Expressive HCXAI: An Art-Science Design Framework for Ethical and Usable AI Systems

Dashiel Carrera
University of Toronto
Toronto, ON, Canada
dcarrera@dgp.toronto.edu

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'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

Artificial Intelligence (AI) systems are often difficult for users to understand despite their widespread usage on digital platforms. This hinders users' ability to understand and therefore effectively criticize unethical behavior from AI systems. In this paper, we propose a new design framework for thinking about HCXAI which extends Michael Mateas's notion of Expressive AI design in order to address the challenges of HCXAI design for lay users. Rather than translating multi-billion-parameter models into meaningful and accessible explanations to lay users, this framework takes a cue from the arts and helps designers of AI systems express a reasonable but limited "performance" of an AI system in order to foster an appropriate mental model. This design framework shows how a productive but limited articulation of an AI system may facilitate more ethical AI design.

Author Keywords

HCXAI; AI; Expressive AI; art-science collaboration; mental models; conceptual metaphors; interface metaphors; folk theories; interpretive affordances.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** *HCXAI;* Literature review; Please use the 2012 Classifiers and see this link to embed them in the text:

https://dl.acm.org/ccs/ccs_flat.cfm

Introduction

Recently, concern has mounted as to how, unbeknownst to us, AI is effecting the ways in which we consume information and make decisions [17, 4]. According to the Biden Administration's recently released "Blueprint for an AI Bill of Rights" [1], AI can effect hiring and credit decisions and embed harmful bias and discrimination. Recent work in HCI studying social media has noted how AI can prevent us from consuming media outside of our AI-assessed tastes [14] and has been responsible for rapidly generating and distributing fake news accounts [13]. The presence of AI in high-stakes situations continues to grow, with large-scale social implications [15].

In order to encourage transparency and accountability of AI systems, HCI practitioners have begun to explain AI systems to the public through Human-Centered Explainable AI (HCXAI). HCXAI attempts to explain to users why particular decisions were made by an AI while remaining sensitive to the social context in which the explanation will be interpreted [6, 23, 9]. HCXAI explanations face several challenges with lay users. Notably, explanations targeted at lay users are limited by users' grasps of AI and education is difficult [12]. Another challenge, described in this workshop, is "translating multi-billion-parameter models into meaningful and accessible explanations for lay users." HCXAI systems responsible for explaining decisions that have a large number of parameters often rely instead on additional AI models to generate post-hoc explanations [21], but cannot generate direct explanations for the behavior of the system.

In this paper, I propose a tentative design framework that extends Michael Mateas's concept of "Expressive AI" to pave the way for informed AI interactions with lay users [18]. This framework shifts focus from providing an "explanation" about how the AI works to performing a story about how the AI works to its user. This story may or may not reflect how the AI system works but nonetheless can be held to certain ethical standards.

Expressive AI is a useful framework because it acknowledges designers' limited capacity to articulate complex LLMs to users. Rather than aiming to prescribe an understanding of an AI system, the proposed Expressive HCXAI design framework advocates performing the AI in a manner which affords users a reasonable and ethical understanding of the system's behaviour. For instance, a version of ChatGPT designed with the HCXAI framework in mind may consider displaying text generation with a search animation that looks through different entries on the Internet and pieces them into one text. Even though this is not how ChatGPT generally generates a text response, this animation would express to users that ChatGPT is generating its responses based on numerous sources of questionable reliability, using the Internet as a familiar metaphor [3]. This performance may therefore guide users toward a more productive and ethical conception of the system that spurs them to interrogate it more deeply.

Like previous work presented in this HCXAI workshop [16], this design framework builds on the idea of a "mental model." A mental model is a user's internal representation of how a system works and is formed through experience, perception, and learning [8]. Mental models help users predict outcomes, solve problems, and develop new concepts. They do not necessarily accurately reflect how a system actually works, but serve a valuable utility [20]. For example, simplified models of the atom are often presented to high school students so that they can gain a partial understanding of the fundamentals of the atom at an early stage of their ed-

ucation [8]. These models are educationally useful even though only approximately accurate.

Mateas's concept of "Expressive AI" is similar to mental models in that it focus on the importance of a user's partial understanding of a system. Pulling from an art-science background, Mateas crucially describes interactions with AI systems as an act of theatre. An AI designer performs a story about how the system works to its user through interactions. He argues this is analogous to how a playwright might tell a story to an audience through the performers of a play. For example, the Eliza AI chatbot was perceived as more complex than the system actually was because the AI performed the role of a psychologist [22]. If instead ELIZA had been asked to take on a different role, the system's limited complexity may have been more clear.

Mateas also crucially introduces the concept of "authorial affordances" in his work to describe the tools available to AI designers to tell the story of how an AI works and the concept of "interpretive affordances" to describe the tools available to users to understand that story. In the case of the ELIZA chatbot, some of the authorial affordances available to the designer were the framing that the AI was a psychologist and the algorithm used to turn user responses back into question form. Some of the interpretive affordances available to users of ELIZA were their previous knowledge and expectations of therapy sessions and their interactions with previous chat systems.

In order to develop the Expressive HCXAI framework, several tasks will need to be performed: (1) understanding existing "authorial affordances" for an AI system (2) understanding existing "interpretive affordances", and (3) understanding what mental models are fostered by each of these affordances.

This design framework will therefore need to address the following research questions:

1. What mental models does an HCXAI system hope to foster? What tradeoffs between accuracy and utility are ethical?
2. What existing "authorial affordances" for HCXAI may be used to foster mental models which encourage healthy critique and user awareness of AI systems?
3. What "interpretive affordancs" do users have for understanding AI agents? How and when are these interpretive affordances used or abused?
4. How do particular pairings of authorial and interpretive affordances influence the mental models users develop for AI systems?
5. What new and appropriate authorial affordances might be created to design expressive HCXAI systems?

Literature Review

We now introduce examples from HCI research that help outline existing authorial and interpretive affordances for AI. In general, research about authorial and interpretive affordances of AI are found in two previous areas of work: conceptual metaphors and folk theories.

Authorial Affordances: Conceptual Metaphors

Conceptual metaphors in Human-Computer Interaction enable a designer to articulate how a computing system should be used by exploiting knowledge the user has from another domain. Interface metaphors, a type of conceptual metaphor articulated to users through a UX, are critical for helping everyday people develop quick, unconscious

understandings of digital systems without additional education [19]. For instance, early in the history of computing the "Desktop Metaphor" enabled lay users to understand the computer as a physical work table [2] by embodying virtual elements (e.g., files, folders, trash cans) that mimicked the physical office environment. Recent work shows that these quick, unconscious understandings of AI can lead to a more rich understanding of how an AI system behaves in practice than their slower, more considered developed counterparts [8]. Finding effective conceptual metaphors for AI is therefore critical for finding authorial affordances for Expressive HCXAI systems.

Recent work suggest that conceptual metaphors can have a serious effect on how users perceive AI systems. For example, the paper *Conceptual Metaphors Impact Perceptions of Human-AI Collaboration*, [11] explored how staging different metaphors for an AI agent, like "toddler," "middle schooler," "young student," or "shrewd travel executive," influenced user expectations and subsequent evaluations of AI agents. Metaphors associated with lower competence were rated higher for usability, adoption intention, and cooperation, despite the common practice of portraying AI agents as highly competent. Another recent work, the paper *Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing*, Jung et al [10], adopted the 'Great Chain of Being' framework to explore the impact of non-human metaphors on worker engagement with AI chatbots in crowdsourcing environments. The research team found that metaphorical representations, particularly non-human ones, significantly affected worker engagement, cognitive load, intrinsic motivation, and trust in the agents. For instance, using an inorganic object metaphor (like a book) can reduce cognitive load but might negatively impact worker motivation.

Interpretive Affordances: Folk Theories

The concept of "folk theories" in Human-Computer Interaction (HCI) refers to the informal, often subconscious beliefs and assumptions that users hold about how technologies work [7]. These theories are not typically based on technical knowledge but are rather constructed from users' everyday experiences, cultural context, and interactions with technology. The particular cultural narratives, popular conceptual metaphors, and shared beliefs of a society all strongly influence how folk theories are formed. They may or may not align with how a particular technology actually works. Because folk theories describe how users reason about an AI system, they are a type of interpretive affordance.

There are a a couple works in the HCI literature that discuss folk theories for AI. *What's the Folk Theory? Reasoning About Cyber-Social Systems* discusses how folk theories effect people's understanding of Twitter and Facebook feed curation [7]. The authors identified four common folk theories: the rational assistant, the unwanted observer, the transparent platform, and the corporate black box. Another work, *How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation* studied the formation of folk theories and found that individuals draw from various information sources to form their folk theories, suggesting these beliefs are more complex, multifaceted, and adaptable than previously understood [5].

Conclusion

This paper presents a developing design framework called Expressive HCXAI. This art-science inspired framework encourages the design of AI that "expresses" an ethical story about how an AI system works rather than trying to provide a perfect explanation. This expression must sensitively con-

sider the interpretive affordances of users and the authorial affordances of AI designers in order to help foster reasonable and ethical mental models in lay users. This framework contributes to this workshop by addressing the question "How do we address the difficulty in translating multi-billion-parameter models into meaningful and accessible explanations for lay users?"

REFERENCES

[1] 2022. Blueprint for an AI bill of rights. (Oct 2022). https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[2] Anand Agarawala and Ravin Balakrishnan. 2006. Keepin'it real: pushing the desktop metaphor with physics, piles and the pen. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1283–1292.

[3] Ted Chiang. 2023. ChatGPT is a blurry JPEG of the web. *The New Yorker* 9 (2023), 2023.

[4] Daniel Wigdor Dashiel Carrera, Robert Soden. 2023. Beyond Intelligence: Finding New Metaphors for AI through a Collaborative Workshop with Artists and Scholars. In *Proceedings of the 2023 DIS Conference Towards Mutual Benefit: Reflecting on Artist Residencies as a Method for Collaboration in DIS Workshop*.

[5] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[6] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Yokohama, Japan, 19. DOI: http://dx.doi.org/10.1145/3411764.3445188

[7] Megan French and Jeff Hancock. 2017. What's the folk theory? Reasoning about cyber-social systems. *Reasoning About Cyber-Social Systems (February 2, 2017)* (2017).

[8] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, and others. 2020. Mental models of AI agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[10] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[11] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[12] Bran Knowles. 2023. Explainable AI: Another Successful Failure. In *Proceedings of the 2023 CHI conference on human factors in computing systems HCXAI Worksghop*.

[13] Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117.

[14] Christoph Krönke. 2020. Artificial Intelligence and Social Media. In *Regulating Artificial Intelligence*. Springer, 145–173.

[15] Richard N Landers and Tara S Behrend. 2022. Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist* (2022).

[16] Mareike Petrosjan Lisa Graichen, Matthias Graichen. 2001. How to Facilitate Mental Model Building and Mitigate Overtrust Using HCXAI. (2001).

[17] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[18] Michael Mateas. 2001. Expressive AI: A hybrid art and science practice. *Leonardo* 34, 2 (2001), 147–153.

[19] Dennis C Neale and John M Carroll. 1997. The role of metaphors in user interface design. In *Handbook of human-computer interaction*. Elsevier, 441–462.

[20] Donald A Norman. 1988. *The psychology of everyday things.* Basic books.

[21] Julian Tritscher, Markus Ring, Daniel Schlr, Lena Hettinger, and Andreas Hotho. 2020. Evaluation of post-hoc XAI approaches through synthetic tabular data. In *Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings*. Springer, 422–430.

[22] Noah Wardrip-Fruin. 2007. Three Play Effects–Eliza, Tale-Spin, and Sim City. *Digital Humanities* (2007), 1–2.

[23] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 563–574.