Patterns

CellPress
OPEN ACCESS

Perspective

Explainability pitfalls: Beyond dark patterns in explainable AI

Upol Ehsan[1,*] and Mark O. Riedl[1]
[1]Georgia Institute of Technology, Atlanta GA, USA
*Correspondence: ehsanu@gatech.edu
https://doi.org/10.1016/j.patter.2024.100971

THE BIGGER PICTURE Explainability pitfalls (EPs) are negative effects of AI systems that arise without the intention to deceive end users. Defining EPs and differentiating them from dark patterns, which, in contrast, are negative features intentionally designed to manipulate and deceive users, sets the stage for designing and implementing strategies to safely adopt explainable AI technologies.

SUMMARY

To make explainable artificial intelligence (XAI) systems trustworthy, understanding harmful effects is important. In this paper, we address an important yet unarticulated type of negative effect in XAI. We introduce explainability pitfalls (EPs), unanticipated negative downstream effects from AI explanations manifesting even when there is no intention to manipulate users. EPs are different from dark patterns, which are intentionally deceptive practices. We articulate the concept of EPs by demarcating it from dark patterns and highlighting the challenges arising from uncertainties around pitfalls. We situate and operationalize the concept using a case study that showcases how, despite best intentions, unsuspecting negative effects, such as unwarranted trust in numerical explanations, can emerge. We propose proactive and preventative strategies to address EPs at three interconnected levels: research, design, and organizational. We discuss design and societal implications around reframing AI adoption, recalibrating stakeholder empowerment, and resisting the "move fast and break things" mindset.

INTRODUCTION

Engendering appropriate reliance on artificial intelligence (AI) is a major goal of explainable AI (XAI), a research area that aims to provide human-understandable justifications for the system's behavior.[1,2] This is vital because AI systems increasingly power decision-making in high-stakes domains like healthcare[3–6] and criminal justice.[7–9] However, adding AI explainability does not always guarantee positive effects; there can also be detrimental ones. Thus, we need to properly understand detrimental effects of AI explanations—intentional and unintentional ones.

Despite commendable progress in XAI, emerging work has highlighted detrimental effects of explanations.[10–14] For instance, deceptive practices can be intentionally used to design placebic explanations (lacking justificatory content) and engender trust in AI systems.[15–17] However, not all detrimental effects are intentional. Despite the designer's best intentions, AI explanations can have unintentional negative effects on the end user. In such cases, what can we do? What do we call these unintentional negative effects in XAI? How might we address them? Unintentional negative effects of AI explanations are underexplored both conceptually and practically in XAI.

To address this problem, we introduce and operationalize the concept of Explainability Pitfalls (EPs), which are unanticipated negative downstream effects from adding AI explanations that emerge even when there is no intention to manipulate anyone. EPs can cause users to act against their own self-interests, align their decisions with a third party, or unsuspectingly exploit their cognitive heuristics. Note that explanations themselves are not the pitfalls. When AI explanations are added to enhance explainability, it changes the terrain of the XAI design space. This changed terrain can have unsuspecting pitfalls even when there are the best of intentions. Examples of these downstream negative effects include user perceptions like misplaced trust, over-estimating the AI's capabilities, and over-reliance on certain forms of explanations.

EPs are different from dark patterns (DPs), which are a set of deceptive practices intentionally and "carefully crafted to trick users into doing things … They do not have the user's interests in mind."[18] Emerging work showcases how DPs in XAI can create a false sense of security to trick users to over-trust systems.[19] With roots in user experience design (UX design),[20] DPs have been explored in multiple contexts, like games[21] and medical records.[22] A major difference between EPs and DPs is around intentionality—DPs have bad-faith actors intentionally trying to trick users. The intentionality can not only manifest from deceptive practices but can also emerge from perverse incentives, such as prioritizing product growth and adoption at the expense of responsible stewardship. Despite being different, EPs and DPs are related; you can turn pitfalls into DPs by intentionally setting the traps (to trick the user).

We use the metaphor of "pitfalls" to signal unsuspected or hidden difficulties or dangers that are not easily recognized. Developers of AI and XAI systems might encounter these difficulties due to lack of information, understanding, or oversight (or a combination). Without the awareness of how to identify and avoid pitfalls, there are increased risks for end users who may never be aware of being affected. Moreover, EPs may only be symptomatic, thus detectable, after users interact with the explanations, and there is a misalignment between their behavior and designers' expectations. This implies a high bar of accountability for designers to be "pitfall aware" when designing XAI systems (at first glance, the lack of intentionality in EPs may be misinterpreted as an exonerating force). Taking the metaphor of pitfalls further, we can envision navigation strategies to detect and avoid them. If we are pitfall aware in our navigation of the design space and aware of the possibility of unanticipated and unintentional negative effects, we can proactively build resilience against the pitfalls.

The motivation for conceptualizing EPs is not purely theoretical. It is practically motivated: we cannot mitigate the ill effects of something without detecting it. We cannot detect it without conceptualizing it. The conceptualization makes things tractable, thereby actionable. The motivation is also empirically situated in emerging work (e.g., Ehsan et al.[23]) that showcases how, despite no intentions of deception, unsuspecting negative effects can emerge from AI explanations. While unintentional negative effects of technology are not new,[24,25] the introduction of EPs into the discourse in XAI provides essential conceptual vocabulary and practical strategies that allow us to address unintentional detrimental effects of XAI. The idea of EPs addresses a timely challenge in the XAI community because it affords us to not have to cluster all negative XAI effects as "DPs" or face the oxymoronic implications of unintentional DPs. Existing at the translational space between HCI and XAI, our contributions are fourfold.

(1) We highlight an intellectual blind spot in XAI by bringing awareness to previously unarticulated downstream negative effects of AI explanations that are unintentional.
(2) We address this blind spot by conceptually introducing and practically operationalizing the notion of EPs. We propose pitfall-aware navigation strategies at the research, design, and organizational levels to build resiliency.
(3) Reflecting on our strategies, we share implications around reforming AI adoption, recalibrating stakeholder empowerment, and resisting the "move fast and break things" mindset.

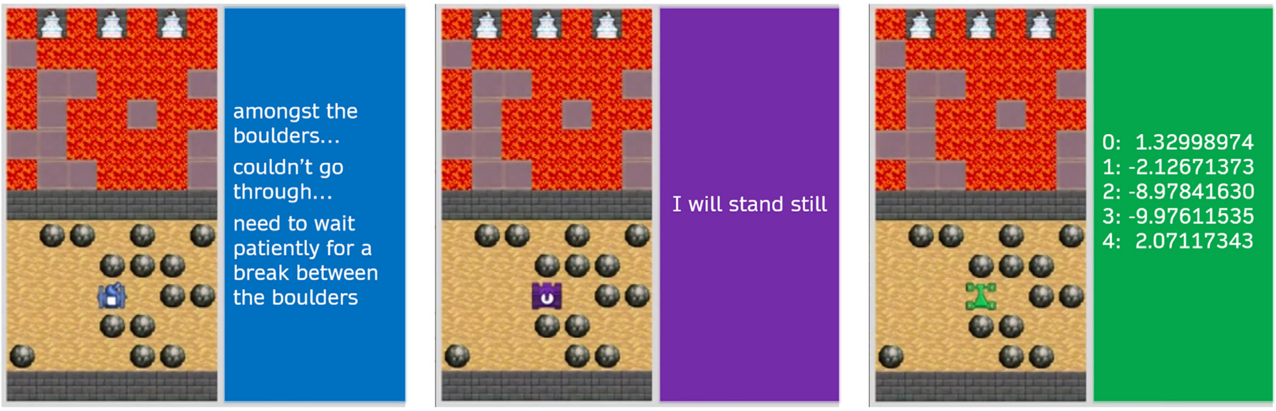
In the rest of this paper, we operationalize the concept of EPs through a case study where EPs manifested and were discovered through qualitative analysis of perceptions to AI user explanations. Reflecting on the findings, we then propose formative strategies to address EPs, followed by sharing design and societal implications. This paper is not a full treatise of EPs; rather, it takes a foundational step toward operationalizing the notion conceptually and practically.

CASE STUDY: SITUATING EPs

For the case study to situate and operationalize EPs, we choose Ehsan et al.'s work[23] for three main reasons. First, it provides an example where their qualitative analysis revealed unexpected negative effects of a particular type of AI explanation on end users. It showcases how, even when there is no intention to deceive anyone, negative effects can emerge in "surprising, non-intuitive" and "unanticipated ways."[23] Second, it highlights how "merely producing well-designed explanations is not enough to guarantee people would perceive them as designed,"[23] which can point to origins of EPs. Third, it provides the what, how, and why of the potential negative effects, which provides a deeper understanding of EPs. We highlight the most relevant parts of their findings, reframing them through the lens of EPs.

This case study investigated how two different groups—people with and without a background in AI—perceive different types of AI explanations. It probed for user perceptions on three types of AI-generated explanations: (1) natural language with justification, (rationale-generating robot, blue in Figure 1), (2) natural language without justification (action-declaring robot, purple in Figure 1), and (3) numbers that provide uncontextualized transparency into agent's actions (numerical reasoning robot, green in Figure 1). For our purposes, to situate the notion of EPs, we only need to focus on how and why both groups reacted to the numbers from the numerical reasoning (NR) robot (#3).

In the study, participants provided both qualitative and quantitative perceptions after watching videos of three robots (AI agents) using reinforcement learning to navigate a sequential decision-making environment: a field of rolling boulders and flowing lava to retrieve essential food supplies for trapped space explorers (more details in Ehsan et al.[23]). Participants were asked to imagine themselves as space explorers faced with a search-and-rescue mission where they are trapped inside a protective dome and must rely on an autonomous robot to reach a remote supply depot. They cannot control the autonomous robots, which must navigate a treacherous field of boulders and a river of lava to retrieve the supplies (Figure 1). Participants could only see a non-interactive video stream of their activities through their "space visors." This non-interactiveness aimed to heighten their sense of lack of control (and thereby reliance on the robots). Since the robots took identical actions during the task, participants were asked to pay special attention to the only differentiating factor, which in this case is the way each robot explained its actions by "thinking out loud." The NR robot (the relevant one for this paper) "thinks out loud" by simply outputting the numerical Q values for the current state (Figure 1). Q values[26] can provide some transparency into the agent's beliefs about each action's relative utility ("quality") but do not contain information on "why" one action has a higher utility than another. Participants, by design, had no idea about how each robot generated its expressions; for instance, participants were not informed that NR's numbers are Q values. To them, it was just another different way of explaining the actions. NR was meant to be used as a comparative baseline. Both groups experienced the same set of videos in a 2 (groups) by 3 (types of explanations) factorial design.



amongst the boulders...
couldn't go through...
need to wait patiently for a break between the boulders

I will stand still

```
0:  1.32998974
1: -2.12671373
2: -8.97841630
3: -9.97611535
4:  2.07117343
```

Figure 1. Screenshot from[23] depicting three robots navigating the task environment and explaining their actions

From left to right, the robots and their colors are as follows: rationale generation (blue), action declaring (purple), and NR (green). In the screenshot, each robot is taking the same action, but the robots are explaining it differently. The focus in this paper is the NR robot, used as a foil against two other robots, with natural language explanation strategies.

In terms of findings, both groups had unwarranted faith in numbers but demonstrate it to different extents and for different reasons. See Ehsan et al.[23] for a detailed description of data collection and analysis. To illustrate the reasons behind the effect, the authors use cognitive heuristics (rules of thumb or mental shortcuts), which leads to biases and error if applied inappropriately.[27–29] They elucidate how different heuristics of people with different AI backgrounds can lead to undesired outcomes.

For the AI group, the presence of numbers triggered heuristic reasoning that associated mathematical representations with a logical algorithmic thinking process even though they could "'not fully understand the logic behind [NR's] decision making.' (A43)."[23] Contradictorily, they voted the least understandable robot (NR) to be more intelligent! To them, "'Math [… had] an aura of intelligence', which 'made the NR robot feel smarter' (A16, A75)."[23] Not only did they over-value numerical representation, but this group also viewed numbers as potentially actionable even when their meaning was unclear. Actionability refers to what one might do with the information in terms of diagnosis or predicting future behavior. Many highlighted that even if they could not "make sense of numbers right now, [they] should be able to act on them in the future (A39)."[23] We should critically ask: how actionable are NR's numbers? As highlighted before, Q values cannot indicate the "why" behind the decision. These numbers do not allow much actionability beyond an assessment of the quality of the actions available. That is, despite a desire to correct errors, having the numbers on hand would not help them determine the cause of failures that could be corrected. Instead, they engendered over-trust and misplaced assessment of the robot's intelligence.

For the non-AI group, the very inability to understand complex numbers triggered heuristic reasoning that NR must be intelligent. NR's "'language of numbers', *because* of its 'cryptic incomprehensibility', signaled intelligence (NA6, NA1)."[23] Given that NR's numbers were inaccessible, the non-AI group never posited actionability. They did not exhibit any intent to fix the robot or use the numbers to diagnose its behavior. Note that this line of reasoning is different from the AI group's heuristic, which posited a future actionability (despite lack of understandability).

The authors underscore the unexpected nature of these negative effects. For either group, they did not anticipate that unlabeled, seemingly incomprehensible numbers would increase trust and assessment of the agent's intelligence. Moreover, they presented "the Q-values in good-faith."[23] What if these numbers were manipulated? Imagine bad-faith actions exploiting these pitfalls to manifest DPs; for instance, an XAI system that explains in (manipulated) numbers (to induce trust). Given the heuristic faith in numbers, it can induce over-trust and incorrect perceptions of the system. Operationalizing the concept of EPs, this case study showcases how unanticipated negative effects (i.e., over-reliance on numbers) can arise even when there are the best of design intentions.

NAVIGATION STRATEGIES FOR EPs

Given their nature, it is unlikely that we can completely eliminate EPs. Recall the uncertainty around EPs, just because they exist does not guarantee the downstream harms will happen. We do not yet know enough to predict when, how, and why a given AI explanation will trigger unanticipated negative downstream effects. While we are vulnerable to pitfalls, and there is no silver bullet solution, we can increase our resiliency by adopting pitfall-aware strategies—proactive and preventative measures that help us understand where pitfalls tend to be found, how they work, and how they can be avoided. To expand on the pitfall metaphor, we want to probe the areas ("grounds") of the explanation design space (where pitfalls are likely to occur) to increase our likelihood of being on sturdy ground. We can be pitfall aware in our approaches at three interconnected levels: research, design, and organization.

At the research level, we need to conduct more situated and empirically diverse human-centered research to obtain a refined understanding of the stakeholders[1,30] as well as the dimensions of explanations that affect different stakeholders in XAI.[23] This is because pitfalls become symptomatic—and thereby identifiable—when downstream effects (like user perceptions on AI explanations) manifest. For instance, the case study reviewed in the prior section revealed that different AI backgrounds in end users can trigger the same pitfall (over-trust in numbers) but for different heuristic reasons.

Without running the study, these pitfalls could not have been identified. Fortunately, the case study was a controlled lab experiment and not a real-world deployment, which limits the potential harm done by EPs. However, the exploration in the case study revealed an important blind spot around the divergent interpretations of explanations based on one's AI background. Building on these insights, we can do further research on a range of relevant areas. For instance, how combinations of user characteristics (e.g., educational and professional backgrounds) impact susceptibility to EPs, how different heuristics can combine to manifest harmful biases, and how different users appropriate explanations in unexpected manners. For further guidance on what human-centered research might look like in XAI and how to do it, we can draw from the domain of human-centered XAI (HCXAI) that has outlined a robust research agenda bridging methods from human-computer interaction (HCI) and theory from critical AI studies.[31–35] Taking a pitfall-aware mindset in these explorations can generate actionable insights about how end-user reactions to AI explanations may diverge from designer intentions.

At the design level, we seek design strategies that are resilient to pitfalls. One possible strategy can be to shift our explanation design philosophy to emphasize user reflection (as opposed to acceptance) during interpretation of explanations. Recent HCXAI work has also advocated for conceptualizing ways to foster trust via reflection.[31] In terms of origins, some pitfalls are a consequence of uncritical acceptance of explanations. Langer et al.[36] point out that people are likely to accept explanations without conscious attention if no effortful thinking is required from them. In Kahneman's dual-process theory[27] terms, this means that if we do not invoke mindful and deliberative (system 2) thinking with explanations, we increase the likelihood of uncritical consumption. To trigger mindfulness, Langer et al.[36] recommend to design for "effortful responses" or "thoughtful responding." To help with mindfulness, we can incorporate the lenses of seamful design,[37] which emphasize configurability, agency, appropriation, and revelation of complexity.[38] Seamful design is the complement of the notion of "seamlessness" in computing systems[37–39] and has conceptual roots in ubiquitous computing.[37]

The notion of seamfulness aligns well with XAI because (1) AI systems are deployed in what Vertesi calls seamful spaces,[40] and (2) the approach can be viewed as a response to "seamless" black-boxed AI decisions with "zero friction" or understanding. In terms of form and function, seams strategically reveal complexities and mechanisms of connection between different parts while concealing distracting elements. This notion of "strategic revealing and concealment" is central to seamful design because it connects form with function.[38] Understanding such connections can promote reflective thinking.[37] Seamful explanations, thus, strategically reveal relevant information that augments system understanding and conceal information that distracts. They shed light on both the imperfections and affordances of the system, awareness of which can add useful cognitive friction and promote effortful and reflective thinking. Examples of seamful explanations include interactive counterfactual explanations where we prompt the user with what-if scenarios. In simple terms, counterfactual examples can help users decipher unknown decision-making processes by understanding the hypothetical input conditions under which the outcome changes.[41] For instance, what if the AI group members were prompted to reflect using counterfactual scenarios on Q values? Making the counterfactuals explicit can help the user to be aware of the variability around the system's decision, which can help build better mental models of the system.[42] Moreover, by facilitating contrastive thinking, counterfactuals elicit engagement and deliberative thinking,[43] which can potentially help navigate around EPs. Recent work on seamful XAI offers a design process that stakeholders can use to anticipate "seams," locate them in the AI's life cycle stage, and leverage them in a way that boosts user agency and AI explainability.[44]

At an organizational level, we can introduce educational (training) programs (e.g., pitfall literacy programs) for both designers and end users. Having an ecosystem perspective is important because EPs have sociotechnical complexities in XAI.[45] Recent work has shown that literacy of DPs can promote self-reflection and mitigate harms.[46] We can develop EP literacy programs that empower (1) designers to proactively address EPs and (2) end users to identify being trapped in pitfalls. These programs can include simulation exercises using speculative design[47] and reflective design[48] to envision "what could go wrong"[49] in facets of XAI systems. The programs should empower designers to critically reflect on problematic organizational incentives that might be pitfall inducing. For instance, if there are incentives to prioritize adoption at all costs, these programs should give agency to stakeholders to resist these incentives to avoid future EPs. Designers can use participatory design[50] with end users to gather effects of potential EPs. While participatory methods are not a silver bullet, they aim to bring parity to power dynamics in design where developers do not necessarily have more voice than other stakeholders. A participatory approach can thus bring diverse voices more equally into the conversation and highlight blind spots that could have been missed otherwise.[44,51–55] They can also utilize case studies (like ours) to think through the effects of the pitfalls. Insights from these programs can facilitate the navigation around pitfalls at both the design and evaluation levels of the ecosystem.

Taken together, these strategies can help us address EPs in a proactive manner, fostering resiliency against them. These strategies are neither exhaustive nor normative but take a formative step toward addressing the EPs.

IMPLICATIONS: DESIGN AND SOCIETAL

The proposed pitfall-aware strategies carry implications at both design and societal levels. The challenging nature of EPs entails an increased need for accountability and consideration of implications from stakeholders across the XAI pipeline (researchers, practitioners, and organizational leaders). Here, we reflect on implications around three areas: reframing AI adoption, recalibrating stakeholder empowerment, and resisting the "move fast and break things" mindset. The implications are informed by Agre's critical technical practice,[56,57] a design philosophy that promotes a self-reflective outlook; for the foreseeable future, work around EPs can benefit from a "split identity" where we simultaneously (1) push the boundaries of pitfall-aware strategies and seamful explanations while (2) reflexively work on our intellectual blind spots. Adopting such an identity balances design (action) and reflection when we think of implications.

Reframing AI adoption

A pitfall-aware (like seamful explanation) design mindset can reframe our thinking around trust building and AI adoption, which have upstream (e.g., research) and downstream (e.g., industry practices) implications. There is an oft-unspoken yet dominant assumption that connects adoption with acceptance, where we view AI adoption emerging from user trust, which, in turn, emerges from user acceptance.[31] Exemplified by multiple technology acceptance models (e.g., TAM[58] and UTAUT[59]), acceptance is seen as a core tenet of trust building (thereby adoption). Given this assumption, there is often an uncritical push toward trust building without asking a fundamental question: is the AI worthy of our trust? Moreover, we should critically reflect: is acceptance the only way to build trust and get adoption? How would we, in a human-human relationship, feel if we are told the only way to trust the other party was exclusively through accepting whatever they said? There are many ways to build trust. In XAI, the principles of reflective HCXAI suggest diverse foundations for trust building, such as healthy skepticism from informed users, awareness of variability of around system decisions, and knowledge of what the AI cannot do (as opposed to the typical what the AI can do).

Connected with the ethos of going beyond user acceptance, pitfall-aware strategies such as seamful explanation design can catalyze a mindset shift that emphasizes critical reflection during AI explanation interpretation (as opposed to unmindful acceptance). Thus, being pitfall aware diversifies our ways of building trust while minimizing the propensity of unmindful acceptance of AI, which, in turn, can mitigate pitfalls such as over-trust (or reliance). Opting to promote reflection in the user can begin the process of defamiliarization from acceptance-first approaches, reframing the discourse around AI adoption. Such mindset shifts in AI adoption have cascading societal ripple effects at both upstream (e.g., research) and downstream (e.g., industry practices) levels. For instance, if an organization embodies a critically reflective AI adoption mindset (as opposed to an acceptance-driven AI adoption one), then it could mitigate perverse organizational incentives, such as prioritizing growth and adoption at all costs. This can reflexively catalyze a change in organizational culture toward responsible and accountable AI governance. Note that we are not advocating to eliminate building trust via acceptance; rather, we are encouraging reforms that go beyond acceptance. More importantly, reflection and acceptance are not mutually exclusive and can work in tandem with each other. Reforming the dialog around AI adoption and trust building promotes our resilience against detrimental pitfalls like over-estimating (or hyping) AI capabilities, which has cascading societal implications around holding these systems accountable.

Recalibrating stakeholder empowerment

Being pitfall aware implies that we cannot afford to treat explainees (users) as passive members in the design space, which has implications on stakeholder empowerment and voice. Recall that EPs are downstream effects, emerging after users interpret explanations. Thus, if we do not incorporate the voices of users as active partners in XAI design and evaluation, then we are unlikely to effectively address pitfalls. This is a shift in perspective from one that treats explanations as a one-directional information transfer to one that views explanations as a co-constructed meaning-making activity.[60–62] Given that EPs are also hard to proactively identify, we are unlikely to succeed with a one-shot (one-and-done) explanation design approach. We need an iterative approach that allows insights from evaluation to feedback to design. We can accomplish this using an ‘‘AI life cycle’’ perspective, one that weaves in relevant stakeholders at different AI touchpoints in the pipeline.[52] Our proposed strategies offer ways to activate user agency, like developing pitfall literacy programs, in a participatory manner. Improving user agency in the process also aligns with a core goal of explainability—informed actionability. Explanations are useful when they are actionable (what can one do with the information), and even more so when the actionability is informed. Through the training programs, stakeholders can become informed. Informing stakeholders alone might not guarantee perfect decision-making, but it can improve decision-making. Combined with their improved agency, their informed status can improve their actionability. Informed user actionability can add to their resilience around pitfalls, which reinforces their empowerment in the design space.

Resisting the ‘‘move fast and break things’’ mindset

An EP-aware XAI strategy advocates to move steadily and critically reflect. It is a departure from the ‘‘move fast and break things’’ (MFBT) mindset,[63,64] one that has had problematic consequences in shaping our notions of innovation.[65] Mindful and measured EP-aware interventions can provide checks and balances against the short intervention-to- application AI cycle (often catalyzed by an MFBT mindset). These checks and balances are important to accommodate the increased accountability required of XAI creators and practitioners to address EPs. Despite the benefits, pitfall-aware steady approaches can face tensions in organizational environments around sustainable implementation. Practitioners in organizational environments can face ‘‘market’’ pressures that disincentivize measured approaches.

To resist the MFBT mindset and combat short-term thinking, we can focus on the costs of not being accountable stewards of technology. AI systems today are powerful; when misused, they can amplify systemic inequities and discriminatory practices.[66,67] Even if organizations might not be intrinsically motivated to do social good, monetary and legal costs are good motivators to promote accountability. Engaging stakeholders, especially leadership, in activities around ‘‘what could go wrong?’’[49] can promote practices that resist the MFBT mindset. The cost of a scandal that destroys user trust or a lawsuit is high. We can engage stakeholders (e.g., leadership) using scenario-based design[55,68] activities (like Tarot Cards of Tech[69]) to envision plausible futures and pitfalls in a holistic manner. For instance, thinking through the question ‘‘what’s the worst news headline you can imagine about your product?’’ can provide generative thoughts that make a strong case for measured EP-aware strategies. Cost-benefit analyses[70] and algorithmic impact assessments[71] from these scenarios can also quantify the negative effects, which can incentivize long-term and sustainable measures. While there will be friction as we steer away from the MFBT mindset, the aforementioned mechanisms can assist in our journey toward a paradigm where we move steadily and ask the right questions.

LIMITATIONS AND FUTURE WORK

This paper begins a cross-disciplinary dialogue around the unintended downstream negative effects of AI explanations by introducing the notion of EPs. Given the formative nature of our work, there are limitations that scope the coverage of our insights. We acknowledge and encourage future work to further refine our ideas around EPs. To that end, we propose three research questions that have emerged from our work. (1) How can we develop a taxonomy of EPs to better diagnose and mitigate its negative effects? (2) How might we use seamful explanations to account for the temporal evolution of pitfalls? (3) How might we assess the impact of training programs to mitigate the effects of pitfalls?

For future refinements around EPs, we are guided by Agre's critical technical practice[56,57] and the idea of a "split identity" that simultaneously (1) push the boundaries of pitfall-aware strategies and seamful explanations while (2) reflexively work on our intellectual blind spots. By operationalizing EPs, we begin the design journey. Now we seek to learn from and with the HCI and AI communities through foundational and applied research to further develop the conceptual and practical facets of EPs.

CONCLUSIONS

Being able to appropriately classify negative impacts of AI explanations is crucial to making XAI systems safe and reliable. By starting the conversation about EPs, this paper brings conscious awareness to the (previously unarticulated) possibility of unintended negative effects of AI explanations. By broadening the scope of harmful effects in XAI, EPs expand the dialog that has already started around DPs. The operationalization of EPs and proposed mitigation strategies provide actionable insights that can improve accountability and safety in XAI systems. We believe that further understanding where, how, and why unintended pitfalls reside in the design space of XAI can lead to improved safety and user empowerment in AI systems.

ACKNOWLEDGMENTS

This project was partially supported by the National Science Foundation under grant 1928586. We are grateful to Zhiyu Lin, Sarah Wiegreffe, Amal Alabdulkarim, Becky Peng, and Kaige Xie for their feedback on the ideas. We are indebted to Samir Passi, Vera Liao, Larry Chan, Ethan Lee, and Michael Muller for their contributions to the case study, which informed the current work. We give special thanks to Rachel Urban for generously providing proofreading feedback.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M.O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In Proceedings of the 24th international conference on intelligent user interfaces, pp. 263–274.

2. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Comput. Surv. *51*, 1–42.

3. Holzinger, A., Biemann, C., Pattichis, C.S., and Kell, D.B. (2017). What do we need to build explainable AI systems for the medical domain?. Preprint at arXiv. https://doi.org/10.48550/arXiv.1712.09923.

4. Katuwal, G.J., and Chen, R. (2016). Machine learning model interpretability for precision medicine. Preprint at arXiv. https://doi.org/10.48550/arXiv.1610.09045.

5. Loftus, T.J., Tighe, P.J., Filiberto, A.C., Efron, P.A., Brakenridge, S.C., Mohr, A.M., Rashidi, P., Upchurch, G.R., and Bihorac, A. (2020). Artificial intelligence and surgical decision-making. JAMA Surg. *155*, 148–158.

6. Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., and Zatloukal, K. (2022). Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. N. Biotech. *70*, 67–72.

7. Hao, K. (2019). AI is sending people to jail—and getting it wrong. Technol. Rev. *21*.

8. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. Q. J. Econ. *133*, 237–293.

9. Rudin, C., Wang, C., and Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. Harvard. Data. Science. Review. *2*, 1.

10. Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. In Proceedings of the 25th international conference on intelligent user interfaces, pp. 275–285.

11. Ghai, B., Liao, Q.V., Zhang, Y., Bellamy, R., and Mueller, K. (2021). Explainable active learning (xAI) toward AI explanations as interfaces for machine teachers. Proc. ACM Hum. Comput. Interact. *4*, 1–28.

12. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–14.

13. Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D.S., and Findlater, L. (2020). No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In Proceedings of the 2020 chi conference on human factors in computing systems, pp. 1–13.

14. Stumpf, S., Bussone, A., and O'Sullivan, D. (2016). Explanations considered harmful? user interactions with machine learning systems. In Proceedings of the ACM SIGCHI conference on human factors in computing systems (CHI).

15. Eiband, M., Buschek, D., and Hussmann, H. (2021). How to support users in understanding intelligent systems? Structuring the discussion. In 26th International Conference on Intelligent User Interfaces, pp. 120–132.

16. Eiband, M., Buschek, D., Kremer, A., and Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. In Extended abstracts of the 2019 CHI conference on human factors in computing systems, pp. 1–6.

17. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., and Hussmann, H. (2018). Bringing transparency design into practice. In 23rd international conference on intelligent user interfaces, pp. 211–223.

18. Brignull, H., Miquel, M., Rosenberg, J., and Offer, J. (2015). Dark patterns-user interfaces designed to trick people. In Proceedings of the Poster Presentation, Australian Psychological Society Congress, pp. 21–23.

19. Chromik, M., Eiband, M., Völkel, S.T., and Buschek, D. (2019). Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In IUI workshops *2327*.

20. Gray, C.M., Kou, Y., Battles, B., Hoggatt, J., and Toombs, A.L. (2018). The dark (patterns) side of UX design. In Proceedings of the 2018 CHI conference on human factors in computing systems, pp. 1–14.

21. Zagal, J.P., Björk, S., and Lewis, C. (2013). Dark patterns in the design of games. In Foundations of Digital Games 2013.

22. Capurro, D., and Velloso, E. (2021). Dark patterns, electronic medical records, and the opioid epidemic. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.08870.

23. Ehsan, U., Passi, S., Liao, Q.V., Chan, L., Lee, I., Muller, M., and Riedl, M.O. (2021). The who in explainable ai: How ai background shapes perceptions of AI explanations. In In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24), May 11–16, 2024, p. 43. https://doi.org/10.1145/3613904.3642474.

24. Noble, S.U. (2018). Algorithms of oppression: How search engines reinforce racism. In Algorithms of oppression (New York University Press).

25. Verma, S. (2019). Weapons of math destruction: how big data increases inequality and threatens democracy. Vikalpa *44*, 97–98.

26. Watkins, C.J., and Dayan, P. (1992). Q-learning. Mach. Learn. *8*, 279–292.

27. Kahneman, D. (2011). Thinking, Fast and Slow (Macmillan Publishers).

28. Petty, R.E., Cacioppo, J.T., Petty, R.E., and Cacioppo, J.T. (1986). The Elaboration Likelihood Model of Persuasion (Springer), pp. 1–24.

29. Wason, P.C., and Evans, J. (1974). Dual processes in reasoning? Cognition *3*, 141–154.

30. Liao, Q.V., Gruen, D., and Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–15.

31. Ehsan, U., and Riedl, M.O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. In HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, Proceedings 22 (Springer International Publishing), pp. 449–466.

32. Ehsan, U., Wintersberger, P., Liao, Q.V., Mara, M., Streit, M., Wachter, S., Riener, A., and Riedl, M.O. (2021). Operationalizing human-centered perspectives in explainable AI. In Extended abstracts of the 2021 CHI conference on human factors in computing systems, pp. 1–6.

33. Ehsan, U., Wintersberger, P., Liao, Q.V., Watkins, E.A., Manger, C., Daumé III, H., Riener, A., and Riedl, M.O. (2022). Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In CHI conference on human factors in computing systems extended abstracts, pp. 1–7.

34. Ehsan, U., Wintersberger, P., Watkins, E.A., Manger, C., Ramos, G., Weisz, J.D., Daumé Iii, H., Riener, A., and Riedl, M.O. (2023). Human-Centered Explainable AI (HCXAI): Coming of Age. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–7.

35. Liao, Q.V., and Varshney, K.R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. Preprint at arXiv. https://doi.org/10.48550/arXiv.2110.10790.

36. Langer, E.J., Blank, A., and Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. J. Pers. Soc. Psychol. *36*, 635–642.

37. Chalmers, M., and MacColl, I. (2003, January). Seamful and seamless design in ubiquitous computing. In Workshop at the Crossroads: The Interaction of HCI and Systems Issues in UbiComp, *8*.

38. Inman, S., and Ribes, D. (2019). "Beautiful Seams" Strategic Revelations and Concealments. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–14.

39. Broll, G., and Benford, S. (2005). Seamful design for location-based mobile games. In International conference on entertainment computing (Springer Berlin Heidelberg), pp. 155–166.

40. Vertesi, J. (2014). Seamful spaces: Heterogeneous infrastructures in interaction. Sci. Technol. Hum. Val. *39*, 264–284.

41. Del Ser, J., Barredo-Arrieta, A., Díaz-Rodríguez, N., Herrera, F., Saranti, A., and Holzinger, A. (2024). On generating trustworthy counterfactual explanations. Inf. Sci. *655*, 119898.

42. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. *267*, 1–38.

43. Byrne, R.M. (2019). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. IJCAI, 6276–6282.

44. Ehsan, U., Liao, Q.V., Passi, S., Riedl, M.O., and Daume, H., III (2022). Seamful XAI: Operationalizing Seamful Design in Explainable AI. Proceedings of the ACM on Human-Computer Interaction *8*, 29. https://doi.org/10.1145/3637396.

45. Ehsan, U., Saha, K., De Choudhury, M., and Riedl, M.O. (2023). Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. Proc. ACM Hum. Comput. Interact. *7*, 1–32.

46. Magnusson, J. (2023). Improving Dark Pattern Literacy of End Users. The Department of Mathematics and Computer Science (Karlstad University), pp. 1–3.

47. Auger, J. (2013). Speculative design: crafting the speculation. Digit. Creativ. *24*, 11–35.

48. Sengers, P., Boehner, K., David, S., and Kaye, J.J. (2005). Reflective design. In Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility, pp. 49–58.

49. Colusso, L., Bennett, C.L., Gabriel, P., and Rosner, D.K. (2019). Design and Diversity? Speculations on what could go wrong. In Proceedings of the 2019 on Designing Interactive Systems Conference, pp. 1405–1413.

50. Muller, M.J., and Kuhn, S. (1993). Participatory design. Commun. ACM *36*, 24–28.

51. Delgado, F., Yang, S., Madaio, M., and Yang, Q. (2023). The participatory turn in ai design: Theoretical foundations and the current state of practice. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–23.

52. Dhanorkar, S., Wolf, C.T., Qian, K., Xu, A., Popa, L., and Li, Y. (2021, June). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In Proceedings of the 2021 ACM Designing Interactive Systems Conference, pp. 1591–1602.

53. Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O., and Weisz, J.D. (2021). Expanding explainability: Towards social transparency in Ai systems. In Proceedings of the 2021 CHI conference on human factors in computing systems, pp. 1–19.

54. 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, pp. 1–14.

55. Wolf, C.T. (2019). Explainability scenarios: towards scenario-based XAI design. In Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 252–257.

56. Agre, P.E. (2014). Toward a critical technical practice: Lessons learned in trying to reform AI. In Social science, technical systems, and cooperative work (Psychology Press), pp. 131–157.

57. Agre, P. (1997). Computation and Human Experience (Cambridge University Press).

58. Davis, F.D., Bagozzi, R.P., and Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. Manag. Sci. *35*, 982–1003.

59. Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F.D. (2003). User acceptance of information technology: Toward a unified view. MIS Q. *27*, 425–478.

60. Lombrozo, T., and Gwynne, N.Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. Front. Hum. Neurosci. *8*, 700.

61. Lombrozo, T., Wilkenfeld, D., Lombrozo, T., and Wilkenfeld, D. (2019). Mechanistic versus functional understanding. Varieties of understanding: New perspectives from philosophy, psychology, and theology *209*.

62. Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). Minds Mach. *29*, 441–459.

63. Chesterman, S. (2021). Move fast and break things": Law, technology, and the problem of speed. Singapore Acad. Law J. *33*, 5–23.

64. Taneja, H. (2019). The era of "move fast and break things" is over. Harv. Bus. Rev. *22*.

65. Taplin, J. (2017). Move Fast and Break Things: How Facebook, Google, and Amazon Have Cornered Culture and what it Means for All of Us (Pan Macmillan).

66. Benjamin, R. (2020). Race after Technology: Abolitionist Tools for the New Jim Code (Polity Books).

67. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Comput. Surv. *54*, 1–35.

68. Rosson, M.B., and Carroll, J.M. (2007). Scenario-based design. In The human-computer interaction handbook (CRC Press), pp. 1067–1086.

69. Hopton, S.B. (2021). The Tarot of Tech. *Equipping Technical Communicators for Social Justice Work: Theories, Methodologies, and Pedagogies*, p. 158.

70. Hurley, N.J., O'Mahony, M.P., and Silvestre, G.C. (2007). Attacking recommender systems: A cost-benefit analysis. IEEE Intell. Syst. *22*, 64–68.

71. Moss, E., Watkins, E.A., Singh, R., Elish, M.C., and Metcalf, J. (2021). Assembling accountability: algorithmic impact assessment for the public interest. SSRN, 3877437.