
Conceptualizing the Relationship between AI Explanations and User Agency

Iyadunni Adenuga

Penn State University
University Park, PA 16802, USA
ija5027@psu.edu

Jonathan Dodge

Penn State University
University Park, PA 16802, USA
jxd6067@psu.edu

Abstract

We grapple with the question: *How, for whom and why should explainable artificial intelligence (XAI) aim to support the user goal of agency?* In particular, we analyze the relationship between agency and explanations through a user-centric lens through case studies and thought experiments. We find that explanation serves as one of several possible first steps for agency by allowing the user convert forethought to outcome in a more effective manner in future interactions. Also, we observe that XAI systems might better cater to laypersons, particularly “tinkerers,” when combining explanations and user control, so they can make meaningful changes.

Author Keywords

Explainability; Agency; AI systems

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Copyright held by the owner/author(s).
CHI Workshops '23, April 23–28, 2023, Hamburg, Germany
ACM 978-1-4503-6819-3/20/04.
<https://doi.org/10.1145/3334480.XXXXXX>

Introduction

Complex technologies are commonplace in today’s society, with examples including reinforcement learning, deep neural networks, or other forms of artificial intelligence (AI). Criticisms have plagued the acceptance of these technologies due to the opaque nature of the algorithms and the erasure of user influence (i.e., creating an automated experience). For example, high-stakes scenarios (e.g. law enforcement, medicine, etc.) traditionally require human experts that go through rigorous training, who are then accountable to human stakeholders. Thus, it is unsurprising that such decision makers prefer worse-performing, interpretable models over opaque models [44].

Beyond experts, laypeople also desire a level of control and understanding of the complex AI systems that affect them [47, 39]. Legal regimes (e.g., European Union General Data Protection Regulation [17] and White House Executive Order [22]) align with such observations by highlighting the importance of human agency over these systems and the need for these systems to explain and justify their results.

However, making AI systems more *agentic* is not as widely studied as making them *explainable*. This paper attempts to describe how designing for agency fits with XAI, namely: 1) the relationship between agency and explanations and 2) agency in scenarios with two and three user groups.

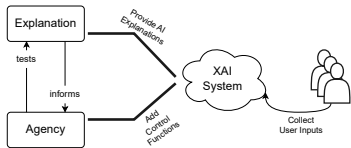


Figure 1: Relationship between Agency and Explanations in an XAI system

What is an Explanation?

Explanation is a human phenomenon that strongly relates with peoples’ mental models, understanding and knowledge of “why an outcome happened” [25]. Its social interactive characteristic [25] means there’s some level of communication (which may be continuous) occurring between the explain-er and explain-ee.

Today, AI systems are a major part of our environment. If the target users do not understand the model, they usually cannot assess or appropriately rely [37] on it. To address this, post-hoc methods aim to make opaque AI methods (e.g. neural networks, ensemble models, etc.) more “understandable” without compromising accuracy [3, 13]. There are two main approaches employed by post-hoc techniques: *opaque box* (operates on the input/output boundary; e.g., LIME [35], LORE [20]) and *transparent box* (operates on the internal structures; e.g., deconvnet model method [49] and network dissection [7, 8]).

Existing XAI systems that utilize the opaque- and transparent-box approaches described above do not fit the requirements laid out in prior work for “everyday” explanations understandable to the layperson [32]. AI explanations created based on human characteristics (e.g. preferences, reasoning and perception methods) are more relatable and effective [45, 50, 28]. In their work about connecting existing XAI techniques to user expectations for explanations, Liao et al. [29] propose a “question-driven framework” that encourages an interactive explanation experience [29] via meaningful interrogative dialogue [32].

What is Agency?

People have an innate need to control the course of their lives and predict the outcomes of situations, no matter the difficulty [5]. Humans feel a sense of agency when we

believe that our “conscious intention caused a voluntary action” [46]. Agency is an internal “human” feeling that is outwardly expressed by intentional actions. If people do not feel in control, they might abandon the on-going task or distrust their actions, especially in hard situations [4].

A technology that affords agency is “flexible” to the user’s interactions inputs and interests such that they can modify their experience [48, 42]. The control a person has in a typical environment (such as while utilizing technology) can be weighted by: 1) the presence of relevant actions; 2) the relationship between the actions of a user and the outcome in the environment; 3) the ability of a user to predict the outcome of their actions, and; 4) the ability of the user to trace the cause of an outcome [41].

Researchers have shown agentic interactions have positive effects such as improved user experience and satisfaction and more appropriate trust [18, 23, 43]. The many AI systems stakeholders with low technical knowledge should also experience these benefits, as per the ACM Code of Ethics: “...all people are stakeholders in computing” [1].

How are Agency and Explanation Related?

The answer to this question is not straightforward, but we will attempt an answer for AI systems. Existing human-centered XAI systems prioritize providing explanations in an understandable, visually appealing format with an assumption of improved agency in the represented artifact. There is no *direct* measure for the “actual” agency a user experienced while interacting with such a system. *Self-reporting* only measures agency *perceived* by users, which is a proxy for “actual” agency. Teasing out the modalities of the relationship between explanation and agency is the first step in *deducing* the “actual” agency in XAI systems.



Figure 2: From Zhang and Lim paper [50], user interface of Counterfactual Explanation for Voice-Emotion Recognition system (best viewed digitally).

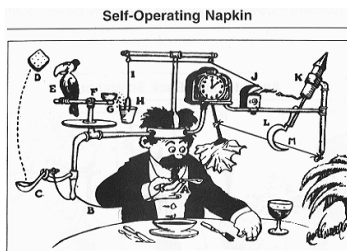


Figure 3: The original Rube Goldberg machine, as depicted in the cartoon “Professor Butts and the Self-Operating Napkin.” The machine functions as follows: “Soup spoon (A) is raised to mouth, pulling string (B) and thereby jerking ladle (C), which throws cracker (D) past toucan (E). Toucan jumps after cracker and perch (F) tilts, upsetting seeds (G) into pail (H). Extra weight in pail pulls cord (I), which opens and ignites lighter (J), setting off skyrocket (K), which causes sickle (L) to cut string (M), allowing pendulum with attached napkin to swing back and forth, thereby wiping chin.”

https://en.wikipedia.org/wiki/Rube_Goldberg_machine

The agency process starts from a person’s *forethought* to their performance of an action and then, observation of the action’s outcome. The aim of explanations is to improve the consumers’ understanding of their environment. Providing explanations can contribute to consumers’ sense of agency by informing their initial *forethought* so they perform the appropriate actions to successfully complete their task. People with higher need for control are more likely to seek more information and clarifications in a work environment [19]. This shows that even before the introduction of explanations to a scenario, an individual has an inherent agency requirement—and that such requirements will vary among users.

Studies on designing agency in AI systems, such as interactive machine learning, have primarily focused on users with technical know-how, and in its absence, requires additional technical training for end-users for them to understand and use the provided agentic functions [40, 14]. End-users with no access to technical training can still benefit from an agentic experience with explanations.

Users can take an active role in their absorption of an explanation. Zhang and Lim [50] studied providing relatable explanations for a vocal-emotion recognition system, the participants preferred and utilized more effectively the “Counterfactual Sample + Cues” explanation (Figure 2). The user interface for this explanation required active play-through and listening to alternative voices to detect vocal differences. Another method for involving people in the explanation process is to obtain input from them to create “selective” explanations [28]. Here, the user customizes the types of received explanations to their taste.

Tastes vary, for example GenderMag [9] identified facets describing people’s cognitive styles. One important axis is *learning style*, with people who gain understanding by

“tinkering” with the technology on one end. To cater to tinkerers, XAI system designs should have control functions. These functions would accept different kinds of user inputs and then provide appropriate visible outcomes, allowing the system to “*be actionable*” [27]. On the other end of the learning style axis are people who gain understanding by *process*. Process-oriented learners may benefit from assessment *processes*, such as *After-Action Review for AI* (AAR/AI) [12]. Later, Khanna et al. [26] found participants helped participants examine and effectively use explanations to identify AI faults, observing a moderate-sized practical effect.

There are situations when XAI systems cannot honor user inputs [40]. How should the system react? For low-stakes scenarios, illusory agency may be a useful tool. Game designers use this as a complementary mechanism to preserve their rigid game-story narrative [11, 31]. To allow for continued user agency, the user is able to observe the effect of their input but the input has no effect on the underlying algorithms. Vaccaro et al. [43] showed in a social media setting that users “*felt more satisfied with the presence of controls*” regardless of their effectiveness. Some everyday systems that already utilize illusory agency include crosswalk buttons and elevator close-door buttons. Illusory agency should only be designed to supplement the already present “real” agentic experience in low-stakes scenarios so as to avoid user deception and minimize the effects of ethical issues. Example of such scenarios that might benefit from illusory agency include XAI systems in training environments [16].

How Does One Increase or Decrease Agency?

We will use two examples to illustrate adjusting agency. The first example is to consider wiping your mouth with a napkin using direct manipulation vs with a Rube Goldberg

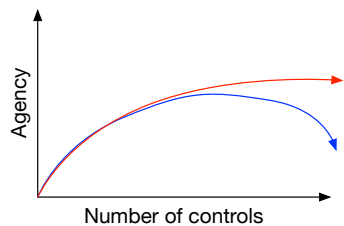


Figure 4: Notional curve depicting agency as a function of number of controls. When moving from no controls to few controls, agency gain enjoys a direct relationship. But at some point, the extra controls will overwhelm the user, either taking the form of a plateau (red curve) or even a downturn (blue curve).

- 1) Add/remove source documents
- 2) Add/remove sections, where sections are subtopics of the document title
- 3) Add/remove words and/or sentences
- 4) View the order relation of the summary sentences in a concept graph
- 5) View actual sentences contributed by each document to the overall summary output

Table 1: Key functionality found in Living Documents.

machine, depicted in Figure 3. Rube Goldberg machines are famous for having a simple input, which then initiates a complicated chain reaction generating a simple output. In this case, the simple input is lifting a spoon, the long chain reaction is via the crackers, toucan, string, etc., and the simple output is wiping one's mouth.

As Figure 3 shows, the machine automates the functioning of the napkin to the point that its use is involuntary. Suppose we changed the simple input to be pushing a button, which is more typical of modern technology. Now, consider how much agency the user has in each case. It seems fairly obvious that agency would be highest with direct manipulation, then with button-interface Rube Goldberg machine, and finally the unmodified Rube Goldberg machine. The reasoning is that with direct manipulation, one could manifest whatever wiping approach they desire: arbitrary direction, length of time, and so on. Note that all of the previously compared interactions lead to same *outcome*, but are different in terms of controllability [36]. According to Shneiderman [38], high levels of automation and human control can co-exist in a technical artifact. They illustrate this in their description of the digital camera and elevator where agency is afforded by the inclusion of a button and settings page respectively. This is similar to the surface-level agency button introduced above, to the Rube Goldberg machine. Would increasing agency require addition of extra buttons/settings or introducing a more manual and influential process (i.e., less automation)? If we assume adding a feature and accompanying button increases perceived agency, what is the amount of UI complexity at which agency gains diminish or turn negative (See Figure 4 for an illustration)?

Thus, we have illustrated a tension between manual processes (which have the highest agency) vs automation

(which reduces agency). The open question is, how much agency does one lose when a process undergoes automation? To answer this question, we turn to our second example: Living Documents [2]. Living Documents is an interactive multi-document text summarization system, providing the control functions found in Table 1.

What would agency treatment levels look like in Living Documents? The highest-level of agency is full access to all functionality in the system (full-agency, see Table 1) while the automation level has no user controls (no-agency). This means it would work like a typical text summarization system such that the only input-output operation is the user providing the source documents and receiving the summary result. The interaction designer can decide on intermediate level(s) based on specified criteria. Our criterion is "magnitude of impact" (document → sections → sentences/words), so our some-agency treatment has functionality 3 to 5 in Table 1.

Agency and/or Explanations, for whom?

Now, we would like to broaden the previously discussed system-user agency cases to where there are *multiple types* of users, leading to a more complex tradeoff relationship. From an explanation perspective, we know that explanations may need to account for different domain expertise, cognitive abilities, context of use, and audience. Users have varied needs for an XAI system [29] and do not have a homogeneous process for interacting with models [40]. The agency perspective is less well studied.

Consider the case of a rideshare application called Co-opRide, which is an algorithmic manager for *two* user groups: drivers and riders. What is the right agency balance to strike between these three parties (the third is Co-opRide)? Suppose Co-opRide offers a design feature

Hotelling's Game [21]

Suppose two competing shops are located along the length of a street, with customers spread equally along the street. Each customer will always choose the nearer shop.

With two shops, the consumer ideal has shops at $\frac{1}{4}$ and $\frac{3}{4}$. However, this is unstable, since both shops can claim more customers by moving toward the middle. The stable solution has both at $\frac{1}{2}$. (E.g., this is why Lowe's and Home Depot are often co-located.)

that drivers can veto riders. This would increase the agency of the drivers at the expense of the riders' agency, as well as that of the system provider. Imagine being a rider receiving vetoes from several drivers, based on your low population density location or even worse, cultural markers present in your name. This might lead to long wait times and negative customer sentiment. Should the XAI system alert riders that a driver vetoed them? Each time? How should the XAI system provide notification of the veto and/or explain the decision? If there are no satisfying agentic actions available, why should the system provide explanation at all? In the case where the rider's waiting time increases over time as a result of receiving multiple driver vetoes, provision of explanations by the algorithmic XAI platform becomes even more imperative.

The example of the driver veto feature suggests "The Agency Tradeoff Game" might be zero-sum, though this it is not totally clear that it cannot be positive-sum or negative-sum. It is also an open question whether or not a stable solution exists. As an example, "Hotelling's game" (see margin) has a stable solution with two players, but the three-player version has no stable solution ([34], Chapter 3).

XAI platforms interact with groups of humans; as a result, agency occurring on a *collective* basis becomes relevant. When individuals perform a joint action, they can feel individual and/or collective (joint) agency [30]. The individual perceived self-efficacy of multiple members of a group forms collective efficacy which can lead to meaningful social change [4, 6]. People are usually interested in the experiences of their fellow people and this has led to calls for social explanations [29, 15]. Enabling social explanations means there can be a joint platform for "knowledge sharing" and "social learning" [15]. People can confidently contest the decisions by an AI system and if

some form of collective agency has been designed in the system, they can effect popular meaningful change.

For an example of collective agency as a result of the conditions in an AI system, consider work by Calacci and Pentland [10]. Shipt, a grocery delivery service with an AI algorithmic manager, was initially explainable and transparent about its wage calculation process. The introduction of a wage-processing opaque-box algorithm to Shipt led to the implementation of social explanations, albeit on a platform (called Shipt Calculator [10]) external to the AI system. Workers anonymously provided screenshots of their payment history and the Shipt Calculator aggregated the payment information and provided the observed wage difference to the workers. These authors discovered a paycut for 41 percent of the workers in their study. Similar occurrences with Doordash led to change in the pay and tipping model as well as a class-action lawsuit [33, 24]. XAI systems that allow social explanations and collective agency would ensure a cooperative approach so that only beneficial improvements are implemented on the platform.

Concluding Remarks

We do not claim that our statement of the relationship between agency and explanation is complete. As an example, perhaps agency and explanations might share multiple simultaneous relationships. We believe that recognizing and formalizing these relationship(s) would ensure that XAI designers take the closely-related extra step of designing for *agency* while working on explainability.

REFERENCES

- [1] ACM. 2018. ACM Code of Ethics and Professional Conduct. (2018).
<https://www.acm.org/code-of-ethics>

- [2] Iyadunni J Adenuga, Benjamin V Hanrahan, Chen Wu, and Prasenjit Mitra. 2022. Living Documents: Designing for User Agency over Automated Text Summarization. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.
- [3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and others. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [4] Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American psychologist* 37, 2 (1982), 122.
- [5] Albert Bandura. 1996a. Reflections on human agency: Part I. *Constructivism in the Human Sciences* 1, 2 (1996), 3.
- [6] Albert Bandura. 1996b. Reflections on human agency: Part II. *Constructivism in the Human Sciences* 1, 3/4 (1996), 5.
- [7] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6541–6549.
- [8] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30071–30078.
- [9] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software’s gender inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787.
- [10] Dan Calacci and Alex Pentland. 2022. Bargaining with the Black-Box: Designing and Deploying Worker-Centric Tools to Audit Algorithmic Management. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 428 (nov 2022), 24 pages. DOI : <http://dx.doi.org/10.1145/3570601>
- [11] Timothy Day and Jichen Zhu. 2017. Agency informing techniques: Communicating player agency in interactive narratives. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. 1–4.
- [12] Jonathan Dodge, Roli Khanna, Jed Irvine, Kin-ho Lam, Theresa Mai, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Andrew Anderson, Sai Raja, Caleb Matthews, Christopher Perdriau, Margaret Burnett, and Alan Fern. 2021. After-Action Review for AI (AAR/AI). *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 29 (sep 2021), 35 pages. DOI : <http://dx.doi.org/10.1145/3453173>
- [13] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.
- [14] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiIS)* 8, 2 (2018), 1–37.

- [15] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [16] Krzysztof Fiok, Farzad V Farahani, Waldemar Karwowski, and Tareq Ahram. 2022. Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation* 19, 2 (2022), 133–144.
- [17] GDPR. 2018. European Union General Data Protection Regulation, Article 15 - “Right of access by the data subject”. (2018).
<http://www.privacy-regulation.eu/en/article-15-right-of-access-by-the-data-subject-GDPR.htm>
Accessed: 1/16/2019.
- [18] Andreas Girgensohn, Sara A Bly, Frank Shipman, John S Boreczky, and Lynn Wilcox. 2001. Home Video Editing Made Easy-Balancing Automation and User Control.. In *INTERACT*, Vol. 1. 464–471.
- [19] Michael Goller and Christian Harteis. 2017. Human agency at work: Towards a clarification and operationalisation of the concept. In *Agency at work*. Springer, 85–103.
- [20] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.
- [21] Harold Hotelling. 1929. Stability in Competition. *The Economic Journal* 39, 153 (1929), 41–57.
<http://www.jstor.org/stable/2224214>
- [22] White House. 2022. Blueprint for an AI Bill of Rights. (2022). <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>
Last accessed: 10/13/22.
- [23] Jina Huh, Martha Pollack, Hadi Katebi, Karem Sakallah, and Ned Kirsch. 2010. Incorporating user control in automated interactive scheduling systems. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. 306–309.
- [24] Catie Keck. 2019. DoorDash tip-skimming scheme prompts class action lawsuit seeking all those tips that didn’t go to drivers. (Jul 2019).
<https://gizmodo.com/door-dash-tip-skimming-scheme-prompts-clash-action-lawsuit-1836820630>
- [25] Frank C Keil. 2006. Explanation and understanding. *Annual review of psychology* 57 (2006), 227.
- [26] Roli Khanna, Jonathan Dodge, Andrew Anderson, Rupika Dikkala, Jed Irvine, Zeyad Shureih, Kin-ho Lam, Caleb R Matthews, Zhengxian Lin, Minsuk Kahng, and others. 2022. Finding AI’s faults with AAR/AI: An empirical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 1 (2022), 1–33.
- [27] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- [28] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *arXiv preprint arXiv:2301.09656* (2023).

- [29] Q Vera Liao, Daniel Gruen, and Sarah Miller. 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*. 1–15.
- [30] Janeen D Loehr. 2022. The sense of agency in joint action: An integrative review. *Psychonomic Bulletin & Review* (2022), 1–29.
- [31] Esther MacCallum-Stewart and Justin Parsler. 2007. Illusory agency in vampire: The masquerade–Bloodlines. *Dichtung Digital. Journal für Kunst und Kultur digitaler Medien* 9, 1 (2007), 1–17.
- [32] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.
- [33] Andy Newman. 2019. Doordash changes tipping model after uproar from customers. (Jul 2019). <https://www.nytimes.com/2019/07/24/nyregion/doordash-tip-policy.html?action=click&module=Intentional&pgtype=Article>
- [34] Martin J. Osborne. 2004. *An introduction to game theory*. Oxford Univ. Press, New York, NY [u.a.]. http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+369342747&sourceid=fbw_bibsonomy
- [35] 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.
- [36] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. DOI : <http://dx.doi.org/10.1145/3290605.3300750>
- [37] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422. DOI : <http://dx.doi.org/10.1145/3581641.3584066>
- [38] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
- [39] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. Grand Challenges for HCI Researchers. *Interactions* 23, 5 (aug 2016), 24–25. DOI : <http://dx.doi.org/10.1145/2977645>
- [40] Alison Smith-Renner, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2020. Digging into User Control: Perceptions of Adherence and Instability in Transparent Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 519–530. DOI : <http://dx.doi.org/10.1145/3377325.3377491>
- [41] Suzanne C Thompson, Wade Armstrong, and Craig Thomas. 1998. Illusions of control, underestimations, and accuracy: a control heuristic explanation. *Psychological bulletin* 123, 2 (1998), 143.

- [42] David Thue, Vadim Bulitko, Marcia Spetch, and Trevon Romanuik. 2010. Player agency and the relevance of decisions. In *Joint International Conference on Interactive Digital Storytelling*. Springer, 210–215.
- [43] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. DOI : <http://dx.doi.org/10.1145/3173574.3173590>
- [44] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 440, 14 pages. DOI : <http://dx.doi.org/10.1145/3173574.3174014>
- [45] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [46] Daniel M Wegner and Thalia Wheatley. 1999. Apparent mental causation: Sources of the experience of will. *American psychologist* 54, 7 (1999), 480.
- [47] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. DOI : <http://dx.doi.org/10.1145/3173574.3174230>
- [48] Peta Wyeth. 2007. Agency, tangible technology and young children. In *Proceedings of the 6th international conference on Interaction design and children*. 101–104.
- [49] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.
- [50] Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.