3 (2023), 504–518.

[14] Daniel Dominguez Figaredo and Julia Stoyanovich. 2023. Responsible AI literacy: A stakeholder-first approach. *Big Data and Society* (2023). forthcoming.

[15] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[16] Urs Gasser and Virgílio A. F. Almeida. 2017. A Layered Model for AI Governance. *IEEE Internet Comput.* 21, 6 (2017), 58–62. https://doi.org/10.1109/MIC.2017.4180835

[17] Government of Canada. 2019. Directive on Automated Decision-Making. (2019). https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

[18] Chris Griffiths, Edwina Pio, and Peter McGhee. 2022. Tempered radicals in manufacturing: Invisible champions of inclusion. *Journal of Management & Organization* (2022), 1–22.

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

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

[21] Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: the system causability scale (SCS). *KI-Künstliche Intelligenz* (2020), 1–6.

[22] Qian Hu and Huzefa Rangwala. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. https://eric.ed.gov/?id=ED608050

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

[24] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. Artificial Intelligence: the global landscape of ethics guidelines. *CoRR* abs/1906.11668 (2019). arXiv:1906.11668 http://arxiv.org/abs/1906.11668

[25] Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance* 72, 3 (2017), 967–998.

[26] Gill Kirton, Anne-Marie Greene, and Deborah Dean. 2007. British diversity professionals as change agents–radicals, tempered radicals or liberal reformers? *The International Journal of Human Resource Management* 18, 11 (2007), 1979–1994.

[27] Michele Loi and Matthias Spielkamp. 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.

[28] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[29] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

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

[31] DE Meyerson. 2003. Tempered Radicals: how everyday leaders inspire change at work Boston. *MA: Harvard Business School Press, xi–xii* 41 (2003), 59–60.

[32] Debra Meyerson and Megan Tompkins. 2007. Tempered radicals as institutional change agents: The case of advancing gender equity at the University of Michigan. *Harv. JL & Gender* 30 (2007), 303.

[33] Luke Munn. 2023. The uselessness of AI ethics. *AI and Ethics* 3, 3 (2023), 869–877.

[34] Faith Wambura Ngunjiri, Sharon Gramby-Sobukwe, and Kimberly Williams-Gegner. 2012. Tempered radicals: Black women's leadership in the church and community. *Journal of Pan African Studies* 5, 2 (2012), 84–109.

[35] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[36] Shruti Phadke, Mattia Samory, and Tanushree Mitra. 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*, Vol. 16. 770–781.

[37] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[38] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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.

[40] Lionel P Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. 2020. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction* 35, 5-6 (2020), 545–575.

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

[42] Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. 2017. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of FATREC Workshop on Responsible Recommendation at ACM RecSys*. https://api.semanticscholar.org/CorpusID:573639

[43] Andrew Selbst and Julia Powles. 2018. "Meaningful Information" and the Right to Explanation. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 48. http://proceedings.mlr.press/v81/selbst18a.html

[44] Andrew D Selbst. 2017. Disparate impact in big data policing. *Ga. L. Rev.* 52 (2017), 109.

[45] Helen Street, David Hoppe, David Kingsbury, and Tony Ma. 2004. The Game Factory: Using Cooperative Games to Promote Pro-social Behaviour Among Children. *Australian journal of educational & developmental Psychology* 4 (2004), 97–109.

[46] UNICRI. 2020. Towards Responsible Artificial Intelligence Innovation. *European Comission* (2020). http://www.unicri.it/index.php/topics/ai_robotics

[47] Sara Walton and Jodyanne Kirkwood. 2013. Tempered radicals! Ecopreneurs as change agents for sustainability–an exploratory study. *International Journal of Social Entrepreneurship and Innovation* 2, 5 (2013), 461–475.

[48] Yiwei Yang, Eser Kandogan, Yunyao Li, Prithviraj Sen, and Walter S Lasecki. 2019. A study on interaction in human-in-the-loop machine learning for text analytics. In *IUI Workshops*.