Designing AI Expressivity: An Art-Science Design Framework for Ethical and Usable AI Systems

Dashiel Carrera
dcarrera@dgp.toronto.edu
DGP Lab, University of Toronto

ABSTRACT

Artificial Intelligence (AI) systems are often difficult for users to understand despite their widespread usage on digital platforms. This hinders usability and can prevent users from effectively criticizing AI systems or identifying new problem domains in which AI may be useful. In my proposed doctoral work, I present a new design framework for AI that extends Michael Mateas's "Expressive AI" to address poor AI understanding. Rather than attempting to educate the public about AI, this framework helps AI systems express a reasonable but limited "performance" of an AI system to its user that fosters an appropriate mental model. I aim to create this design framework by: (1) cataloging existing and discovering new authorial affordances and conceptual metaphors are available ti designers of AI systems (2) discovering what interpretive affordances and folk theories laypeople have for interpreting AI systems and (3) understanding how these combinations of authorial and interpretive affordances foster various mental models.

KEYWORDS

art-science collaboration, artificial intelligence, ai literacy, conceptual metaphors, mental models, folk theories, expressive ai, explainable ai, xai, hcxai

ACM Reference Format:
Dashiel Carrera. 2023. Designing AI Expressivity: An Art-Science Design Framework for Ethical and Usable AI Systems. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Artificial intelligence is steadily becoming a core component of technologies that directly impact users. Yet, the underlying algorithms of these prevalent platforms often remain unclear to the individuals using them, leading many to be unaware of their engagements with AI systems [8]. This lack of understanding can hinder users' capacity to utilize, work in tandem with, and critically evaluate AI technologies [9]. Users who encounter AI systems with unclear capabilities or motivations for their behavior are often left dissatisfied or frustrated with their interactions [2], leading some to lose confidence in these technologies and ultimately abandon them [16]. This lack of continued use may prevent users from finding other domains in which AI systems can be useful, and make it more difficult for users to hold these systems accountable [6].

In order to address user dissatisfaction and encourage transparency and accountability of AI systems, HCI practitioners have taken to educate or explain AI systems to the public, notably through the field of "Explainable AI" (XAI). XAI systems attempt to explain to users why particular decisions were made by an AI [6, 13, 36]. However, both educational and explanatory interventions have their limitations. AI edication takes time and resources both on the part of the educator and the lay user. Additionally, lay user's ability to develop a strong grasp of AI or interpet AI explanations may be limited. AI Explainable systems responsible for explaining decisions made by AI systems that have a large number of parameters like LLMs or DNNs often rely instead on additional AI models to generate post-hoc explanations [32], but cannot generate direct explanations for the behavior of the system. While there are many domains in which these limitations may not be a concern, when dealing with lay users who use AI systems less frequently, they may be a more significant barrier to helping the public understand AI.

In my doctoral work, I propose a design framework that extends Michael Mateas's concept of "Expressive AI" to offer a new strategy to combat poor AI understanding [24]. The aim of the design framework in this proposed doctoral work is to pave the way for AI systems which can be used in an educated manner without additional explanation. To do this, instead of putting the onus on the user to learn about AI, this framework puts the onus on the designer to "express" a reasonable story about how the AI works to the user. In doing so, this framework sacrifices accuracy for utility, fostering a reasonable mental model for the system in users rather than trying to explain exactly how it works.

Mateas's work crucially describes interactions with AI systems as an act of theatre. An AI system, he argues, conveys a constellation of ideas and experiences from the creators of the AI system to the audience through a cultural artifact: the AI system. He argues this is analogous to how a playwright might convey a constellation of ideas and experiences to an audience through a play. He then crucially introduces the concept of "authorial affordances" to describe the tools available to the creators of AI systems to convey the story of how an AI systems works to an audience and the concept of "intepretive affordances" to describe the tools available to the users of AI systems to interpret and understand that story.

My proposed doctoral work will show how an AI system can "perform" intelligence to its user and foster an appropriate mental model for how the system works. I aim to do this by: (1) cataloging existing and discovering new "authorial affordances" for an AI system, like conceptual metaphors for AI agents, (2) discovering

Oral Qualifying Exam, Department of Computer Science, University of Toronto

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

what "interpretive affordances" users have for AI systems, like folk theories, and (3) understanding what mental models are fostered by these affordances. My doctoral work will likely combine methods from design theory, Explainable AI, interview studies, media theory, systems building, and art-science collaboration.

1.1 Research Questions

I propose the following set of research questions:

(1) What inaccuracies in mental models for AI systems should be prioritized over others?
(2) What existing "authorial affordances", like conceptual metaphors, are being used to design AI agents today? How and when are these metaphors used?
(3) What existing "interpretive affordances," like cultural narratives and folk theories, are users using to interpret AI agents? How and when are these interpretive affordances used?
(4) What new and appropriate authorial affordances or conceptual metaphors might be created for AI systems?
(5) How do particular pairings of authorial and interpetive affordances influence the mental models users develop for AI systems?

2 LITERATURE REVIEW

This literature review is split into four sections: (1) Core Ideas, AI Expressivity, and Mental Models (2) AI Explainability (XAI), (3) Authorial Affordances: Interface Metaphors, and (4) Interpretive Affordances: Folk Theories. The first section focuses on establishing core concepts for this dissertation. The second, AI Explanability, highlights a large subfield of HCI which is doing research which most strongly resembles that of my own work. Recent work in human-centered AI explainability (HCXAI), in particular, aims to generate explanations that account for the cultural context in which a explanation is provided. The last two sections, Authorial Affordances and Interpretive Affordances, focus on the existing body of work in HCI which describes how designers can convey how systems work to users and how users make sense of unfamiliar systems.

2.1 Core Ideas: AI Literacy, Mental Models, and Expressivity

In this section I focus on core ideas for this proposed dissertation:

(1) *What is AI Literacy?*, Long et Al (2020)
(2) *Mental Models for AI Agents*, Gero et al (2020)
(3) *Expressive AI: A Hybrid Art and Science Practice*, Michael Mateas (2001)

The core ideas that undergird my proposed design framework are (1) that there are a set of competencies which all people should have when they use an AI system, (2) that users can develop useful mental models for AI systems even if they only approximately satisfy these competencies, and (3) that these mental models can be fostered through design. *What is AI Literacy?* lays out these competencies; *Mental Models for AI Agents* describes the value of having mental models for AI systems even if they are not perfectly accurate; *Expressive AI* lays out a model for Human-AI interaction that discusses how designers can guide user understanding using techniques from the arts.

AI literacy refers to a set of competencies that enable individuals to critically evaluate and understand the principles behind AI [22]. This understanding includes the ability to interact effectively with AI, comprehend its implications, recognize its capabilities and limitations, and understand ethical considerations. AI literacy encompasses not just the functional use of AI but also an understanding of its underlying concepts, enabling individuals to be informed users, consumers, and potentially contributors to AI-based technologies.

What is AI Literacy? [22] discusses the growing integration of AI in everyday technology and the general public's limited understanding of these systems. The authors argue the necessity for more research in Human-Computer Interaction (HCI) focused on what competencies users need to effectively interact with and critically evaluate AI and how to design AI technologies that enhance user understanding. They propose a concrete definition of AI literacy, synthesizing interdisciplinary literature into a set of core AI literacy competencies and design considerations for developing learner-centered AI. These insights are organized into a conceptual framework derived thematically from various literature sources, aiming to initiate discussions and guide future AI literacy research within the HCI community. The paper emphasizes the importance of AI literacy in the face of common misconceptions and the potential societal impact of AI technologies, advocating for educational strategies that can foster a deeper understanding of AI among all users.

AI Literacy is of particular importance to this proposed doctoral work because it describes what knowledge and competencies AI lay users need in order to be effective users of AI systems. AI Literacy, as it's currently conceived of in the HCI literature, frames poor AI understanding as an educational problem. Some of the co-authors of *What Is AI Literacy?* [22] have gone on to host workshops and write additional papers that describe more specific educational methods and intervention for helping the public achieve AI literacy [21, 23]. Further educational interventions have been created in recent years, occasionally in the form of games, which attempt to help users understand AI reasoning [25]. However, in my proposed doctoral work, I focus on what can be done to foster reasonable and appropriate mental models for these systems without additional explanation or education. While the AI Literacy paper encourages educational initiatives to achieve AI competency, I aim to foster quick, approximate competencies through culturally-aware design.

This design vision is based on the idea of a "mental model." A mental model is a user's internal representation of how a system works formed through experience, perception, and learning [12]. Mental models help us predict outcomes, solve problems, and develop new concepts. They do not necessarily accurately reflect how a system actually works, but serve a valuable utility [27]. In UX Design, understanding user's mental models helps designers create interfaces that are more intuitive [20].

In Gero's CHI paper talk about the paper *Mental Models of AI Agents in a Cooperative Game Setting* she draws an analogy between mental models and STEM education. In STEM education, a mental model is a student's conception of how a particular scientific concept works, while a conceptual model is the scientific community's consensus about how a particular scientific concept works. Mental

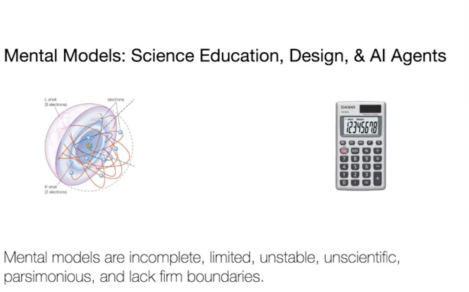

Figure 1: CHI talk of *Mental Models of AI Agents in a Cooperative Game Setting*

models are incomplete, limited, unstable, unscientific and lack firm boundaries. Yet the aim of STEM education is to purposefully foster these mental models because each subsequent mental model will bring the student a step closer to the correct conceptual model. For example, though it is the case that Newtonian physics is no longer taken seriously within the physics community, Newtonian physics continues to be taught in high school STEM education [10]. This is because though Newtonian physics is less accurate than more broadly accepted models in physics like General Relativity and Quantum physics, Newtonian physics provides students a reasonable and accessible understanding of physics at an earlier age. The students can use this understanding, even if it isn't perfect.

The paper *Mental Models of AI Agents in a Cooperative Game Setting*, Gero et al, explores how individuals construct mental models of AI systems in the context of a cooperative word guessing game. The research team conducted think-aloud studies where participants played a game with an AI agent, and through thematic analysis, identified characteristics of the mental models formed by the participants. A large-scale study was also conducted online, where participants played the game with an AI agent and completed a post-game survey to probe their mental models. One key finding was that participants who were more successful in the game tended to have more accurate estimations of the AI agent's abilities. The paper notes that understanding the underlying technology is insufficient for developing appropriate mental models. The paper introduces three key components for the development of large scale mental models: (1) Global behavior, which encompasses actions and reactions of the AI system, and how it might adapt to various situations and other contexts, (2) Local Behaviour, which encompasses responses to AI's decisions and actions at the micro level, and (3) Knowledge Distribution, which refers to what domains of knowledge the AI has access to and how it interprets and uses this information. These components are of particular interest to this proposed doctoral work because they describe what is needed in order for a lay user to form a strong mental model of an AI system. By synthesizing the competencies laid out by Long and the key components to the mental model for an AI agent in Gero et al, my doctoral work will in part aim to map out the space of AI competencies so that designers can make informed decisions about trade offs between utility and accuracy.

Part of the advantage of using mental models to consider Human-AI interaction is that for large AI models like Deep Neural Nets or Large Language Models, there is no broadly accepted conceptual model for all of the system's behaviour [35]. While the techniques for training these models were consciously designed, the behaviour these AI systems exhibit after they have been trained can often be difficult to explain. Therefore, mental models may be the only option for these large options.

In *Expressive AI: A Hybrid Art and Science Practice* by Michael Mateas [24] explores the intersection of artificial intelligence (AI) and art, introducing the concept of "Expressive AI." Pulling from an art-science hybrid practice and drawing an analogy between the dynamic between artist and audience, Mateas proposes thinking of the interaction between AI creator and user like that between play director and audience member. The AI itself can be thought of as an artifact built by creators that communicates a constellation of ideas and experiences to an audience, much like how a play communicates a constellation of ideas and experiences to its audience members. Mateas argues that this contrasts from previous models of building AI, which aim to build intelligent agents which are intelligent independent of a particular observer or cultural context. Rather, in Expressive AI, an agent is intelligent insofar as it can perform as intelligent within a particular cultural context. Mateas notes that the experience of a user of an AI system is therefore influenced by the "authorial affordances" of the creators, the tools the creators have available to articulate a particular story about what the AI is doing, and the "interpretive affordances" of the audience, meaning the resources the user has available to understand such a story. This is significant to my doctoral work because this system aims to change what lay users of AI think and know about these systems by building off of what lay users already know. Lay users already have a variety of cultural touchstones for interpreting AI and behavior more broadly based on their interactions with other people and animals, and previous depictions of AI agents like HAL in 2001 a Space Odyssey or the replicants from Blade Runner. By acknowledging cultural context, this design framework proposes building systems which navigates and builds off this prexisting knowledge in order to foster apprropriate mental models in the user.

There are already several examples of AI systems which, through a combination of authorial and interpretive affordances, shape the user's mental model. The Eliza AI chatbot, created by Joseph Weizenbaum in 1966 at the MIT AI Lab, was a bot which mimicked the behavior of a Rogerian psychologist by taking small key phrases from user responses and turning them into question forms. Users often had lengthy conversations with this very simple AI agent, and perceived it to be more complex than it actually was [30]. Noah-Wardrip Ruin describes three effects which systems can have on user mental models: the Eliza effect, the Tale-Spin effect, and the Sim City effect. The Eliza Effect causes a user to think that a system is more complex than it actually is, the Tale-Spin less complex, and the Sim City effect causes the user to understand the system's internal operations [34]. Recent studies about metaphorical representations of agents have revealed similarly drastic effects depending on user mental models, and will be discussed in more depth later in the literature review [14, 15].

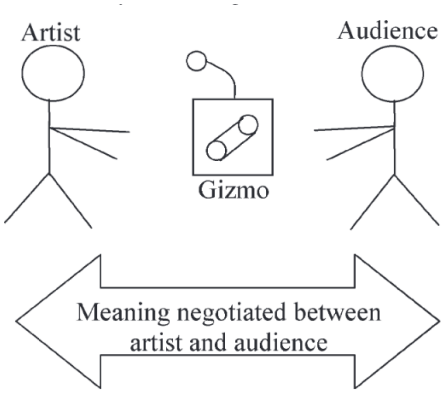

Figure 2: Expressive AI: A Hybrid Art-Science Practice

2.2 Explainable AI (XAI)

In this section I focus on core ideas for this proposed dissertation:

(1) *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*, Liao et al (2020)
(2) *Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI*, Ehsan et al (2023)

Explainable Artificial Intelligence (XAI) refers to methods and techniques in the application of artificial intelligence technology (such as machine learning models) that make the results of the solution understandable and interpretable by human experts. XAI is crucial for critical decision-making processes, especially in sectors like healthcare, finance, and defense, where a human understanding of an AI's decision-making process is necessary for safety, compliance, and trust. This field seeks to create a suite of new AI techniques that produce more explainable models while maintaining a high level of learning performance (accuracy, precision, recall, etc.), and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners. XAI is gaining prominence as a solution to the "black box" problem, helping to bridge the gap between AI's advanced capabilities and the human need for trust and comprehension.

More recently, attention has shifted to the ways in which XAI fails to provide explanations that are useful to humans in practice. This has spurred the creation of another subfield known as Human-Centered Explainable AI (HCXAI) [19]. HCXAI emphasizes the importance of explanations being intuitive and easily digestible for the end-users, regardless of their technical expertise. It involves interdisciplinary research, combining cognitive science, human-computer interaction, design, and ethics to develop AI systems that support collaboration, build trust, and make complex AI systems accessible and comprehensible. By focusing on these aspects, human-centered XAI facilitates more effective human-AI collaboration, ensuring that automated decisions enhance human decision-making processes rather than obscure them with difficult to comprehend machine logic.

Questioning the AI: Informing Design Practices for Explainable AI User Experiences, Liao et al [18], discusses the challenges of HCXAI. The authors conducted interviews with 20 UX and design practitioners working on various AI products to understand the real-world user needs for AI explanations and to identify gaps between existing XAI algorithmic work and practical applications. The paper highlights that while there is a surge in algorithmic work aimed at making AI systems explainable, there is a significant disconnect between these technical explanations and the actual information needs of users, particularly those without deep technical knowledge. One key issue identified is that explanations generated by current XAI approaches often do not satisfy the practical needs of users, such as doctors seeking to understand AI-based diagnostic suggestions. To bridge this gap, the authors advocate for a user-centered approach to XAI, calling for interdisciplinary collaboration and the development of systems that consider the user's perspective and context.

Another paper, *Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI* [7] focuses on the gap between what can be technically supported by XAI and the actual social needs of the users. The authors argue that understanding and addressing this gap is crucial for the effective implementation of XAI, especially as these systems are increasingly deployed in high-stakes domains like healthcare, finance, and criminal justice. To chart this sociotechnical gap systematically, the authors introduce a framework derived from a series of workshops in two different domains: sales and mental health. This framework connects AI guidelines in the context of XAI, providing actionable insights to improve explainability. Once the framework was developed, the research team tried it in a new domain, cybersecurity, to test its efficacy. This paper is useful to my proposed doctoral work in part because it highlights the difference between explainability and actionability. Actionability describes the propensity for an explanation to spur a user to take on further action. This is similar to my proposed design framework, which emphasizes the utility of approximate mental models for AI over accuracy. I can now consider actionability as one of the ways in which a mental model may be more useful.

2.3 Authorial Affordances: Conceptual Metaphors

In this section I focus on work around conceptual metaphors in AI systems:

(1) *Conceptual Metaphors Impact Perceptions of Human-AI Collaboration* Khadp et al (2020)
(2) *Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing* Junge et al (2022)

The idea of a "conceptual" metaphor dates back to the late 1970s, where linguists like George Lakoff, Mark Johnson, and Michael Reddy began to expand the definition of a metaphor beyond its usage as an abnormal part of speech or literary device [17, 29]. These scholars argued that any form of analogous thinking, in which one concept or set of experiences is compared to another, could be considered a metaphor. This definition of a "conceptual" metaphor was much broader and described a fundamental part of learning and thinking. When a new or unfamiliar concept or experience is introduced, others will often try and interpret or

compare this concept or experience with more familiar ones in order to understand it. Conceptual metaphors are a crucial part of the learning process.

Conceptual metaphors are a form of metaphor in Human-Computer Interaction which enables a user to understand how a computing system should be used or works by exploiting knowledge the user has from another domain. Interface metaphors are critical for helping everyday people develop quick, unconscious understandings of digital systems without additional education [26]. 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 [12]. Finding effective conceptual metaphors for AI is therefore critical.

The most famous example of a conceptual metaphor in HCI is the "desktop metaphor." Introduced by Xerox PARC in the 1970s and popularized by Apple's Macintosh in the 1980s, this metaphor enabled lay users to understand the computer as a physical work table [1] by embodying virtual elements (e.g., files, folders, trash cans) that mimicked the physical office environment. This made computers more accessible to the general public at an time in history when they still needed to be won over. Conceptual metaphors can also be important because strong underlying metaphorical structures to UIs can help users adapt to new changes with updates [33].

The concept of an interface is related to that of a skeuomorph, which is an object that retains ornamental design cues from structures that were necessary in previous version of a technology [28]. For example, most "Phone" app on smartphones still use the icon of a phone receiver with a handle and most "Mail" apps use a physical envelope as an icon even though neither receivers with handles nor physical envelopes are used with smartphone technology. These structures—at least when they were first introduced—cued users to interpret cell phones and email using previous experiences with landlines and phyiscal mail;. They are therefore a form of conceptual metaphor conveyed through an interface to help users understand how to approach a new technology.

Conceptual metaphors could be used to help users gain a reasonable understanding of an AI system. A recent example of a novel metaphor for AI comes from the science-fiction author Ted Chiang who recently penned a widely publicized article in *The New Yorker* stating that the Internet is a reasonable metaphor for ChatGPT: it possesses an incredible breadth of knowledge, but is also laden with inaccuracies, follies, and biases [4]. This metaphor doesn't fully capture how ChatGPT works, but reasonably helps the public develop a mental model for the system. ChatGPT's interface could be redesigned in such a fashion that the user got the impression that, rather than speaking to an AI agent, they were communicating with the Internet at large. This could be done with an animation of a browser clicking through web pages that suggests ChatGPT is actively consulting sources on the Internet before it responds, or by presenting ChatGPT's responses in the form of a web page rather than a chat response. While none of these interfaces would perfectly accurately articulate how ChatGPT works, they would give the user a metaphor that might appropriately set their expectations for the capabilities and behavior of the system: namely, that ChatGPT has a tremendous breath of knowledge, but mixed accuracy.

Recent work suggest that conceptual metaphors can have a serious effect on how users perceive AI systems. For example, in the paper *Conceptual Metaphors Impact Perceptions of Human-AI Collaboration*, Khadpe et al [15] ran a study in which users interacted with what they thought were chatbots staged with differing conceptual metaphors but were secretely people. The researchers explore how presenting different metaphors for the 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 are rated higher for usability, adoption intention, and cooperation, despite the common practice of portraying AI agents as highly competent. The researchers advocate for a nuanced approach to communicating AI capabilities, balancing attractive AI agent metaphors with realistic ones to prevent user disappointment and system abandonment. This study underscores the significant impact conceptual and interface metaphors can have on user experience and provides a preliminary set of conceptual metaphors and authorial affordances that could be incorporated into the design framework I aim to create in my proposed doctoral work.

Another recent work, the paper *Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing*, Jung et al [14], adopts the 'Great Chain of Being' framework to systematically explore the impact of non-human metaphors on worker engagement with AI chatbots in crowdsourcing environments. The study focused on how different human and non-human metaphors influence worker engagement, cognitive load, intrinsic motivation, and trust in the agents. The researchers found that metaphorical representations, particularly non-human ones, significantly affect these factors. For instance, using an inorganic object metaphor (like a book) can reduce cognitive load but might negatively impact worker motivation. The study highlights the trade-offs involved in using different metaphors and underscores the significant impact these metaphors can have on user experience. The large difference in user experience suggests that users may also have variant mental models for the behavior of these AI systems. This could be researched further in my proposed doctoral work.

2.4 Interpretive Affordances: Folk Theories

In this section I focus on work with folk theories in HCI:

(1) *What's the Folk Theory? Reasoning about Cyber Social Systems*, French et al (2017)
(2) *How People Form Folk Theories of Social Media Feeds*, DeVito et al (2018)

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 [11]. These theories are not typically based on technical knowledge but are rather constructed from users' everyday experiences, cultural context, and interactions with technology. They are intuitive and change as new information and experiences are introduced. The particular cultural narratives, popular conceptual metaphors, and shared beliefs of a society all strongly influence how folk theories are formed. They may nor may not align with how a particular technology actually works.

Folk theories have a great deal of overlap with mental models as they were originally described by Donald Norman in the HCI literature. But while Norman describes mental models as a representation of how a person believes a system operates through experiences with a technological system [27], folk theories are intuitive explanations users have for systems which come in large part from belief systems and cultural narratives about how those systems work.

In HCI, folk theories are of particular importance because they can help designers create more intuitive and user-friendly technologies by aligning with or building off of existing user beliefs. Folk theories help predict how a user might interact with a system and what aspect of the system they might find confusing. Designers can then use this knowledge to anticipate problems the user may encounter while using their system, and find ways to ease these problems for the user.

In my proposed doctoral work, folk theories for AI are of particular importance because they are a type of interpretive affordance. Folk theories for AI describe how users reason about an AI system intuitively within a particular cultural context. Because the aim of my proposed doctoral work is to allow lay users to gain a reasonable understanding of how an AI system works without additional labor, folk theories most closely approximate the type of reasoning which I anticipate my users having to go through. If I can discover some folk theories for AI systems, than I have discovered some interpretive affordances.

There are a few works in the HCI literature that discuss folk theories. In *What's the Folk Theory? Reasoning About Cyber-Social Systems* French et al [11], the research team focuses on how folk theories effect people's interactions with Twitter and Facebook. The authors introduce a three-phase paradigm for identifying and understanding folk theories. The first phase involves discovering conceptual metaphors that users associate with a system, using a unique survey that encourages participants to propose their metaphors. The second phase identifies underlying folk theories through factor analysis of these metaphors. The third phase specifies the characteristics of each folk theory using semantic differentials. The authors apply this paradigm to study folk theories about the Facebook News Feed and Twitter Feed, and identify four primary folk theories: the rational assistant, the unwanted observer, the transparent platform, and the corporate black box. This paper is of particular use to my proposed doctoral work because it lays out a method via which one can both identify conceptual metaphors for a technological system and then use these conceptual metaphors in order to identify the folk theories for a particular system. One reasonable study to pursue in my proposed doctoral work would be to use the same methods to determine conceptual metaphors for AI and then identify folk theories for AI systems.

Another important paper on folk theories in HCI is the paper *How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation* DeVito et al [5]. In this paper, the authors argue that proprietary algorithms curating social media feeds on platforms like Facebook and Instagram create challenges for users who wish to manage how they present themselves online. These algorithms, often opaque and unpredictable, curate content in a user's feed, affecting which posts are visible to others and potentially influencing perceptions of the poster. For instance, the authors argue that curation algorithms effect the context in which a particular post is perceived and effects what likes and comments the post will get, which in turn effects user's perceptions of the post. The authors conducted a semi-structured interview study with 28 participants to learn how folk theories about social media feeds curation formed. They 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. This paper is of particular use for my proposed doctoral research because it describes how folk theories are formed. Knowing how folk theories are formed allows designers of AI systems to engineer their designs to provoke particular folk theories. This technique could be used to create the type of AI tools I envision, which lay users can understand the capabilities and behaviours of quickly without having to pursue additional education or be provided additional explanations.

3 ANALYSIS

In this section, I synthesize what I have learned from the literature review in order to describe a tentative path forward for this project. My aim in this analysis is to identify specific methods from these papers which may be appropriate to use in my own work. I also compare and contrast the merits of these different subfields and approaches and describe how I see my own project fitting in with this existing body of work. Finally, I propose a few prospective projects which I may pursue in my doctoral research and highlight what contributions I hope to make to the field.

The body of work on "folk theories" in HCI already identifies that there is a relationship between conceptual metaphors, folk theories, and mental models for HCI [11]. Each of these terms describes how a user who is unacquainted with a new technology forms an understanding of how the system works without additional knowledge or explanation. Conceptual metaphors are most closely aligned with the lineage of design theory, and are often conveyed to users through UX features; folk theories are closely aligned with scholarship from communications and media studies which describe how a particular object is depicted in culture; mental models are most closely aligned with cognitive science, which describes the relationship between a user and a technological device as a closed system in which a symbolic mental representation forms in the user from interaction.

All three of these approaches and methods used in these respective studies have something valuable to contribute to my proposed doctoral work. One of the stated goal of my research is to understand how users make sense of systems (interpretive affordances) so that designers of AI systems can design in such a fashion that users can obtain a reasonably accurate understanding of the system without additional work. UX features, cultural context, and the cognition of the user will all influence how users make sense of the system and all must be considered when designing AI systems so that they can easily be understood. Mental models notably develop over the course of a user's experience with a particular system, whereas folk theories are often rooted in exogenous factors, such as narratives about the system in the media and conversations with other users. The former likely has more significance with more experienced users of an AI system, while the latter may have more

influence on less experienced or casual users. One aim of this proposed doctoral work may be to determine the bearing each has on mental models. Is it necessarily the case that more experienced users of an AI system develop a more accurate mental model for how an AI system works? What factors might influence how well a user forms a particular mental model or folk theory over time?

By extending Michael Mateas's model Expressive AI model, I hope my project contributes a design framework for Human-AI interaction that expands our conception of HCXAI. Instead of asking how we can effectively explain AI or educate the public about AI, my work will ask how we can reasonably convey AI to the public through design decisions. This will be the first HCI work of its kind to synthesize concepts from HCXAI and Expressive AI. Until now, Expressive AI has largely been referenced in game studies and discussions of AI art hybrid practices. My work is the first to propose Expressive AI may have utility in modern XAI discourse.

3.1 Future Work

I have ideas for several prospective papers. In one paper, I may catalog the existing conceptual metaphors and folk theories for AI. In order to do this, I will take a cue from a paper previously discussed in this review [11] and use a questionnaire based on the unique probing questionnaire used by that research team and perform a similar factor analysis of the metaphors in order to identify folk theories. Another paper of mine aims to generate new conceptual metaphors for AI by hosting a collaborative workshop with artists and scholars. Pulling from prior work of mine [3] by other HCI practitioners [31], I aim to run a workshop with artists across disciplines and scholars with the aim of having artists devise new conceptual metaphors for AI beyond those that are traditionally part of public discourse around AI (for instance: AI as oracle, AI as assistant, etc..) Lastly, I would like to one of these new metaphors generated from the workshop and incorporate into the UX of an AI system. For instance: suppose that one of the new metaphors that was generated for the curation of posts on a user's feed on Instagram was that of a series of cars in different lanes trying to merge or cut into one lane in the middle, signifying the actual feed. In a future study, I would aim to make a new version of the Instagram app that displays an animation to the user that indicates this is what's happening. For instance, when the user starts up the faux-Instagram app, a large number of posts may appear on different parallel lanes which scroll down and merge into one single feed. After showing this app to users, I would to solicit feedback about how this changed the folk theory or mental model which the users used to explore the research. Lastly, I would like a synthesis paper that consolidates some of these studies to form an overarching set of design principles that allow an AI practitioner to evoke particular mental models.

REFERENCES

[1] 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.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Dashiel Carrera, Gitanjali Bhattacharjee, and Robert Soden. 2023. " We're Not Decorators": Fostering Interdisciplinary Exchange in STEM–Artist Collaborations. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1398–1410.

[4] Ted Chiang. 2023. ChatGPT is a Blurry JPEG of the Web. *The New Yorker* (2023).

[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. https://doi.org/10.1145/3411764.3445188

[7] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[8] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[9] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[10] Paola Ferrarelli and Luca Iocchi. 2021. Learning Newtonian physics through programming robot experiments. *Technology, Knowledge and Learning* 26, 4 (2021), 789–824.

[11] 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).

[12] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 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.

[13] 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).

[14] 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.

[15] 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.

[16] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[17] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

[18] 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.

[19] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[20] Diana Loeffler, Anne Hess, Andreas Maier, Joern Hurtienne, and Hartmut Schmitt. 2013. Developing intuitive user interfaces by integrating users' mental models into requirements engineering. In *27th International BCS Human Computer Interaction Conference (HCI 2013) 27*. 1–10.

[21] Duri Long, Takeria Blunt, and Brian Magerko. 2021. Co-designing AI literacy exhibits for informal learning spaces. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[22] 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.

[23] Duri Long, Anthony Teachey, and Brian Magerko. 2022. Family Learning Talk in AI Literacy Learning Activities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.

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

[25] KATELYN MORRISON, MAYANK JAIN, JESSICA HAMMER, and ADAM PERER. 2023. Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game With a Purpose. (2023).

[26] 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.

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

[28] Tom Page. 2014. Skeuomorphism or flat design: future directions in mobile device User Interface (UI) design education. *International Journal of Mobile Learning and Organisation* 8, 2 (2014), 130–142.

[29] Dan Saffer. 2005. The role of metaphor in interaction design. *Information Architecture Summit* 6 (2005).
[30] Vibhor Sharma, Monika Goyal, and Drishti Malik. 2017. An intelligent behaviour shown by chatbot system. *International Journal of New Technology and Research* 3, 4 (2017), 263312.
[31] Robert Soden, Perrine Hamel, David Lallemant, and James Pierce. 2020. The Disaster and Climate Change Artathon: Staging art/science collaborations in crisis informatics. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1273–1286.
[32] 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.
[33] J. Vanderdonckt and P. Berquin. 1999. Towards a very large model-based approach for user interface development. *Proceedings User Interfaces to Data Intensive Systems* (1999), 76–85. https://doi.org/10.1109/UIDIS.1999.791464
[34] Noah Wardrip-Fruin. 2007. Three Play Effects–Eliza, Tale-Spin, and Sim City. *Digital Humanities* (2007), 1–2.
[35] S Wolfram. 2023. What is chat gpt doing… and why does it work? Wolfram Research.
[36] 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.

Received 30 May 2023