Using folk theories of recommender systems to inform human-centered explainable AI (HCXAI)

L'utilisation de théories populaires des systèmes de recommandation pour éclairer l'IA explicable centrée sur l'humain (HCXAI)

Michael Ridley
University of Guelph

Abstract: This study uses folk theories of the Spotify music recommender system to inform the principles of human-centered explainable AI (HCXAI). The results show that folk theories can reinforce, challenge, and augment these principles facilitating the development of more transparent and explainable recommender systems for the non-expert, lay public.

Keywords: Folk theories, human-centered explainable artificial intelligence (HCXAI), recommender systems, explanations

Résumé : Cette étude utilise les théories populaires du système de recommandation de musique de Spotify pour éclairer les principes de l'IA explicable centrée sur l'humain (HCXAI). Les résultats montrent que les théories populaires peuvent renforcer, remettre en question et augmenter ces principes, facilitant le développement de systèmes de recommandation plus transparents et explicables pour le public non initié et non expert.

Mots clés : Théories populaires, intelligence artificielle explicable centrée sur l'humain (HCXAI), systèmes de recommandation, explications

Introduction

Recommender systems like Facebook, TikTok, Amazon, and Spotify are ubiquitous in our lives. They are the "public face" of machine learning and form part of the "digital everyday" (Kant 2020). However, machine learning systems are complex, opaque, often hidden, and most significantly, consequential. The recommendations, decisions, and predictions of these systems have a material impact on our lives. The black box nature of machine learning has highlighted the need for explanations, resulting in the rise of explainable AI (XAI) as a field of research and development. XAI has largely been focused on system performance rather than the user experience. In

La Revue canadienne des sciences de l'information et de bibliothéconomie, Vol. 46 (2023), No. 2
https://doi.org/10.5206/cjilsrcsib.v46i2.15723

particular, the explanatory needs of the non-expert, lay public have been ignored. The rise of human-centered XAI (HCXAI) is a response to this, putting the user at the center of XAI techniques, strategies, and approaches. While the needs, expectations, and perceptions of users should be used to inform HCXAI, the lack of user studies remains problematic. Folk theories, also known as mental models, are the subjective perceptions people have about how things work. The folk theories of recommender systems can be used to inform how HCXAI can facilitate the development of more transparent and explainable systems.

As greater portions of our lives are "algorithmically mediated" (Anderson 2020), "the danger is not so much in delegating cognitive tasks, but in distancing ourselves from–or in not knowing about–the nature and precise mechanisms of that delegation" (de Mul and van den Berg 2011, 59). Understanding how a system works is a good and perhaps even a necessary thing. However, if users hold relatively intractable ideas about how a system works (i.e., their folk theories), then one way to achieve explainability is to ensure that HCXAI recognize those folk theories. Folk theories can be seen as a bridge that "let us meet the user where they are in terms of understanding and literacy, regardless of how contradictory, sparse, or fragmented these understandings may be" (DeVito 2021, 339:4). Applying the insights from folk theories to the principles of HCXAI can help machine learning developers create better systems, educators address algorithmic literacy, policy makers devise consumer protection, and the public navigate the complexities of using these systems.

The objective of this research is to discover and apply user folk theories to support and enhance the HCXAI principles that guide the development of explanatory systems. The following research questions informed this research:

> RQ#1: What are the folk theories of users that explain how a recommender system works?
>
> RQ#2: Is there a relationship between the folk theories of users and the principles of HCXAI that would facilitate the development of more transparent and explainable recommender systems?

Folk theories

Folk theories refer to "the mental representations that humans use to structure experience" (Gelman and Legare 2011, 380). They allow people to "systematically investigate what [they] believe to be true about particular domains" and provide "a mental structure of possible states of the world that the user can search in order to plan their behavior" (Payne 2003, 152). Importantly, they are "not neutral or passive snapshots of experience; they embody cognitive biases that influence thought and action" (Gelman and Legare 2011, 380).

Folk theories are "surprisingly meager, imprecisely specified, and full of inconsistencies, gaps, and idiosyncratic quirks" (Norman 1983, 8) and yet they are centrally "causal and explanatory" (Gelman and Legare 2011, 380) and "must be functional" (Norman 1983, 7). They are important because of their "utility for the user, rather than their verisimilitude" (Hamilton et al. 2014, 638). Viewed exclusively in the

context of algorithmic systems, Bucher calls these beliefs the "algorithmic imaginary": "the algorithmic imaginary is not to be understood as a false belief or fetish of sorts but, rather, as the way in which people imagine, perceive and experience algorithms and what these imaginations make possible" (Bucher 2017, 31). The algorithmic imaginary describes the "productive and affective power" of users which enables people to act on and influence algorithms in addition to being passive recipients (Bucher 2017, 41).

To align with the focus on recommender systems, this study uses the technologically specific definition of folk theory as "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems, which guide reactions to and behavior towards said systems" (DeVito, Gergle, and Birnholtz 2017, 3165)

XAI and HCXAI

According to the widely referenced US Defence Advanced Research Projects Agency (DARPA) description, the purpose of XAI is for AI systems to have "the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" (DARPA 2016) and to "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (Turek 2016). XAI is a set of strategies, techniques, and processes that include testable and unambiguous proofs, various verification and validation methods that assess influence and veracity, and authorizations that define requirements or mandate auditing (Mueller et al. 2019; Mohseni, Zarei, and Ragan 2021; Speith 2022; Lopes et al. 2022).

Human-centered explainable AI (HCXAI) is a specific area of XAI that responds to the DARPA definition with a focus on explainable AI for the lay, non-expert public. HCXAI has been widely discussed (Chari et al. 2020; Ehsan et al. 2022; Ehsan and Riedl 2020; Liao and Varshney 2022; Shen and Huang 2021; Vaughan and Wallach 2021; Wang et al. 2019). The emergence of HCXAI was motivated by the lack of user studies, a focus on researchers and developers rather than the lay public, an almost exclusive emphasis on the technical aspects and techniques of XAI, the lack of pedagogical methods, the importance of actionable explanations, and the need to reduce the complexity of explanations. Machine learning systems are "often not tested to determine whether the algorithm helps users accomplish any goals" (Mueller et al. 2021). As a result, there have been numerous calls for more user studies of XAI (Burkart and Huber 2021; Miller 2019; Ribera and Lapedriza 2019; Samek and Muller 2019).

Mueller et al. (Mueller et al. 2021) have attempted to consolidate a set of principles to guide HCXAI development. These principles will be used to assess how folk theories can facilitate explainability in recommender systems. While termed "principles," the document more accurately serves as a set of "broad guidelines" (Wang et al. 2019) upon which HCXAI developers can base specific implementations.

Among the HCXAI principles are the importance of context (regarding user objectives, decision consequences, timing, modality, and intended audience), the value of using hybrid explanation methods that complement and extend each other, and the

power of contrastive examples and approaches. Developers are urged to "build explanatory systems, not explanations" recognizing the dynamic nature of intelligent systems and that XAI cannot simply be "one-off." While the principles are centrally about "knowledge transformation and sense-making", they also highlight the importance of "changing previous beliefs and preconceptions."

Prior research

Folk theories and recommender systems

Previous research regarding user folk theories of recommender systems identified differing beliefs about agency and effect. Some user folk theories described a shared agency with the system ("relational," and "transparent") and a largely benign relationship ("practical") (French and Hancock 2017; Ytre-Arne and Moe 2021). Others believed the recommender systems were opaque and exclusively in control exposing users to surveillance and exploitation ("black box," "control system," and "unwanted") (French and Hancock 2017). Siles et al., who studied Spotify, noted that some user folk theories viewed agency as continually contested with users moving between "submission and resistance" (Siles et al. 2020). None of these papers applied their findings to XAI.

Folk theories and HCXAI

Four papers have identified the folk theories of algorithmic systems and made recommendations relevant to the objectives of HCXAI (Villareale and Zhu 2021; Ngo and Krämer 2021; Gentile, Jamieson, and Donmez 2021; Wang et al. 2019). While the folk theories in these papers are less fully developed than those reviewed above, the authors offer specific, if narrow, recommendations for addressing XAI for non-expert, lay users. These include designing for user and system failure in explanatory systems (Villareale and Zhu), focusing explanations on a limited number of specific misunderstandings (Ngo and Krämer), determining algorithmic literacy as a prerequisite to explanations (Gentile et al.), and aligning HCXAI with cognitive factors in decision-making (Wang et al.).

Methodology

Spotify, the music streaming system, was selected to explore the folk theories of recommender systems because of its size, reach, experience, and relative transparency regarding the machine learning processes it uses for music recommendations (Spotify 2021; *Spotify Stream on [Video]* 2021; Fleischer and Snickars 2017). Spotify's recommender techniques are varied, ranging from simple heuristics to commonly used matrix factorization and collaborative filtering, and lastly to state-of-the-art deep learning neural networks and reinforcement learning incorporating extensive data elements (Chodos 2019; Eriksson et al. 2019; Stål 2021; Whitman 2012). Most importantly, machine learning is central to both the business processes and the user experience. Tony Jebara, Vice President of Engineering, describes machine learning as "the heart of everything we do at Spotify" (Jebara 2020).

This study used a multistage process. Users of Spotify were surveyed and subsequently individually interviewed. The elicited folk theories from the survey and interviews were analyzed in conjunction with the principles of HCXAI as presented in Mueller et al. (Mueller et al. 2021). The intent of the user survey and interviews was to obtain the everyday experiences of a convenience sample of Spotify users. The everyday experiences of users reflect the "messiness of real life" (Braun and Clarke 2006) complete with responses that are "complex, nuanced, playful, glib" (Kant 2020).

Using Twitter, 19 Spotify users were recruited. The survey, conducted using the Qualtrics survey tool, collected some baseline information about the usage of Spotify and initial responses to three key questions that were explored in more detail during the interviews. The individual interviews were conducted and recorded using Zoom. Interviews lasted approximately 60 minutes. The survey was analyzed using SPSS (descriptive statistics) and Q Methodology (factor analysis), and the interviews were analyzed using NVivo (thematic analysis).

Spotify user survey

Following the guidance that a survey be "appropriately brief and simple to complete" (Hank, Jordan, and Wildemuth 2009, 257), the user survey consists of 7 closed, contextual questions, two open-ended questions, and a final section of 22 scalar statements. One open-ended question asked participants to describe how they think Spotify uses information to make personalized recommendations. The other open-ended question asked participants about what strategies they might use to influence (i.e., change) those recommendations. The scalar statements ask participants about the influence 22 data elements have on the recommendations they receive from Spotify. Examples of these data elements include, "What I listen to," "How long I listen," "What my friends are listening to," and "What I'm doing while listening." Participants rated the influence of these data signals as "very important," "important," "somewhat important," and "not important." The order of the statements was randomized for each participant. The survey results informed the direction and focus of the user interviews.

Spotify user interviews

As "purposeful conversations," semi-structured interviews are recommended when researchers "are aware that individuals understand the world in varying ways. They want to elicit information on their research topics from each subject's perspective" (Luo and Wildemuth 2009, 233). This process allows the interview to follow the specific context and experience of the participant.

Throughout the interviews, participants were reminded of their responses to survey questions (participants were encouraged to download their responses after they had completed the survey). In all cases users provided additional details and observations when prompted with their survey responses. In some cases, users reconsidered aspects of their response making different assessments of how Spotify used information or how they could shape recommendations.

As part of "member check" validation (Lincoln and Guba 1985), users were given the opportunity to review their interview transcript and make changes, adjustments or

clarifications that would better represent their ideas and opinions. Most participants acknowledged receipt of the transcript and requested no changes. Two participants requested that clarifications and elaborations be included.

Difficulties in eliciting folk theories

Eliciting user folk theories can be difficult. The folk theory research literature reveals a relatively narrow set of approaches, dominated by reflective interviews or questionnaires (Lin et al. 2012; Park and Gittelman 1995; Staggers and Norcio 1993). Such methods can provide incomplete information, contain erroneous information (people say one thing but do another), obscure belief structures, require reasons where users have none, and elicit user responses based on what they think the researchers want to hear (Norman 1983). Questionnaires and interviews rely on retrospective reflection and highlight the concern that "it is not how well users remember their past experience which is of relevance to design but why certain details are reconstructed, and not others" (Doherty and Doherty 2018, 68).

While this study uses the common elicitation practice of self-reporting, it attempted to mitigate some of the concerns of this method. Using two different elicitation methods, separated by several days or weeks, enabled participants to express their views differently and more clearly. It also allowed participants to reflect on the answers from the survey prior to and during the interview. This encouraged a greater degree of self-reflection. The separation of survey and the interview allowed the researcher to consider the survey responses in guiding the direction of the interview based on the known views of the user. The semi-structured interviews with users were designed using a form of laddering (Price 2002; Reynolds and Gutman 1988) that draws the participant into more detailed descriptions of their views. Factual questions were coupled with counterfactual questions, and these were grouped thematically to ensure key aspects of Spotify recommendations were considered.

Findings and discussion

Most of the participants in this study describe themselves as "passionate" or "keen" about music who listen to Spotify every day or most days. They have been a Spotify user for over 12 months, with many having used the system for over five years.

The folk theories were identified through a thematic analysis of the user interviews. Thematic analysis attempts "to identify or examine the *underlying* ideas, assumptions, and conceptualizations—and ideologies—that are theorized as shaping or informing the semantic content of the data" (Braun and Clarke 2006, 84). The thematic analysis of the user interviews proceeded through five stages: data familiarization, coding, theme development, reviewing themes, and defining themes (Braun and Clarke 2006; Terry et al. 2017).

Coding is a process of both data reduction and synthesis (Terry et al. 2017). Using NVivo, an iterative process of coding and recoding was undertaken, focusing on key concepts, consolidation, patterns, and finally the identification of themes (Jackson

2019; Saldaña 2021). Seven folk theories emerged from the themes identified in this analysis.

The folk theories are expressed as verbs (e.g., "Spotify Complies") and are grouped into four categories:

Agency: Compiles, Decides, Dialogues
Conflicting Perceptions: Surveils, Exploits
Black Box: Withholds & Conceals
Feelings: Empathizes

The categories were inferred from the folk theories and brought together similar concepts or highlighted a central idea. The agency category represents theories concerned with power and control. The conflicting perceptions category denotes folk theories that held contradictions. The last two categories, each with a single folk theory, describe the central theme of the respective theories.

Agency: Complies

For some participants, Spotify complies with their deliberate directions and actions: "The only cues that it's getting are the ones that I'm feeding it" (User 3). Users are in control and the resulting recommendations, as the factor analysis revealed, are "about me." The most highly rated data signals in the survey reflected the focus on the active and explicit actions a user takes on the system. The importance of "what I listen to" as a data signal was rated "very important" by 95% of the participants and the importance of "how many times I listen" was rated "very important" by 89% of the participants. As User 19 noted, "Spotify only works because they [listeners] are teaching it to work."

Agency: Decides

For some participants the opposite was true, Spotify decides for the user based on its own objectives. Some users are happy to acquiesce by putting Spotify "on cruise control" and letting the system "take the wheel" (User 5). Others view this more problematically: Spotify "silos me into a particular style" (User 16) and when "in doubt" Spotify will "give me the thing they're being paid to promote" (User 18). Users believe they have little control over Spotify's recommendations: "It's all this giant black box, I don't know anything and there's nothing I can do about it either" (User 13).

Agency: Dialogues

Other participants believe a more cooperative relationship exists with shared agency where Spotify dialogues with the user about recommendations: "I'm feeding it, it feeds me" (User 19). In this perception, Spotify is a "feedback loop" (User 16) which does "a good job of matching my music tastes" (User 12) and is "good at anticipating what kind of music I would be into" (User 14). However, some believe the dialogue is limited and want a richer exchange: "Give me a bigger vocabulary and then make it meaningful. Then prove to me that you've heard me" (User 10).

Conflicting perceptions: Surveils and Exploits

Two folk theories, Spotify Surveils and Spotify Exploits reflect both negative and positive perceptions. Users believe Spotify does extensive data collection about them and view this negatively: "I don't like that they know me, I don't like that they're collecting data, I don't like that they also make assumptions about me that are incorrect. I don't like that they know so much about me" (User 2). However, they also believe that this is necessary: "There's a surrender of personal information that it needs in order to make recommendations that you want. I think that's part of the deal. And that's a world that I've accepted" (User 20).

Users perceive that Spotify exploits them: "My choices, my preferences, are being harvested for their algorithm ... [and this is] the product people are paying for" (User 15). However, as with surveillance, users believe that data collection and sharing is a necessary part of the "bargain" to ensure satisfactory recommendations (User 3). Folk theories are not exclusively positive or negative. As the Spotify Surveils and Spotify Exploits folk theories indicate, perceptions are contextual.

Black box: Withholds and Conceals

Participants believe Spotify is "a complete black box" (User 3). More than just opacity, users perceive there is a deliberate attempt to obscure and restrict (Withholds and Conceals). Users are "not exactly fully cooperating here because Spotify is still doing a lot that we don't necessarily know" (User 13). Whatever users try to do to shape the recommendations "doesn't seem to influence algorithms too much" (User 20). As a result, users want to "see behind the curtain and see what they have me pinned as" (User 11). This folk theory leads some to "game their algorithm" (User 10) as a means of resistance (Kant 2020; Bucher 2018). The Withholds and Conceals folk theory aligns with the Decides folk theory with users believing that "training the algorithm is a lot of effort" (User 3).

Feelings: Empathizes

While prior research has often identified user personification or anthropomorphization of the system (Siles et al. 2020), Spotify users had a more specific belief: Spotify empathizes. The importance of "what I'm feeling" as a data signal that influences recommendations was rated "very important" or "important" by 32% of the participants. When asked if Spotify infers feelings to make recommendations, User 14 responded "Yeah, I think so." Even those who doubt that feelings are inferred by Spotify are not completely sure: "I don't think that it really could get a beat on how I was feeling or anything like that. I wouldn't be surprised if I'm wrong" (User 11).

Folk theories and HCXAI

The elicited folk theories from Spotify users provide a unique view into how users of machine learning-based recommender systems believe they work. These beliefs describe not only how a user understands the system but how they must engage and interact with it. While some of the individual folk theories align with and support each other, taken collectively they do not form a unified whole. They do not aggregate to a

singular theory of recommender systems. Rather these folk theories are separate vectors that sometimes intersect and at other times remain distinct. They contain contradictions and commonalities. However, as a window into the complex user beliefs that inform their interactions with Spotify, they offer insights into how HCXAI systems can more effectively provide machine learning explainability to the non-expert, lay public. The following examples indicate where the folk theories elicited from Spotify users reinforce, challenge, and augment the HCXAI principles.

Where folk theories reinforce HCXAI principles

Self-explanation

The HCXAI principle of "active self-explanation" shifts the balance of power and agency toward the user. By giving the user more information and context, they are empowered to make their own assessments and explanations rather than only receiving an algorithmic explanation. This recognizes that folk theories go through an "exploration and elaborative phase" (Villareale and Zhu 2021) where users are questioning and interrogating the system. The "supplementary data" and "situational data" (Wang et al. 2019) that enable self-explanation align with user desires for a "bigger vocabulary" with which to engage Spotify. Facilitating self-explanation responds to the "Withholds and Conceals" folk theory where users have questions and concerns but don't trust Spotify as a "data company" to be fully forthright in providing explanations or answers.

Design for failure

The HCXAI principles note that trust is damaged when "AI fails in a way that a human would never fail." Users are concerned that Spotify has "all that information" and "should know me" yet fails to provide what the user believes are obvious recommendations. This is experienced by users as a breakdown in the "Dialogues" folk theory where the system is no longer "listening" or they are no longer "training" the system effectively. Users can blame the system or themselves. The "Withholds and Conceals" folk theory attributes failure to a deliberate attempt by Spotify to "push or deter" songs or artists which would favour the interests of the company over the user. The impact of the "Withholds and Conceals" theory and the breakdown of the "Dialogues" theory reinforces the need for "design failure" (Villareale and Zhu 2021) (i.e., designing HCXAI that acknowledges user or system failures) and the importance of the "self-explain" HCXAI principle to encourage users to interrogate their own folk theories.

Explanations are not always necessary

The principle that explanations are "not always necessary" is clear from Spotify users who didn't notice or didn't care about the explanations the system provides. This is consistent with the "Decides" folk theory where users relinquish agency to the system (put it on "cruise control") and don't want or expect an explanation. Spotify's explanations for specific user recommendations are concise and innocuous. Spotify's overall explanation of the recommendation process is provided in a relatively obscure dropdown menu that was only added to the system in 2021. In both cases explanations

are unobtrusive for those who don't care or don't care in the moment. They are examples of "hidden design" features for HCXAI (Ngo and Krämer 2021).

Explanations as verifications

The HCXAI principles acknowledge that an explanation can have "different consequences" and address different needs. The principles emphasize that different methods are required to respond to those needs. For some Spotify users explanations (e.g., "Because you liked ….") were a way to validate the accuracy of the recommendations (e.g., "to see if they were right"). The need for an explanation is not about "why" or "how" but rather for a set of confidence or performance metrics. The folk theories "Decides" and "Withholds and Conceals" reflect the concern about agency and the need for verification. The explanations provided by Spotify do not adequately address these sorts of questions and this is often true with other consumer recommender systems. Spotify users were interested in the performance of the recommender model not just for them but across the user community.

Where folk theories challenge HCXAI principles

Triggered explanations

The HCXAI principles indicate that the need for an explanation is "triggered" by "surprise and violations." However, the folk theories such as "Surveils", and "Exploits" have both positive and negative connotations for users suggesting that the need for an explanation is less a "trigger" than an issue of a threshold or level of intensity that is context specific.

Recommender systems function more effectively when they know more about a user (experienced in some contexts as "surveillance") and when they can use that information to enhance the experience of all users (experienced in some contexts as "exploitation"). Users are aware these are part of the "bargain" with the system.

Neither "Surveillance" nor "Exploitation" is a trigger event, and neither is a violation of expectations. Instead, they are contextual perspectives that diverge from the user's conception of the "platform spirit" (DeVito 2021). Spotify both surveils and exploits, and users are aware and accepting of this. However, in certain contexts and at certain times, concern reaches a temporary threshold or level of intensity where users require an explanation or a justification. The folk theories of Spotify users suggest a more nuanced view of what motivates the need for an explanation.

Multistakeholder contexts

The HCXAI principles are human centered but who are the humans being prioritized? The focus is clearly on the end-user, particularly the non-expert, lay population but there are others involved who should be addressed in the principles. This is especially apparent in recommender systems which are multisided marketplaces with a diverse set of stakeholders. All the folk theories explicitly reference the presence and influence of Spotify, but users also acknowledge that there are other stakeholders beyond Spotify whose interests also influence the system and, indirectly, its explanations (e.g., artists, music companies, music distributors, advertisers, and even other users). The explanatory system is informed by the needs, requirements, and

preferences of these others. As a result, they inform the explanatory system, and mostly consequentially, the nature and extent of the information provided.

The "common ground" being sought is more complex than the principles would suggest. The "explainer" involved is a network or aggregation of many, filtered through the explanatory system to the explainee. While the explanatory system is what the user encounters and engages with, it is a system guided and governed by multiple and diverse stakeholders.

Where folk theories augment HCXAI principles

Consumer protection

Are policy issues relevant to the HCXAI principles? Selbst and Barocas (2018) note that "questions about justifying a model are often just questions about policy in disguise" (1133). The folk theories "Surveils," "Exploits," "Withholds and Conceals," and "Decides" all raise issues of consumer protection: privacy, data protection, risk, and harm. While not all HCXAI implementations relate to consumer applications, the emphasis on the non-expert, lay population suggests a consumer focus. The HCXAI principles are silent on consumer protection.

The principles recognize that explanatory systems must be "accompanied by other things to succeed." A unique suggestion is to develop consumer-facing labels for data and algorithmic models analogous to the nutritional labels mandated by regulation for the food industry (Stoyanovich, Van Bavel, and West 2020). Given that Spotify users enter into a contract with Spotify when they subscribe to the service and agree to the Terms of Use, it is reasonable to position the HCXAI principles within a consumer protection framework.

Right to explanation

While Mueller et al. reference the EU General Data Protection Regulation (GDPR) and the "right to explanation," this right or policy does not appear in the HCXAI principles. The folk theories "Decides" and "Withholds and Conceals" both arise in part because users believe that Spotify is not completely forthcoming. "Decides" reflects the belief that Spotify's recommendation process is opaque. "Withholds and Conceals" amplifies this with the belief that Spotify deliberately hides information from users. Users believe Spotify should provide them with explanations although they described this as an expectation rather than a right. The principles should emphasize, if not a right, a user expectation and a provider obligation.

Manipulation

The folk theories "Surveils," "Exploits," "Empathizes," "Withholds and Conceals," and "Decides" all raise concerns about manipulation in the system and potentially in the explanatory system. The HCXAI principles indicate that explanations should or could be "persuasive" leading to "unjustified trust," but do not caution that they could also be manipulative, deceptive, or coercive. Perhaps this is assumed, but in consumer applications where the typical power imbalance favours the provider not the user, explanations can easily be deployed against the interests of the user. As a result of this

possible manipulation in explanations, researchers have suggested that a "right to explanation" is "not a sufficient condition for ensuring fair, accountable, and transparent use of AI" (Schoeffer, De-Arteaga, and Kuehl 2022, 4). Perceptions of deception and manipulation are evident in the "Decides" and "Withholds and Conceals" folk theories. Given the opacity of the system, users harbour beliefs that Spotify is making recommendations that favour Spotify's financial interests and not the preferences of the user. While not raised by Spotify users, "Withholds and Conceals" suggests the growing concern with "shadow banning" where users of recommender systems or social media believe some of their comments, selections, and preferences are deliberately not collected or are hidden by the system because they reflect undesirable choices or preferences (Savolainen 2022).

However, the folk theory "Empathizes" suggests an emerging avenue for manipulation where the emotional state of users is captured or inferred and utilized to shape or direct their actions. The danger and consequences of such data capture is widely criticized (Crawford 2021; Stark and Hoey 2021).

The principles criticize "transparency" in explanations as insufficient and argue for "apparency" (i.e., systems and explanations that are "readily understood and not hidden") but acknowledge that this is "still not enough." The principles focus on obtaining and sustaining trust but not on the actions or explanations that would undermine that trust. Human-centered XAI should guard against manipulation and deception and the HCXAI principles should articulate this.

Explanatory systems

The HCXAI principles emphasize explanatory systems over explanations. The folk theories that reflect concerns about Spotify ("Decides," "Surveils," "Withholds and Conceals," and "Exploits") challenge confidence and trust in the system. Users don't trust Spotify as a "data company." The lack of trust in Spotify could extend to the explanatory system and raises the following questions: Whose explanatory system is it? Where does the system reside? Does the user have any influence on the system?

While some XAI techniques are post-hoc and independent from the system, in the consumer-facing applications the explanations (i.e., the explanatory systems) are embedded in the application. The APIs available from Spotify and other recommender systems allow access for limited data extraction but nothing sufficient to enable an external explanatory system. Given the focus on user goals, objectives, and contexts, the principles articulate explanatory systems where there is little or no input or influence from users. The principles outline what developers should provide to users through explanatory systems and not how developers and users are partners in the explanatory process.

The principles could, and perhaps should, preference explanatory systems that have some, or even complete, independence from the target service. Enabling an external, independent explanatory system or an explanatory agent would require technical protocols yet to be defined and policy requirements yet to be devised and imposed. However, such a system or agent could be tuned by the user to reflect their algorithmic literacy, experience with other systems, and preferences for types and

extent of explanatory detail. One way to operationalize the "right to explanation" would be by requiring such access in a way similar to the idea of requiring that machine learning systems be "auditable" (Sandvig et al. 2014). A modest attempt at this is the eXplanatory Interactive Learning (XIL) module that lets a user challenge and improve an explanation as well as allow the system to query the user (Weber et al. 2022).

Reviewing the research questions

With respect to RQ#1, the elicited folk theories provide insights into how users believe the Spotify recommender system works. They hold beliefs about agency, control, and the processes the system uses. In some cases, these folk theories reflect conflicting beliefs. Subjective perspectives about surveillance and exploitation also include user beliefs about the need to extensively collect and share data. Users understand the system as a "black box" and believe this opacity is, in part, deliberate.

Regarding RQ#2, the elicited folk theories both supported and challenged aspects of the principles of HCXAI used to inform system design. Principles, such as self-explanation, were aligned with user beliefs while others, such as explanation triggering, where less nuanced than the folk theories revealed. This suggests that the HCXAI principles for system design, as presented by Mueller et al. (2021), could benefit from modifications and augmentations arising from the folk theories. These principles should reflect a more wholistic, sociotechnical perspective that would include multistakeholder influences, policy perspectives such as consumer protection, and stronger safeguards regarding malicious and deceptive practices.

Informed by folk theories, enhanced HCXAI principles would guide system design in directions that would align with how users believe recommenders systems work and how they utilize those systems. This would further directions towards improved user trust and greater system accountability.

Future research

The findings of this study suggest future research directions. HCXAI developers could be studied to evaluate their adoption of these enhanced, folk theory informed principles. This could include not only how the principles effect system design but also how the resulting HCXAI systems are evaluated by users.

The folk theories elicited in this research reflect a more nuanced and contextual understanding of recommender systems than has been described previously in the literature. Future research could utilize these folk theories as the basis for more specific investigations. Issues of contested power and agency raised in this study warrant further study as machine learning systems become increasingly autonomous.

The applicability of folk theories to instructional strategies regarding algorithmic literacy has been identified in prior research (DeVito 2021). The user beliefs elicited in this study offer new avenues for algorithmic literacy pedagogy by highlighting gaps in user understanding as well as revealing concerns users have about broader sociotechnical issues.

Conclusion

Latanya Sweeney, Director of the Public Interest Tech Lab at Harvard, notes that "technology designers are the new policymakers; we didn't elect them, but their decisions determine the rules we live by" (Sweeney 2018). The principles of human-centered explainable AI (HCXAI) were motivated by the "need for use-inspired human-focused guidelines for XAI" (Mueller et al. 2021) that help these new "policymakers" be more responsive to the needs of users, particularly the non-expert, lay public. Folk theories describe the beliefs users hold about how machine learning systems work. They are a window into the way people and technology interact and communicate. Understanding folk theories provides insights into how XAI can be more effectively designed and deployed resulting in machine learning explainability that is more user focused.

The folk theories of Spotify users describe beliefs about agency, power, process, intent, and relationships. Applied to HCXAI, the folk theories support, challenge, and augment the principles of HCXAI. Taken collectively, the folk theories encourage HCXAI to take a broader view of XAI. The questions and concerns implicit in the folk theories indicate that users have explanatory issues that extend beyond the model veracity and authorization. The objective of HCXAI is to move towards a more user-centered, less technically focused XAI. This requires adopting principles that include policy implications, consumer protection issues, and concerns about intention and the possibility of manipulation.

About the author

Michael Ridley is Librarian Emeritus at the University of Guelph where he was for many years the Chief Information Officer (CIO) and Chief Librarian. He holds an MLS (Toronto), MA (UNB), MEd (Toronto), and PhD (Western). Website: www.MichaelRidley.ca.

References

Anderson, Jack. 2020. "Understanding and Interpreting Algorithms: Toward a Hermeneutics of Algorithms." *Media, Culture & Society* 42 (7–8): 1479–94. https://doi.org/10.1177/0163443720919373.

Braun, Virginia, and Victoria Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101. https://doi.org/10.1191/1478088706qp063oa.

Bucher, Taina. 2017. "The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms." *Information, Communication & Society* 20 (1): 30–44. https://doi.org/10.1080/1369118X.2016.1154086.

———. 2018. *If...Then: Algorithmic Power and Politics*. New York: Oxford University Press.

Burkart, Nadia, and Marco F. Huber. 2021. "A Survey on the Explainability of Supervised

Machine Learning." *Journal of Artificial Intelligence Research* 70: 245–317. https://doi.org/10.1613/jair.1.12228.

Chari, Shruthi, Oshani Seneviratne, Daniel M. Gruen, Morgan A. Foreman, Amar K. Das, and Deborah L. McGuinness. 2020. "Explanation Ontology: A Model of Explanations for User-Centered AI." In *The Semantic Web - ISWC 2020*, edited by Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, and Axel Plleres, 228–43. Cham: Springer.

Chodos, Asher Tobin. 2019. "Solving and Dissolving Musical Affection: A Critical Study of Spotify and Automated Music Recommendation in the 21st Century." PhD Dissertation, University of California San Diego. https://escholarship.org/uc/item/2c27z9xk.

Crawford, Kate. 2021. "Time to Regulate AI That Interprets Human Emotions." *Nature* 592 (7853): 167. https://doi.org/10.1038/d41586-021-00868-5.

DARPA. 2016. "Explainable Artificial Intelligence (XAI)." Arlington, VA: DARPA. http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

DeVito, Michael Ann. 2021. "Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 339:1-339:38. https://doi.org/10.1145/3476080.

DeVito, Michael Ann, Darren Gergle, and Jeremy Birnholtz. 2017. "'Algorithms Ruin Everything': #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3163–74. CHI '17. Denver, Colorado: Association for Computing Machinery. https://doi.org/10.1145/3025453.3025659.

Doherty, Kevin, and Gavin Doherty. 2018. "The Construal of Experience in HCI: Understanding Self-Reports." *International Journal of Human-Computer Studies* 110: 63–74. https://doi.org/10.1016/j.ijhcs.2017.10.006.

Ehsan, Upol, and Mark O. Riedl. 2020. "Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach." In *Proceedings of HCI International Conference on Human-Computer Interaction*. Copenhagen, Denmark. http://arxiv.org/abs/2002.01092.

Ehsan, Upol, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. "Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI." In *CHI Conference on Human Factors in Computing Systems*, 1–7. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3491101.3503727.

Eriksson, Maria, Rasmus Fleischer, Anna Joansson, Pelle Snickars, and Patrick Vonderau. 2019. *Spotify Teardown: Inside the Black Box of Streaming Music*. Cambridge MA: MIT Press.

Fleischer, Rasmus, and Pelle Snickars. 2017. "Discovering Spotify." *Culture Unbound* 9 (2): 130–45. https://doi.org/10.3384/cu.2000.1525.1792.

French, Megan, and Jeff Hancock. 2017. "What's the Folk Theory? Reasoning about Cyber-Social Systems." In *67th Annual Conference of the International Communication Association*. San Diego, CA. https://doi.org/10.2139/ssrn.2910571.

Gelman, Susan A., and Christine H. Legare. 2011. "Concepts and Folk Theories." *Annual Review of Anthropology* 40: 379–98. https://doi.org/10.1146/annurev-anthro-081309-145822.

Gentile, Davide, Greg Jamieson, and Birsen Donmez. 2021. "Evaluating Human Understanding in XAI Systems." In *HCXAI '21: ACM CHI Workshop on Human-Centered Perspectives in Explainable AI*. https://hfast.mie.utoronto.ca/wp-content/uploads/HCXAI2021_paper_25.pdf.

Hamilton, Kevin, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. "A Path to Understanding the Effects of Algorithm Awareness." In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 631–42. Toronto, Ontario: Association for Computing Machinery. https://doi.org/10.1145/2559206.2578883.

Hank, Carolyn, Mary Wilkins Jordan, and Barbara M. Wildemuth. 2009. "Survey Research." In *Applications of Social Research Methods to Questions in Information and Library Science*, edited by Barbara M. Wildemuth, 256–69. Westport, Conn.: Libraries Unlimited.

Jackson, Kristi. 2019. *Qualitative Data Analysis with NVivo*. Third edition. London: Sage Publications.

Jebara, Tony. 2020. "For Your Ears Only: Personalizing Spotify Home with Machine Learning." *Spotify Labs* (blog). January 16, 2020. https://labs.spotify.com/2020/01/16/for-your-ears-only-personalizing-spotify-home-with-machine-learning/.

Kant, Tanya. 2020. *Making It Personal: Algorithmic Personalization, Identify, and Everyday Life*. Oxford: Oxford University Press.

Liao, Q. Vera, and Kush R. Varshney. 2022. "Human-Centered Explainable AI (XAI): From Algorithms to User Experiences." http://arxiv.org/abs/2110.10790.

Lin, Jialiu, Shahriyar Amini, Jason I. Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. "Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing." In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 501–10. Pittsburgh, Pennsylvania: ACM. https://doi.org/10.1145/2370216.2370290.

Lincoln, Yvonne S., and Egan G. Guba. 1985. *Naturalistic Inquiry*. Beverly Hills, Calif.: Sage.

Lopes, Pedro, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. "XAI Systems Evaluation: A Review of Human and Computer-Centred Methods." *Applied Sciences* 12 (19). https://doi.org/10.3390/app12199423.

Luo, Lili, and Barbara M. Wildemuth. 2009. "Semistructured Interviews." In *Applications of Social Research Methods to Questions in Information and Library Science*, edited by Barbara M. Wildemuth, 232–41. Westport, Conn.: Libraries Unlimited.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. 2021. "A Multidisciplinary Survey and

Framework for Design and Evaluation of Explainable AI Systems." *ACM Transactions on Interactive Intelligent Systems* 11 (3–4): 24:1-24:45. https://doi.org/10.1145/3387166.

Mueller, Shane T., Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI." *ArXiv*. http://arxiv.org/abs/1902.01876.

Mueller, Shane T., Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. 2021. "Principles of Explanation in Human-AI Systems." In *Explainable Agency in Artificial Intelligence Workshop. AAAI 2021*. AAAI. http://arxiv.org/abs/2102.04972.

Mul, Jos de, and Bibi van den Berg. 2011. "Remote Control: Human Autonomy in the Age of Computer-Mediated Agency." In *Law, Human Agency, and Autonomic Computing*, edited by Mireille Hildebrandt and Antoinette Rouvroy, 46–63. Abingdon: Routledge.

Ngo, Thao, and Nicole Krämer. 2021. "Exploring Folk Theories of Algorithmic News Curation for Explainable Design." *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2021.1987522.

Norman, Donald A. 1983. "Some Observations on Mental Models." In *Mental Models*, edited by Dedre Gentner and Albert L. Stevens, 7–14. New York: Psychology Press.

Park, Ok-Choon, and Stuart Gittelman. 1995. "Dynamic Characteristics of Mental Models and Dynamic Visual Displays." *Instructional Science* 23 (5–6): 303–20. https://doi.org/10.1007/BF00896876.

Payne, Stephen J. 2003. "Users' Mental Models: The Very Ideas." In *HCI Models, Theories, and Frameworks toward a Multidisciplinary Science*, edited by John M. Carroll, 135–56. San Francisco, CA: Morgan Kaufmann.

Price, Bob. 2002. "Laddered Questions and Qualitative Data Research Interviews." *Journal of Advanced Nursing* 37 (3): 273–81. https://doi.org/10.1046/j.1365-2648.2002.02086.x.

Reynolds, Thomas J., and Jonathan Gutman. 1988. "Laddering Theory Method, Analysis, and Interpretation." *Journal of Advertising Research* 28 (1): 11–31.

Ribera, Mireia, and Agata Lapedriza. 2019. "Can We Do Better Explanations? A Proposal of User-Centered Explainable AI." In *Joint Proceedings of the ACM IUI 2019 Workshops*. New York ; ACM. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.

Saldaña, Johnny. 2021. *The Coding Manual for Qualitative Researchers*. 4th ed. Thousand Oaks, Calif.: SAGE Publications.

Samek, Wojciech, and Klaus-Robert Muller. 2019. "Towards Explainable Artificial Intelligence." In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Muller, 5–22. Lecture Notes in Artificial Intelligence 11700. Cham: Springer International Publishing.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014.

"Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." In *Annual Meeting of the International Communication Association*. Seattle, WA. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

Savolainen, Laura. 2022. "The Shadow Banning Controversy: Perceived Governance and Algorithmic Folklore." *Media, Culture & Society*. https://doi.org/10.1177/01634437221077174.

Schoeffer, Jakob, Maria De-Arteaga, and Niklas Kuehl. 2022. "On the Relationship between Explanations, Fairness Perceptions, and Decisions." In *ACM CHI 2022 Workshop on Human-Centered Explainable AI (HCXAI)*. ACM. http://arxiv.org/abs/2204.13156.

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87 (3). https://ir.lawnet.fordham.edu/flr/vol87/iss3/11.

Shen, Hua, and Ting-Hao Huang. 2021. "Explaining the Road Not Taken." In *ACM CHI Workshop of Operationalizing Human-Centered Perspectives in Explainable AI*. ACM. http://arxiv.org/abs/2103.14973.

Siles, Ignacio, Andrés Segura-Castillo, Ricardo Solís, and Mónica Sancho. 2020. "Folk Theories of Algorithmic Recommendations on Spotify: Enacting Data Assemblages in the Global South:" *Big Data & Society*. https://doi.org/10.1177/2053951720923377.

Speith, Timo. 2022. "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–50. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3534639.

Spotify. 2021. "Annual Report." https://s29.q4cdn.com/175625835/files/doc_financials/2021/AR/2021-Spotify-AR.pdf

Spotify Stream on [Video]. 2021. YouTube. https://youtu.be/Vvo-2MrSgFE.

Staggers, Nancy, and A. F. Norcio. 1993. "Mental Models: Concepts for Human-Computer Interaction Research." *International Journal of Man-Machine Studies* 38 (4): 587–605. https://doi.org/10.1006/imms.1993.1028.

Stål, Oskar. 2021. "How Spotify Uses ML to Create the Future of Personalization." In *TransformX*. Scale AI. https://youtu.be/n16LOyba-SE.

Stark, Luke, and Jesse Hoey. 2021. "The Ethics of Emotion in Artificial Intelligence Systems." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 782–93. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445939.

Stoyanovich, Julia, Jay J. Van Bavel, and Tessa V. West. 2020. "The Imperative of Interpretable Machines." *Nature Machine Intelligence* 2 (April): 197–99. https://doi.org/10.1038/s42256-020-0171-8.

Sweeney, Latanya. 2018. "How to Save Democracy and the World." In *Association of*

Computing Machinery (ACM) Conference on Fairness, Accountability, and Transparency. New York: ACM.

Terry, Gareth, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. "Thematic Analysis." In *The SAGE Handbook of Qualitative Research in Psychology*, edited by C. Willig and W. Rogers, 17–36. SAGE Publications. https://10.4135/9781526405555.n2.

Turek, Matt. 2016. "Explainable Artificial Intelligence (XAI)." Arlington, VA: DARPA. https://www.darpa.mil/program/explainable-artificial-intelligence.

Vaughan, Jennifer Wortman, and Hanna Wallach. 2021. "A Human-Centered Agenda for Intelligent Machine Learning." In *Machines We Trust: Perspectives on Dependable AI*, edited by Marcello Pelillo and Teresa Scantabmurlo, 123–38. Cambridge MA: MIT Press.

Villareale, Jennifer, and Jichen Zhu. 2021. "Understanding Mental Models of AI through Player-AI Interaction." In *HCXAI '21: ACM CHI Workshop on Human-Centered Perspectives in Explainable AI*. http://arxiv.org/abs/2103.16168.

Wang, Danding, 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. CHI '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300831.

Weber, Leander, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. 2022. "Beyond Explaining: Opportunities and Challenges of XAI-Based Model Improvement." *ArXiv*. http://arxiv.org/abs/2203.08008.

Whitman, Brian. 2012. "How Music Recommendation Works — and Doesn't Work." *Variogram* (blog). December 11, 2012. https://notes.variogr.am/2012/12/11/how-music-recommendation-works-and-doesnt-work/.

Ytre-Arne, Brita, and Hallvard Moe. 2021. "Folk Theories of Algorithms: Understanding Digital Irritation." *Media, Culture & Society* 43 (5): 807–24. https://doi.org/10.1177/0163443720972314.