TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Corti, L., & Yang, J. (2025). Desirability vs. feasibility: a research through design inquiry of explainable AI. In W. Lawless, R. Mittu, D. Sofge, & M. Brambilla (Eds.), *Bi-directionality in Human-AI Collaborative Systems* (pp. 75-117). Elsevier. https://doi.org/10.1016/B978-0-44-340553-2.00010-1

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

4

Desirability vs. feasibility: a research through design inquiry of explainable AI

Lorenzo Corti and Jie Yang
Delft University of Technology, Delft, Netherlands

Abstract

Explainability is a key property towards the trustworthiness of AI systems, and scholars have approached the problem from both system- and human-centered angles. However, thus far, these two camps have been researching explainability in isolation without fully accounting for the socio-technical nature of AI systems and the bidirectional effects that emerge when interacting with them. In this chapter, we adopt Research through Design (RtD) – a systematic, reflexive, and documented research approach rooted in design practice – to question our understanding of explainability and how it should be researched. We revisit the explainable AI literature to surface challenges and outline opportunities around bidirectionality, the role of human knowledge, and the availability and temporal dynamics of explanations. We exemplify challenges and opportunities through a speculative exercise grounded in real-world use cases of explainable AI systems. Aligned with RtD, we hope this work helps researchers in explainable AI whom we invite to engage in interdisciplinary and well-documented research processes.

Keywords

Research through Design (RtD), Explainable AI, Human factors, System engineering

4.1 Introduction

Artificial Intelligence (AI) has evolved considerably and can provide benefits across diverse domains. Recently, generative AI systems have risen to scholars' and the broader public's attention for their human-like content-generation capabilities and seamless interaction modalities. Nevertheless, despite the technical advances, such systems remain crippled by reliability (Alcorn et al.,

Bi-directionality in Human-AI Collaborative Systems. https://doi.org/10.1016/B978-0-44-340553-2.00010-1
Copyright © 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

2019; Tocchetti et al., 2025) and transparency (Topol, 2019; Rahwan et al., 2019) issues. Such issues of AI systems stem from several elements like architectural choices (and their formalization), the data used to build these systems, and the evaluation frameworks used to assess their goodness. In turn, these discretionary choices lead to these issues manifesting differently in real-world situations: hallucinated outputs (Ji et al., 2023), poor robustness (Tocchetti et al., 2025), and societal harms such as discrimination (Mehrabi et al., 2021).

In the quest towards responsible and trustworthy AI systems, Explainable AI (XAI) has recently garnered increased attention as a means to assist efforts aimed at addressing the issues, and their manifestations, above. Explanations play a crucial role in informing users about different aspects of AI systems and tools they use. Currently, one can identify two main camps within XAI research: one focused on explaining the underlying algorithms of AI systems and the other focused on researching the humans using such systems. As a result, research works around explainability are diverse and focused on aspects such as designing technical solutions to explain individual outputs of AI systems (Guidotti et al., 2018b; Barredo Arrieta et al., 2020; Guidotti, 2022), promote interpretability (Carvalho et al., 2019), foster appropriate reliance (He et al., 2023), and enable contestation of algorithmic decisions (Yurrita et al., 2023).

Crucially, while there is consensus around explanations being beneficial in diverse settings, these two subcommunities value different, but complementary, facets of explanations but do not actively exchange insights on those (Chen, 2024; Abdul et al., 2018). For instance, the faithfulness of explanations is highly prioritized in system-centered research and has a direct impact on human-centered studies (e.g., affecting users' perceptions and reliance on "imperfect" explanations). Yet, human-centered works commonly employ off-the-shelf explainers for which the "unfaithfulness" has been documented (Bordt et al., 2022). By the same token, the dual holds as well. System-centered works propose methods that lack purpose (Freiesleben and König, 2023) and that, as a consequence, lead to lackluster evaluation protocols and misalignment with user needs (potentially) already surfaced in human-centered discussions.

Our goal is not to downplay the ways research around explainable AI is conducted. Instead, we highlight the lack of active knowledge exchange between the two camps in explainable AI research. Beyond individual system- and human-centered challenges, we believe this gap to be a rather cumbersome and limiting research on explainable AI, a problem that naturally involves humans and

machines. Research and application of explainable AI solutions have progressed at a drastically slower pace compared to those of AI models. For instance, successes around generative large language models have overshadowed efforts around understanding and explaining what such models might be doing. In turn, this unfounded reliance on believable AI-generated outputs raises questions related to the bidirectional effects – humans influencing AI systems and vice versa – that might emerge during human–AI interactions, how these effects could be identified and, if necessary, mitigated to achieve a certain objective.

In this chapter, we adopt a *Research through Design* (RtD) lens (Zimmerman and Forlizzi, 2014) to reflect and reframe the problem of producing explanations for AI systems (Section 4.2). We revisit the interdisciplinary XAI literature, characterizing typical facets of the system- and human-centered research on AI explanations (Section 4.3). Based on findings from the literature, we then discuss the central challenges of explainable AI – both around its research and its application (Section 4.4). Finally, we draft a series of research questions for XAI research encompassing central challenges faced by existing approaches and opportunities for computational means yet to be leveraged (Section 4.5). We complement such considerations with a speculative analysis of potential use cases for explainable AI systems; highlighting both benefits and hindrances.

We do not seek to provide an exhaustive and definitive inquiry of what explainable AI is, should be, or could be. Rather, in line with the distinct and continual reinterpretation process of RtD, we invite other scholars to join us, take up this work, and foster explainable AI.

4.2 Research through Design (RtD) and Explainable AI

Research through Design (RtD) is an approach to conducting research by making use of methods and processes customary to design practice (Zimmerman and Forlizzi, 2014). Compared to other design approaches, such as Human-Centered Design (Cooley, 1989), RtD emphasizes the continuous reinterpretation and reframing of a problem through iteratively proposing solutions and critiquing them. Nevertheless, the ultimate goal of RtD is to generate new knowledge (e.g., perspectives, implications, and methods) by going beyond the sole creation of artifacts. A critical difference – which in our view makes RtD a unique lens to discuss XAI research – with respect to other design practices is the more *systematic nature* and *reflexivity* in the reinterpretation of the understanding

of a problem. Detailed documentation and explanation of rationales accompany the entire process, aiding the reproducibility of the process while retaining freedom around the resulting artefact and knowledge. That is, being able to reproduce the process does not necessarily lead to the same outcome.

While explaining AI systems is an interdisciplinary effort, the two camps of system- and human-centered research face major challenges in achieving the grand goals of explainable AI when continuing to operate in silos (Chen, 2024; Abdul et al., 2018). For one, research on explainable AI has been led by a techno-centric perspective, fostering the development of a wide range of computational explainers, but overlooking the human side of explanations during the design phases. Yet, users are diverse and their explainability needs are disparate and mutable (Tonekaboni et al., 2019; Corti et al., 2024). HCI scholars, in particular, have advocated putting humans back at the center of explainable AI research (Ehsan et al., 2021a, 2023a) by building on top of existing theories from philosophy, social science, and design. While these works have been instrumental in expanding the definition of explainable AI beyond output explanations typical of techno-centric contributions, they employ only a subset of what works on computational methods for XAI that have been proposed.

In our view, Research through Design (RtD) is inherently bidirectional. Its continuous reinterpretation and reframing of a problem hinges on and benefits from a mutual understanding and interdependence between the "researcher" and what or who is "researched", within and across system- and human-centered perspectives. We align ourselves with Dorst and Cross (2001) in that problem and solution spaces (of explainable AI in this case) should coevolve: system-centered contributions might help human-centered works and, in turn, the latter might generate insights that can inform and reshape the former or even suggest all new computational methods.

4.2.1 Method of this paper

In this chapter, we adopt a Research through Design (RtD) lens to discuss challenges and opportunities around XAI research. Since RtD does not impose any particular method or practice, we conduct four main steps in our inquiry.

First, we review the interdisciplinary contributions of scholarly research around explainable AI. Second, we highlight and discuss some pressing challenges around researching and implementing in practice explainable AI systems and solutions. Third, we open up the discussion to the broader research community and draft a

series of research questions for XAI research encompassing central challenges faced by existing approaches and opportunities and computational means. Finally, we engage in a speculative investigation of a selection of real-world use cases where explainable AI systems might be desired, highlighting both benefits and hindrances.

4.3 Related work

To frame our inquiry, we provide an overview of prior work – both system- and human-centered – on explainable AI. Broader accounts of explanations and explainable AI (e.g., from a philosophical viewpoint) are addressed in relation to the individual works included in this section.

4.3.1 System-centered works in Explainable AI

The concept of explainability first emerged in works on expert systems (Buchanan and Shortliffe, 1984). Recent advances around subsymbolic AI approaches (e.g., Deep Learning) fueled interest in explainability as, generally speaking, such approaches favor performance over transparency (Bommasani et al., 2021).[1] Headlining examples of such systems are ChatGPT (OpenAI, 2023) and DALL-E 3 (Betker et al., 2023), as well as other large generative models, e.g., LLaMa 2 (Meta, 2023) and Mistral (Jiang et al., 2023). As a result, such systems have proliferated in diverse domains such as healthcare (Topol, 2019) and finance (Königstorfer and Thalmann, 2020).

Nowadays, research around explainable AI remains mostly system-centered – focused on describing individual outputs and inference (or generation) processes of an AI system – where a plethora of *explainers* have been proposed (Barredo Arrieta et al., 2020; Guidotti et al., 2018b; Guidotti, 2022).[2] Prior works cover *local* (i.e., for individual data instances) and *global* (i.e., for classes of data instances) explainers, either in post-hoc (i.e., without modifying an AI model) (Selvaraju et al., 2020) or self-explaining (i.e., built-in within an AI model) (Zhang et al., 2021; Chen et al., 2018) fashions. Practically, off-the-shelf XAI solutions can disclose the

[1] Subsymbolic AI approaches are more commonly referred to as *black-boxes*, i.e., approaches that do not disclose their internal mechanisms. One thing to notice is that AI approaches regarded as *white-box* (i.e., AI models whose parameters have a precise meaning) are not exempt from interpretability concerns as the number of parameters increases. For instance, decision trees may become harder to interpret as their depth increases.

[2] Explainer: computational method to generate explanations for a given AI model.

importance of individual input features (i.e., feature attribution) (Ribeiro et al., 2016; Selvaraju et al., 2020), identify influential or prototypical training data instances (Koh and Liang, 2017; Guo et al., 2021; Chen et al., 2019; Nauta et al., 2021), provide indications about which input features to change (and by how much) such that different model outputs are obtained (i.e., counterfactuals) (Wachter et al., 2018; Goyal et al., 2019), generate human-like concepts (Ghorbani et al., 2019; Balayn et al., 2021), and produce rule-based explanations (Ribeiro et al., 2018; Guidotti et al., 2018a). Given the broad range of explainers, prior work has formulated different accounts around what makes for a "good" explanation. Such accounts have been operationalized in a selection of properties (Miller, 2019; Carvalho et al., 2019; Sokol and Flach, 2020; Alvarez Melis and Jaakkola, 2018; Kulesza et al., 2015) concerning both system-centered (e.g., fidelity and stability of explanations) and human-centered (e.g., comprehensibility and actionability of explanations) aspects.

However, despite the abundant and insightful body of research, we lack potential solutions to several issues. Explainability methods suffer from brittleness (Slack et al., 2020), intramethod disagreement (Krishna et al., 2022), and misalignment with recipients' (both experts and laypeople) interpretations (Zhang et al., 2019; Balayn et al., 2022b; Alqaraawi et al., 2020).

4.3.2 Human-centered works in Explainable AI

The Human–Computer Interaction (HCI) community (e.g., CHI, CSCW, IUI, etc.) argues for a broader definition of explainability, one that focuses on the recipients of explanations (Green and Viljoen, 2020; Langer et al., 2021; Mohseni et al., 2021) and embraces the socially-situated nature of AI systems (Ehsan and Riedl, 2020; Ehsan et al., 2021a, 2023a). HCI scholars embrace interdisciplinarity to bridge theories and frameworks from cognitive psychology (Lombrozo, 2010, 2012), social sciences (Miller, 2019), design (Yildirim et al., 2022), and philosophy (Buijsman, 2022).

A nascent research direction within HCI is that of *Human-centered Explainable AI* (HCXAI).[3] HCXAI researchers aim to gain an understanding of *who* the recipients of explanations are (Ehsan and Riedl, 2020), such that different perspectives are adequately operationalized at the conceptual, methodological, and technical levels. For that to happen, HCXAI generally rejects the idea of requiring a single definition of explainability (Ehsan and Riedl, 2024) in favor of a *pluralist* view over different social groups and how

[3]The CHI community actively engaged in this topic through workshops (Ehsan et al., 2021b, 2022, 2023b, 2024b).

these might interpret technological artifacts differently (i.e., interpretive flexibility (Bijker et al., 1987)). Prior works in this space rest on the rich literature around framing "XAI stakeholders" (Barredo Arrieta et al., 2020; Preece et al., 2018; Tomsett et al., 2018; Mohseni et al., 2021; Langer et al., 2021) and incorporate reflexive design practices (Ehsan and Riedl, 2020) and prior discussions highlighting the relevance of explanation recipients' context (Carvalho et al., 2019; Miller, 2019; Sokol and Flach, 2020).

Through this lens, we can relate several prior works investigating the technical affordances and end-users of XAI systems (Ehsan et al., 2023a) to the principles of HCXAI. Works targeting developers and practitioners include support tools like Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2021) which provide structures for documenting their work on AI models, datasets, their capabilities, and implications. Liao et al. (2020) proposed a question bank covering prototypical questions around explainability, while Balayn et al. (2024) investigated the diverse roles involved in AI supply chains, their explainability needs and the reasons behind those. On the other hand, examples of works targeting lay users include data-centric explanations (Anik and Bunt, 2021) and empirical studies around the relative importance of evaluation properties of explanations (Vera Liao et al., 2022). Finally, Langer et al. (2021) and Subramonyam et al. (2022) propose frameworks to aid interdisciplinary research and communication around AI systems.

In conjunction with such works, others focused on understanding the XAI needs of end-users. Kim et al. (2023) enquired about the end-users of a real bird identification application and surfaced needs related to improving human–AI collaboration. Corti et al. (2024) focused on pulmonary medicine and inquired how clinicians' uses and purposes for explainable AI evolve throughout patient care, highlighting the relationships between patient conditions, clinicians' desire to maintain agency over the final decisions, and properties of explanations from XAI literature. Similarly, Cai et al. (2019b) investigated the needs of pathologists around AI-based tools for diagnostic decision-making and, later, Cai et al. (2019c) compiled pathologists' information needs emerging during the onboarding phases of prospective AI systems (e.g., technical capabilities and limitations). Instead, Rostamzadeh et al. (20022) adapted Datasheets for Datasets (Gebru et al., 2021) for the documentation of healthcare datasets. Finally, Tonekaboni et al. (2019) explored clinicians' relationship with, and trust in, AI tools for cardiac arrest and triage prediction.

4.4 Challenges of Explainable AI

We now discuss challenges – both system-centered (**SC**) and human-centered (**HC**) – in researching and applying explainable AI systems. We leave for later (Section 4.6) the discussion about their bidirectional interplay and manifestations in specific use cases for explainable AI.

SC1 XAI Requires Full Model Access (Section 4.4.1)
SC2 Applicability of XAI Methods (Section 4.4.2)
SC3 Optimizing for Faithfulness is Hazardous (Section 4.4.3)
HC1 Describing vs. Explaining (Section 4.4.4)
HC2 Understanding XAI Recipients (Section 4.4.5)
HC3 AI Literacy Affects Explanation Understanding (Section 4.4.6)

4.4.1 SC1: XAI requires full model access

Recent trends popularized by industry around the release of large language models (LLMs), black-box models that learn a probability distribution over (sequences of) words, have made this a less common practice.[4] Indeed, in 2023 we saw an increase in the number of open-source (foundation) models being released (Bommasani et al., 2023; Liesenfeld and Dingemanse, 2024). While scholars increasingly call for the openness of AI models and systems, more often than not such systems are closed-source.[5] This lack of openness makes it all the more difficult to understand and assess the robustness of newly released AI models and systems.

For explainability, this is even more crucial as most methods require full model access to be applied.[6] While this general behavior might not represent an issue for developers within the companies releasing such models, it severely limits the work of developers and practitioners on the outside and the agency of users (e.g., general consumers) that leverage those systems in their daily lives – the system creators decide whether or not explanations are present. Yet, to this day, the risks associated with the real-world use of AI systems are difficult to analyze. We currently lack rigorous ways to evaluate and audit AI systems, but the general lack of transparency around AI systems is the first of many causes (e.g., the widespread benchmarking culture (Gururaja et al., 2023)). As a result, we assess AI systems and models in a reactionary manner after incidents have already occurred. Explainable AI methods

[4] In their most current form, LLMs are transformer-based models (Vaswani et al., 2017) capable of generating human-like text.
[5] Meta and BigScience are among the few currently releasing models openly and making use of documentation tools (e.g., (Mitchell et al., 2019)).
[6] There have been attempts at constructing explanations without full model access but those have insofar been either limited or not representative of a model's inner workings (Ribeiro et al., 2020; Madsen et al., 2024).

are viable tools to assist such endeavors (e.g., for debugging purposes (Lertvittayakumjorn and Toni, 2021)) provided they can be applied.

4.4.2 SC2: applicability of XAI methods

The Explainable AI literature comprises hundreds of methods, each with its own peculiarities and purposes. Here we bring attention to the applicability conditions of such methods: model architecture and model classes.

Model architecture

One dimension that can be used to classify XAI methods is whether they are *model-specific* or *model-agnostic*. Model-specific explainability methods work based on specific information about the architecture of the model to which they are applied. For instance, methods like GradCAM (Selvaraju et al., 2020) only work with convolutional neural networks. On the other hand, model-agnostic explainability methods can be applied to any given AI model to yield explanations for its output. A notable example is LIME (Ribeiro et al., 2016), a perturbation-based method that repeatedly interrogates an AI model with perturbed input, collects the output, and builds surrogate models over that data. While similar methods can be used *virtually* with any model, the underlying surrogate models can be relatively simple and thus unfit to represent the complex non-linearities of certain AI models to be explained.

Model class

In the XAI literature, to the best of our knowledge, model classes are seldom discussed, and the majority of prior works focus on discriminative models (i.e., models that are meant to learn a decision boundary over the data). As of now, more recent (large) generative models remain relatively excluded with few works attempting to adapt XAI methods meant for discriminative models to generative, sequence-to-sequence models (Sarti et al., 2023) for Natural Language Processing (NLP). Continuing with NLP, a crucial problem is that while model-agnostic methods could be used to explain generative language models, textual perturbations (e.g., token-level perturbations) often lead to invalid sentences. Indeed, prior work has studied the generation of false positive input sentences for assessing the adversarial robustness of language models (di Giovanni and Brambilla, 2021), but has done so in a classification setting using discriminative language models. For other data types and modalities (e.g., vision), similar questions about the validity of the perturbed samples still apply.

4.4.3 SC3: optimizing for faithfulness is hazardous

The faithfulness (or fidelity) of explanations is concerned with the truthfulness of explanations (and their subcomponents) with respect to the underlying model. It is a commonly used evaluation criterion for post-hoc explanations (e.g., LIME (Ribeiro et al., 2016)) and arguably one that has received the most attention in system-centered works. On the one hand, it is reasonable to give importance to the faithfulness of explanations to guarantee that the information surfaced by an explainer, and later provided to recipients of explanations, is true. However, prioritizing faithfulness could generate tensions with other explanation properties and user needs. For instance, if for an explainability recipient, carrying out a certain task in a certain context where the compactness of explanations is a key property, the faithfulness of the explanation could be compromised. Given a compact explanation, e_c, this might surface features that are faithful only to certain parts of the underlying AI model. However, despite possibly being preferred in practice, it might "score" lower compared to an explanation, e, that relates to every part of the underlying model. As AI models grow more and more intricate and complex, there is an increasing need for weighing the faithfulness of explanations with respect to the number of model parameters and their (usually) dense activation.

Finally, explanations that are faithful to an underlying AI model might not be indicative of the functioning of the overall AI system. As Sculley et al. (2015) discuss, AI models constitute a small part of the overall construction and functioning of an AI system; which is the artefact that stakeholders (not in a development role) usually interact with.

4.4.4 HC1: describing vs. explaining

The first human-centered challenge we would like to discuss is the difference between describing and explaining a topic. Describing something means *providing details about a situation or idea*. A description is normally concerned with answering questions related to the who, what, when, and where. For instance, in a history class, descriptive questions could be "What is Napoleon Bonaparte known for?" or "Where was Napoleon exiled before his death?" On the other hand, explaining entails *providing reasons and evidence about a situation or idea to someone*. An explanation is concerned with questions of why and how. Continuing with the example of Napoleon Bonaparte, similar questions could be "Why did Napoleon fail to invade Russia in 1812?" or "How did Napoleon become emperor of France?"

If we draw parallels between the questions above and what XAI methods can do, we may see that such methods concern the generation of descriptions for the most part. What are often touted in the literature as explanations are simply descriptive reconstructions of what an AI system might be giving more importance to. In psychology and social science literature, on the other hand, several theories of explanations have been introduced, highlighting diverse aspects of explanations such as selectivity (Lombrozo and Liquin, 2023) or stimulating curiosity (Liquin and Lombrozo, 2020) that go beyond what we can currently achieve.

We see this as a central misconception in explainable AI, creating a misalignment between what explainable AI can offer and what it claims to be. Take the example of a physics textbook: how would students be able to build the expected understanding about, e.g., fluid statics if the textbook were to only include formulas? While one may say individual formulas are faithful to a given phenomenon, what makes a textbook useful is the explanation text surrounding the algorithm being written in a language the students can understand.

4.4.5 HC2: understanding XAI recipients

The field of XAI has had the concept of *audience* of the explanations in mind from its inception (Barredo Arrieta et al., 2020; Hind, 2019). The interaction with AI systems and the digestion of information supplied through AI explanations are social processes in nature. However, AI explanations are often used as ends and not means to understand user needs. System-centered works, particularly, often overlook the human factors conditioning the use of explanations (e.g., understandability) or reduce them to a series of desirable properties of explanations (e.g., compactness) (Vera Liao et al., 2022). Such practices can lead to a series of disadvantages: misalignment with recipients' expectations (Hong et al., 2020), high cognitive load (Jesus et al., 2021), and overreliance on the explanations (Kaur et al., 2020) to mention a few.

Understanding users and their context is, on the other hand, one the tenets of several HCI works. Yet, studies around explainable AI solutions and approaches often question the effects explanations have with respect to the completion of specific tasks. Few works have inquired about user needs ahead of prospective implementations by using explanations as *conceptual artifacts* (Gaver and Martin, 2000) from speculative design practice to prompt users about desired explanation types, contents, and modalities (Corti et al., 2024; Kim et al., 2023; Tonekaboni et al., 2019). While such

works provide insights into how people use existing explainability methods, the implications of such findings for the design of future explainers are not always straightforward. Discussions often revolve around what is desired, drafting recommendations or guidelines of what should be done, but do not necessarily relate or connect to the means (computational or not) required to achieve the "desired" explanations.

Overall, understanding the explainability needs of certain user groups is no small feat. This can be done during the early stages of requirements collection or later in the development to understand how recipients use AI explanations. Yet, few prior works can be found at the intersection of the two camps within XAI research adopting an iterative loop of user feedback to refine AI explanations based on both user requirements and usage (e.g., (Xie et al., 2020)).

4.4.6 HC3: AI literacy affects explanation understanding

AI literacy inevitably conditions understanding of explanations; it has been a recurring research topic within the HCI community (Long and Magerko, 2020; Laupichler et al., 2022; Wood et al., 2021). However, empirical studies are often underpinned by the assumption that those engaging in the explanations have a sufficient level of knowledge that enables them to interpret and understand the explanations correctly. In contrast to those, prior work has shown that this is an unrealistic assumption. Explanations vary across contexts, users' backgrounds (Kim et al., 2023), and tasks (Corti et al., 2024). A lack of an adequate level of AI literacy can cause an overreliance on an AI system and the explanations that come with it.

We view *causality* as a crucial point to stress in this case. Oftentimes, there is not a real and justified connection between the output of an AI system and domain knowledge, or domain-specific processes (Holzinger et al., 2019). AI systems do not hold any communicative intent (Bender and Koller, 2020) and, as AI systems advance technically, the risk of plausible-looking outputs and explanations increases as well. As Bender and Koller (2020) argue, systems like large language models solely built on *form*, do not have a way to produce *meaning*. In this sense, prior HCI work has investigated the potential effects of biases and misunderstandings of AI's capabilities (Croskerry, 2013; Bach et al., 2023; Fogliato et al., 2022). However, despite ongoing research efforts in causal ML and XAI (Guidotti, 2022; Biswas et al., 2022), it is still an elusive concept within the current generation of AI systems.

Nowadays, in practice, we see more and more institutions trying to incorporate some degree of technical and AI-related training for their staff. For instance, some healthcare institutions provide residency training (e.g., for radiologists) where AI training is included. Reasonably, broader education support for AI literacy is currently lacking (Karaca et al., 2021), and its operationalization is a significant effort that requires balancing domain-specific and technical knowledge.

4.5 Opportunities and methods

In this section, we will provide what we believe to be a series of opportunities for advancing research on explainable AI. Specifically, we discuss bidirectionality in human–AI interactions, the role human knowledge has on model explainability, the accessibility of AI explanations, and the temporal dynamics of explanations. We conclude by turning to the broader research community with an interdisciplinary call for action.

4.5.1 Bidirectionality as a key principle to design explanations

Bidirectionality in human–AI interactions is fundamental for accomplishing complex tasks and achieving high team performance. It relies on the interdependence of skills, a mutually recognized awareness, shared communication, and intricate teamwork (Butlin et al., 2023; Sliwa, 2021; Holder et al., 2021). Recent advances in Large Language Models have caused significant changes in how one can create plausible and believable text (Park et al., 2023), which for all intents and purposes can also enable AI systems to understand and communicate through human language. However, such advances are often due to an overparameterization of models and a prioritization of fluency of the output text even though issues of robustness persist (Ji et al., 2023; Wang et al., 2023).

Explanations can act as support tools when interacting with AI systems to aid decision-making processes (e.g., fostering credibility) by further disclosing system affordances to humans (Fig. 4.1). However, current explainers and explanations do not abide by the notion of bidirectionality but rather provide information, usually, without offering a chance for humans to follow up. We find this especially crucial when multiple human actors are involved in the interaction with an AI system.

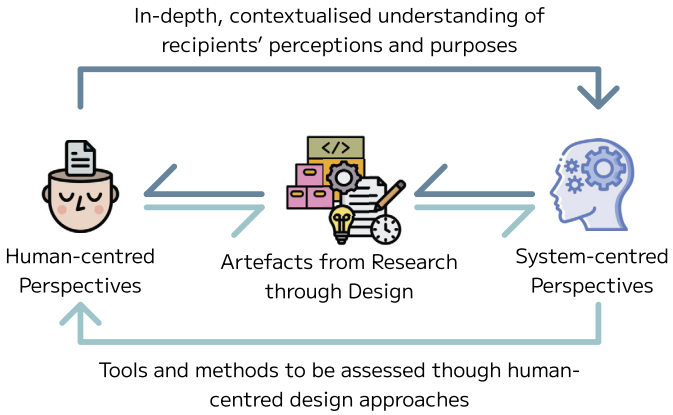


Figure 4.1. Applying Research through Design to researching explainable AI naturally creates a space for bidirectional interplay – underexplored in current explainable AI literature – between recipients of explanations (subjects of human-centered inquiries) and AI systems (built according to system-centered engineering practices).

4.5.1.1 Informational and transactional needs

When interacting with AI systems, humans engage in an information-seeking process regardless of whether this happens in a collaborative setting or not (Sivaraman et al., 2023; Shah and Bender, 2024). In this sense, explanations for AI systems give the opportunity to go beyond the individual outputs of AI systems, even for more articulate large generative models. Different patterns can arise from these interactions. For instance, if *alignment* between human(s) and AI occurs, the explanations can serve as evidence for one's decision-making process. If, instead, *misalignment* is present, external factors like time constraints and pressure could condition the interpretation – or willingness to interpret – of AI explanations. Cases of *partial alignment* represent more nuanced and interesting scenarios. In such situations, human actors might seek the viewpoint of *more knowledgeable* colleagues (e.g., with more years of experience, or a different specialization). As such, the way of satisfying one's information needs shifts from *informational* to *transactional* (Jansen et al., 2008); evolving into a process of locating a potentially different source of information (Ekstrand et al., 2023). These behaviors are aligned with the notion of a *more knowledgeable other* by Vygotsky (2012), e.g., a parent or teacher who has more knowledge of a topic a child might seek to learn.

Future research could further investigate these shifts in information-seeking behaviors for explainability and mitigate critiques around selecting (Lai et al., 2023) and "comparing and contrasting" (Miller, 2023; Cai et al., 2019a) different explanations. We believe gaining such insights could also (partially) address different

biases encoded within an AI system. For instance, anchoring bias (Tversky and Kahneman, 1974) could manifest when AI-generated outputs and explanations are pushed to human actors before their own decision and sense-making processes take place (Bach et al., 2023). Finally, it is conceivable that depending on the nature of the AI system (e.g., proactive vs. reactive), different delivery strategies, and their timing, should be further investigated.

4.5.2 Temporal dynamics of explanations

Despite prior work in explainability soliciting longitudinal perspectives and studies (Ehsan et al., 2019; Maharjan et al., 2021), prior work in explainability has treated explanations as stationary entities and overlooked the *temporal dynamics* of explainable AI. Examples of longitudinal perspectives around studying explainable AI are few and far between (e.g., (Corti et al., 2024)), and often occur at, and focus on, a single moment in time (e.g., (Cai et al., 2019c)). We believe this to be an important knowledge gap given that (1) in "single sitting" studies participants might only get a crude impression of what an AI system can do; and, (2) the potential benefits and shortcomings of certain explanations might emerge after multiple interactions occur over some time. For instance, Corti et al. (2024) found that clinicians see explanations as a tool for learning AI-based CDSSs. That is, explanations should always be present but bear diminishing importance as they learn when and how to use CDSSs.

Longitudinal studies, or studies adopting a longitudinal perspective, per se, are not new to HCI scholars who have organized Special Interest Groups (Vaughan and Courage, 2007), panels (Vaughan et al., 2008), and tutorials (Jameson et al., 2009) around conducting those. For instance, researchers adopted a longitudinal perspective to inquire about the planning and delivery of health interventions (Jo et al., 2022; Blair et al., 2023). Other works cover the usage of conversational agents (Pradhan et al., 2020; Albert et al., 2023), users' behaviors on the web (Teevan et al., 2010; Furtado et al., 2013), personal competence development (Pekarek Doehler and Berger, 2019), or the design and creation of specific tools (Tabard et al., 2008). Such a body of longitudinal research presents, in our view, inspiration and a clear opportunity to understand when and how to apply explainable AI techniques to real-world problems.

4.5.3 Explanations availability to users

Despite the considerable efforts in explainable AI (Section 4.3.1), the fruition of XAI methods remains cumbersome, even for developers and practitioners, and hinders the proper evaluation of AI

models beyond accuracy-based measures. To this, we denote two main contributing factors: (1) the infancy of certain techniques (discussed in Section 4.4.2); and, (2) the lack of support tools that integrate explainability. Here we discuss more in depth the latter.

On a surface level, explainability methods are accessible to developers and practitioners through authors' Github repositories and public collections of software libraries and packages.[7] However, from the perspective of practitioners' workflows, adopting similar methods or libraries could pose some caveats. First, and particularly for development libraries, is that not all of them might support Jupyter notebooks (e.g., (Hoover et al., 2020)) which is instead a prevalent development environment. While this problem might be hinging on individual developers' workflows, we argue that the interactivity of explanations within an environment customary to many practitioners working with AI is a key aspect to bear in mind. Second, explainability techniques might come embedded in prepackaged tools which promote the use of their ad-hoc interfaces to compute and consume explanations (e.g., (Geva et al., 2022; Nori et al., 2019)). To exemplify this, we looked at existing explainability tools for developers working on Natural Language Processing and Language Modeling. We bring relevant examples in Table 4.1. Setting aside the specific purposes of individual methods and tools (e.g., InterroLang (Feldhus et al., 2023) emphasizes exploring AI models), learning new XAI methods and tools, or transitioning across them, can prove cumbersome.

We believe in pursuing research in the direction of interactive explanations (Xu et al., 2023; Mohseni et al., 2021) and, particularly, *selective* and *mutable* (Bertrand et al., 2023) explanations. First, selectivity would enable users to decide when to interact with explanations, change the visualization modality, and tweak the granularity of the information. This can be particularly useful for laypeople who might prioritize the completion of the task at hand. Second, mutability would expand on those interactions to encompass hypothesis testing and comparisons, something that could bring practical benefits to developers and practitioners during AI system design, development, and testing.

4.5.4 Human knowledge for model explainability

Grounding in the real world and human knowledge (historical, domain-specific, etc.) is a missing element in many explainers for

[7] A simple example is PyPI (https://pypi.org/), a searchable repository of community-developed Python software.

Table 4.1 Examples of XAI tools that support (large) language models. *Venue*: system-centered (S), human-centered (H), or neither (N.A.) for non-peer-reviewed tools. *XAI Methods*: we mark the kind of XAI methods that each tool supports.

Paper	Venue	XAI Methods						
		Gradient-based	Surrogate Model	Shapley Values	Hidden layer analysis	Rationalization	Instance Similarity	Perturbation-based
Wallace et al. (2019)	S	✓						✓
Alammar (2021)	S	✓			✓			
Hoover et al. (2020)	S				✓			
Attanasio et al. (2023)	S	✓	✓	✓				
Sarti et al. (2023)	S	✓	✓	✓	✓			
Feldhus et al. (2023)	S	✓				✓	✓	✓
Tenney et al. (2020)	S	✓	✓		✓			✓
Geva et al. (2022)	S				✓			
Feldhus et al. (2021)	S	✓	✓	✓				
Yuan et al. (2022)	H			✓				
Liang et al. (2023)	H	✓	✓	✓				
Song et al. (2023)	H				✓			
Park et al. (2021)	H							✓
Hong et al. (2021)	H							
Nori et al. (2019)	N.A.		✓	✓				
ArizeAI (2023)	N.A.				✓			
Nanda and Bloom (2022)	N.A.				✓			

AI systems. For instance, a common approach to computing explanations is feature importance which, in its simplest formulation, produces a single scalar value indicating the importance of a particular input feature. Then, such scores are visualized for individual input features as image overlays, or as textual token highlights. While similar approaches might give a crude impression of what an AI system might be doing, much is left on the shoulders of the recipients of explanations to determine on their own.

Crowdsourcing offers a viable path towards helping to characterize and explain what AI systems might be doing by eliciting and leveraging people's cognitive abilities at scale. From a broader perspective, prior crowdsourcing works have demonstrated the complementarity of crowd workers and computational means (e.g., clustering (Chang et al., 2016)) on a variety of tasks (Wu et al., 2023). Related to XAI instead, literature shows that involving humans in the explainability process can help to identify what an AI system should and might know (Balayn et al., 2021; Sharifi Noorian et al., 2022) using both explicit and implicit knowledge (Balayn et al., 2022a).

It is important to keep in mind, though, that crowdsourcing is not a panacea for the open challenges in AI and XAI research and development. Plainly asking crowd workers "what should the model know in this case?" is simply not feasible. For one, care should be taken when designing crowdsourcing tasks and incentives so that the desired outcome is achieved (e.g., high-quality annotations) while creating a fair work environment (e.g., proper hourly wage). Second, for the elicited knowledge to be practically useful, procedures need to be put in place such that its informativeness is not compromised, or compromised to an acceptable degree, based on the problem at hand. Lastly, triangulating explanations, input from crowd workers, and existing datasets and curated collections of human knowledge (e.g., ImageNet (Deng et al., 2009) and ConceptNet (Speer et al., 2017)) could provide stronger reassurances when characterizing the behavior of AI systems.

Example: integrating with classical NLP and XAI

Explainable AI methods applicable to NLP models do not present explicit connections with linguistic phenomena. Rather, the explanations are produced *solely* relying on the model and what a particular explainer is built to do. The literature on computational linguistics is rich with methods (and respective implementations) that have been overshadowed by the advent of word embeddings and neural models (Gururaja et al., 2023).[8] While such methods might produce details that require specific knowledge (e.g., grammatical dependencies), we believe those to be instrumental and more insightful than the simple token highlights towards explaining a language model – early results also indicate this to be the case (Tocchetti et al., 2023). To that aim, prior works in HCI can inform the representations to users and the granularity of information on a case-by-case basis.

[8]The work done within Stanford's NLP group is a prime example of the efforts around computational linguistics. Website: https://nlp.stanford.edu/.

4.6 Potential use cases

Here, we adopt a speculative angle and exemplify some potential use cases where explainability could be desired. While the existing literature presents evidence for the need and application of explainable AI systems for the following use cases, those are severely limited. Particularly, we highlight both the computational components and human factors that either benefit or hinder the use of explanations.

4.6.1 AI model diagnosis

It is now well known that AI models and systems exhibit unreliable behavior when used in real-world scenarios which can manifest in different ways: poor robustness to unseen data (Tocchetti et al., 2025), generative models hallucinating (Ji et al., 2023), and – ultimately – producing real-world harms (Mehrabi et al., 2021). Prior works, indeed, have attempted to mitigate such unreliable – or, depending on the domain, even unacceptable – behaviors by developing methods to enhance or assess the robustness of AI systems and models (Tocchetti et al., 2025), debug individual code implementations (Selsam et al., 2017) and AI pipelines (including data-related issues) (Lourenço et al., 2019), or surface misalignments between design requirements and code implementations (Balayn et al., 2023). To aid such activities, explanations have already been leveraged for the purpose of debugging AI models – a process that Lertvittayakumjorn and Toni (2021) define as "explanation-based human debugging" (EBHD).[9]

Benefits

Nowadays, with the increased use of deep learning techniques, an immediate benefit of computing model explanations is to mitigate their opacity. While current explainability techniques might not be perfect, AI developers and practitioners involved in implementing and debugging AI models and pipelines have the necessary knowledge to interpret AI explanations satisfactorily. What we mean by this is that, despite literature indicating practitioners' hurdles in understanding AI explanations (e.g., (Balayn et al., 2023)), they get a crude impression of what the AI model they might be dealing with might be doing and can chart the next course of

[9] According to Lertvittayakumjorn and Toni (2021), EBHD encompasses both "explanatory debugging" (Kulesza et al., 2010) and "human-in-the-loop debugging" (Lertvittayakumjorn et al., 2020).

action, be it further training or collecting additional data following certain criteria. Finally, connected to data, some explainability methods bring to the surface influential (Koh and Liang, 2017; Guo et al., 2021) or prototypical (Chen et al., 2019; Nauta et al., 2021) data instances from training data. These, in turn, can be used to spot problematic subsets of data points which might be underrepresented in the dataset, need a different preprocessing pipeline, or display inconsistent annotations.

Hindrances

While AI developers and practitioners might be equipped with the right knowledge to interpret AI explanations, there are fundamental differences between debugging errors and diagnosing AI models. First, AI model diagnosis is a different and more complex task compared to debugging. Diagnosing AI models relies on the existence of specifications for model behavior, something that is context-specific and generally not readily available. Furthermore, while explanations can surface details concerning what information was seemingly used by a model to predict a label or generate a textual token, those only represent shallow descriptions of what a model might be doing and, by themselves, might be insufficient to adequately characterize and diagnose the behavior of AI models and systems.

4.6.2 Appropriate automation in healthcare

The application of AI in healthcare is received with both enthusiasm and concerns (Topol, 2019; Thieme et al., 2023). Specifically, AI methods and techniques are often tied to the development of Clinical Decision Support Systems (CDSSs). AI-powered or not, CDSSs can "*provide clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered, or presented at appropriate times, to enhance health and health care*" (Osheroff et al., 2007). However, the inherent criticality of clinical decision-making, the lack of strong guarantees of subsymbolic AI approaches (Schaaf et al., 2020; Walsh et al., 2019; Mazo et al., 2020; Velickovski et al., 2014; Durieux et al., 2000), and the broader concerns of accountability (Smith, 2021) aggravate the need for explanations for and from AI systems. For instance, in the context of pulmonary medicine, Corti et al. (2024) illuminated the explainability needs of clinicians across patient care. They found that clinicians seek general explanations of the technical affordances of AI systems as much as local, patient-specific explanations for individual AI-generated recommendations.

Benefits

Before and during the usage of AI-based CDSSs, explanations for those systems bring several benefits that align with clinicians' information and explainability needs. In the early adoption phases, explanations for AI-based CDSSs could be presented as documentation tools (Rostamzadeh et al., 20022; Cai et al., 2019c) that present an overview of a system, including its expected usage, its accuracy, and the patient data used to create it. Such information can then be used to determine whether a particular system is even applicable in a certain clinic or hospital. Similarly, explanations for individual AI-generated output can help clinicians assess such output in relation to the specific conditions of individual patients (Tonekaboni et al., 2019; Corti et al., 2024).

Finally, another benefit is learning how to appropriately use an AI-based CDSS. While, in general, AI techniques can process large amounts of data and compress it into small artifacts (i.e., the models), clinicians' in-situ experience – if captured at all – is rarely reflected appropriately by AI systems. For instance, for certain diseases (e.g., Idiopathic Pulmonary Fibrosis (Ferri and Facp, 2023)), clinicians combine lab test results, physical examinations, and patients' self-expressed desires and values to derive recommendations (e.g., by adjusting treatments) tailored to an individual patient.[10] To that aim, AI explanations can provide additional patient-specific indicators and grounding for AI-generated outputs.

Hindrances

Despite the promising practical benefits and the availability of explanations, the current generation of AI-based CDSSs (e.g., Merative[11]) displays unreliable behavior when applied in real-world practice (Topol, 2019). While guidelines for implementing AI and XAI in healthcare have been discussed (Markus et al., 2021; Lötsch et al., 2022), existing surveys show that in certain medical specialities, AI and explainable AI are underexplored (Pocevičiūtė et al., 2020; Singh et al., 2020). To some extent, this is understandable: many of the off-the-shelf explainability techniques might not adequately capture context and situation-specific details.

In addition to this, clinicians' possible aversion towards algorithmic solutions can impact their perception and use of AI-based CDSSs and their explanations (Dietvorst et al., 2015). Researchers and practitioners working with and implementing explainable AI in practice should be aware of possible barriers to adopting such

[10] These and other similar principles can be found in the literature around *patient-centric care* (Committee on Quality of Health Care in America, 2002; Breen et al., 2009).
[11] Merative (previously IBM Watson Health): https://www.merative.com/.

solutions (e.g., noticeable criticism over AI-generated suggestions vs. human-generated ones). Connected to this, the design of explanations may surface – even unintentionally – interactions typical of "dark patterns": user interfaces that do not benefit the intended recipients (with explanations in this case) but unintended others (Gray et al., 2018). It should be noted that the term "dark pattern" normally has a negative connotation, related to deception and manipulation of users of a particular product being sold. Here, we adopt it to discuss interactions with user interfaces that display AI explanations (Chromik et al., 2019) which could be obstructive or manipulative. The designers of explainable AI-based CDDSSs should be mindful of clinicians feeling forced or nagged to engage with explanations, making the interaction with explanations obstructive (i.e., a convoluted navigation of information) or sneaky (e.g., by withholding information). While it might be tempting to include explanations as a way to show how "good" a system might be, clinicians might want and need explanations to cope with its imperfections in their daily practice (Ehsan et al., 2024a; Qian Yang et al., 2019; Cai et al., 2019b).

4.6.3 AI for accessible resources for second-language learners

Second-language learners (e.g., migrants) struggle to participate actively in everyday society. Embedded clauses, passive voice, noncanonical word order, and specific vocabulary can disrupt the flow of arguments and make common activities (e.g., reading newspapers) complex. For instance, in the Netherlands, about 2.5 million citizens aged between 16 and 65 years old are considered to be of "low-literacy".[12] To mitigate this phenomenon, public authorities have been promoting initiatives to manually simplify texts by leveraging diverse resources: corpus-specific frequency (Bauman and Culligan, 1995) lists, controlled languages, and guidelines for accessible communication (UK Government, 2022).[13]

Text simplification (TS) is a traditional Natural Language Processing (NLP) task. To that end, numerous methods have been proposed to address text simplification at a lexical and syntactical level (Glavaš and Štajner, 2015; Horn et al., 2014; Specia, 2010; Wang et al., 2016; Zhang and Lapata, 2017). By capturing the distributional semantics of words (Mikolov et al., 2013; Devlin et al., 2019), generative large language models have recently shown a human-like ability to simplify text by avoiding complex sentences and preserving text coherence. However, such models are still at an early stage

[12] https://www.lezenenschrijven.nl/informatie-over-laaggeletterdheid-nederland.
[13] An example of a controlled language is ASD-STE100: https://asd-ste100.org/about.html.

and by encoding spurious correlations they might alter the meaning of the text and hallucinate (Ji et al., 2023). For instance, generative LLMs might insert exogenous concepts never mentioned in the original text. While such issues are present across all generative LLMs, they are exacerbated when working with low-resource languages like Dutch (Corti and Yang, 2023).

Benefits

In a similar setting, where the output of AI systems is shown to citizens in their everyday lives, explanations play a crucial role in informing them (1) that an AI system is being used in the first place; and, (2) of instance-level (i.e., on specific fragments of text) that help them appropriately rely on the AI-simplified text. Simple explanation techniques such as token-level feature attribution maps (e.g., (Sundararajan et al., 2017)) could highlight several discrepancies between the original and the simplified text: crucial details that have been obfuscated (Corti and Yang, 2023), temporal inconsistencies that have been introduced, and factually incorrect information (Ji et al., 2023).

Hindrances

Despite the benefits stemming from the inclusion of explanations, these might still prove difficult to parse and understand for second-language learners. If uncontrolled (e.g., by human raters and experts), second-language learners might find AI-generated simplifications still quite complex. Furthermore, AI literacy (Section 4.4.6) heavily shapes the way citizens might understand an explanation. For instance, does the fact that "hitherto" from the complex text contributed 75% towards the generation of "until now" in the simplified text mean that in 75% of all simplifications that is an operation that is made? Or, that the model is 75% sure that this simplification is the correct one?

4.7 An invitation for interdisciplinary collaborations

We hope that the challenges, opportunities, and use cases discussed show that, while explainable AI is inherently an interdisciplinary effort, the system- and human-centered camps do not actively interact with each other despite sharing tools and prioritizing complementary (albeit different) aspects of explainable AI.

In light of this, we would like to extend an invitation for interdisciplinary collaborations, understanding each others' way of researching explainable AI and creating means and artifacts to

aid communication and knowledge sharing across camps. For these collaborations to take place, we suggest the use of Research through Design – just like we did in this paper – as a scaffolding for the overall process. RtD can provide the tools to foster system- and human-centered works in such a way that does not harm either way of conducting research while maintaining the focus on the bigger picture. From a system-centered perspective, RtD is systematic and emphasizes the rigorous documentation of the research process (through academic papers, technical documentation, code repositories, etc.). Furthermore, such documentation can be further tailored to individual stakeholders' needs if necessary. From a human-centered perspective, RtD gives room for researchers' reflexivity to manifest in practice and cultivate the pluralistic views and discussion around explainable AI typical of the HCI literature.

Practically, there is a need to explore more closely the intersection between system- and human-centered works on explainable AI: the similarities help to establish common grounds, and the differences help to formulate potential and fruitful research questions and agendas. Additional aspects to consider are the directions the two camps are moving towards. On the one hand, system-centered researchers have recently started to experience pressure from industry, making their work potentially less useful (e.g., AI systems from industry consistently performing better than research-led efforts) or less accessible (e.g., complex code implementations with ad-hoc setups required). This, in turn, could make such work less appealing to the broader research community. On the other hand, the HCI community has been growing more and more interdisciplinary (Chen, 2024). As a result, we might see an increasing display of system-centered contributions in HCI venues. However, we view this phenomenon as risky. Overfitting to the "customs" of a community might not emphasize the goodness of (the parts of) such contributions and might exacerbate the creation of an "echo chamber" effect, ultimately endangering collaborations beyond the HCI community.

One thing that we would like to make explicit is that we are not arguing for the adoption of RtD in every paper. While possible, the systematicity of RtD, the documentation and artifacts generated throughout the process, and the knowledge being (re)framed would likely not fit within the limits set by individual venues and journals.

4.8 Conclusion

In this chapter, we adopted Research through Design as a lens to question our understanding of explainable AI and how it should be

researched. We discussed contributions to explainable AI research from both the system- and human-centered angles, highlighting, however, the solitary modus operandi of the two camps. To address this, we also suggested Research through Design as an overarching method to research explainable AI. While being rooted in design practice, Research through Design is more systematic, reflexive, and documented such that other researchers and practitioners can follow similar steps without the expectation of achieving the same outcome. We believe this flexibility offered by Research through Design to be instrumental in addressing the socio-technical nature of AI systems, the bidirectional effects emerging in and shaping human–AI interactions, the inherent diversity of the stakeholders of AI systems, and the nuances of specific contexts in which explanations are wanted or necessary. We substantiate our proposal by discussing (1) the challenges that the system- and human-centered camps within explainable AI research face; (2) the opportunities that lie in studying the bidirectionality of human–AI interactions, the temporal dynamics that shape those interactions, the availability of explanations, and the role of human knowledge in creating and understanding explanations; and, (3) a speculative exercise grounded in real-world use cases of explainable AI systems. We call on researchers in either of the camps of explainable AI to collaborate more closely, understanding each others' ways to conduct scholarly research, and creating means and artifacts that aid communication across camps while being situated at the intersection of system- and human-centered perspectives around explainable AI.

Acknowledgments

This research has been partially supported by the TU Delft *Design@Scale* AI Lab.

References

Abdul, Ashraf, Vermeulen, Jo, Wang, Danding, Lim, Brian Y., Kankanhalli, Mohan, 2018. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Montreal, QC, Canada. Association for Computing Machinery, New York, NY, USA, pp. 1–18. https://doi.org/10.1145/3173574.3174156.

Alammar, J., 2021. Ecco: an open source library for the explainability of transformer language models. In: Ji, Heng, Park, Jong C., Xia, Rui (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Association for Computational Linguistics, pp. 249–257. https://doi.org/10.18653/v1/2021.acl-demo.30.

Albert, Saul, Hamann, Magnus, Stokoe, Elizabeth, 2023. Conversational user interfaces in smart homecare interactions: a conversation analytic case study. In: Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23). Eindhoven, Netherlands. Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. https://doi.org/10.1145/3571884.3597140.

Alcorn, Michael A., Li, Qi, Gong, Zhitao, Wang, Chengfei, Mai, Long, Ku, Wei-Shinn, Nguyen, Anh, 2019. Strike (with) a pose: neural networks are easily fooled by strange poses of familiar objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Alqaraawi, Ahmed, Schuessler, Martin, Weiß, Philipp, Costanza, Enrico, Berthouze, Nadia, 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In: Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20). Cagliari, Italy. Association for Computing Machinery, New York, NY, USA, pp. 275–285. https://doi.org/10.1145/3377325.3377519.

Alvarez Melis, David, Jaakkola, Tommi, 2018. Towards robust interpretability with self-explaining neural networks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

Anik, Ariful Islam, Bunt, Andrea, 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Yokohama, Japan. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. https://doi.org/10.1145/3411764.3445736.

ArizeAI, 2023. Phoenix: AI observability & evaluation. https://github.com/Arize-ai/phoenix.

Attanasio, Giuseppe, Pastor, Eliana, Di Bonaventura, Chiara, Nozza, Debora, 2023. ferret: a framework for benchmarking explainers on transformers. In: Croce, Danilo, Soldaini, Luca (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Dubrovnik, Croatia, pp. 256–266. https://doi.org/10.18653/v1/2023.eacl-demo.29.

Bach, Anne Kathrine Petersen, Nørgaard, Trine Munch, Brok, Jens Christian, van Berkel, Niels, 2023. "If I had all the time in the world": ophthalmologists' perceptions of anchoring bias mitigation in clinical AI support. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 16, 14 pages. https://doi.org/10.1145/3544548.3581513.

Balayn, Agathe, Corti, Lorenzo, Rancourt, Fanny, Casati, Fabio, Gadiraju, Ujwal, 2024. Understanding stakeholders' perceptions and needs across the LLM supply chain. arXiv:2405.16311 [cs.HC].

Balayn, Agathe, He, Gaole, Hu, Andrea, Yang, Jie, Gadiraju, Ujwal, 2022a. Ready player one! Eliciting diverse knowledge using a configurable game. In: Proceedings of the ACM Web Conference 2022 (WWW '22). Virtual Event, Lyon, France. Association for Computing Machinery, New York, NY, USA, pp. 1709–1719. https://doi.org/10.1145/3485447.3512241.

Balayn, Agathe, Rikalo, Natasa, Lofi, Christoph, Yang, Jie, Bozzon, Alessandro, 2022b. How can explainability methods be used to support bug identification in computer vision models? In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). New Orleans, LA, USA.

Association for Computing Machinery, New York, NY, USA, 184. https://doi.org/10.1145/3491102.3517474.

Balayn, Agathe, Rikalo, Natasa, Yang, Jie, Bozzon, Alessandro, 2023. Faulty or ready? Handling failures in deep-learning computer vision models until deployment: a study of practices, challenges, and needs. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 11, 20 pages. https://doi.org/10.1145/3544548.3581555.

Balayn, Agathe, Soilis, Panagiotis, Lofi, Christoph, Yang, Jie, Bozzon, Alessandro, 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In: Proceedings of the Web Conference 2021 (WWW '21). Ljubljana, Slovenia. Association for Computing Machinery, New York, NY, USA, pp. 1937–1948. https://doi.org/10.1145/3442381.3450069.

Barredo Arrieta, Alejandro, Díaz-Rodríguez, Natalia, Del Ser, Javier, Bennetot, Adrien, Tabik, Siham, Barbado, Alberto, Garcia, Salvador, Gil-Lopez, Sergio, Molina, Daniel, Benjamins, Richard, Chatila, Raja, Herrera, Francisco, 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bauman, John, Culligan, Brent, 1995. About the general service list. Retrieved, July 21 (1995), pp. 2012.

Bender, Emily M., Koller, Alexander, 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.

Bertrand, Astrid, Viard, Tiphaine, Belloum, Rafik, Eagan, James R., Maxwell, Winston, 2023. On selective, mutable and dialogic XAI: a review of what users say about different types of interactive explanations. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 411, 21 pages. https://doi.org/10.1145/3544548.3581314.

Betker, James, Goh, Gabriel, Jing, Li, Brooks, Tim, Wang, Jianfeng, Li, Linjie, Ouyang, Long, Zhuang, Juntang, Lee, Joyce, Guo, Yufei, et al., 2023. Improving image generation with better captions. Computer Science 2 (3), 8. https://cdn.openai.com/papers/dall-e-3.pdf.

Bijker, Wiebe E., Hughes, Thomas P., Pinch, Trevor, 1987. The Social Constructions of Technological Systems. MIT Press.

Biswas, Shreyan, Corti, Lorenzo, Buijsman, Stefan, Yang, Jie, 2022. CHIME: causal human-in-the-loop model explanations. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 10, pp. 27–39. https://doi.org/10.1609/hcomp.v10i1.21985.

Blair, Johnna, Mukherjee, Dahlia, Saunders, Erika F.H., Abdullah, Saeed, 2023. Knowing how long a storm might last makes it easier to weather: exploring needs and attitudes toward a data-driven and preemptive intervention system for bipolar disorder. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 496, 12 pages. https://doi.org/10.1145/3544548.3581563.

Bommasani, Rishi, Hudson, Drew A., Adeli, Ehsan, Altman, Russ, Arora, Simran, von Arx, Sydney, Bernstein, Michael S., Bohg, Jeannette, Bosselut, Antoine, Brunskill, Emma, Brynjolfsson, Erik, Buch, Shyamal, Card, Dallas, Castellon, Rodrigo, Chatterji, Niladri, Chen, Annie, Creel, Kathleen, Davis, Jared Quincy, Demszky, Dora, Donahue, Chris, Doumbouya, Moussa, Durmus, Esin, Ermon,

Stefano, Etchemendy, John, Ethayarajh, Kawin, Fei-Fei, Li, Finn, Chelsea, Gale, Trevor, Gillespie, Lauren, Goel, Karan, Goodman, Noah, Grossman, Shelby, Guha, Neel, Hashimoto, Tatsunori, Henderson, Peter, Hewitt, John, Ho, Daniel E., Hong, Jenny, Hsu, Kyle, Huang, Jing, Icard, Thomas, Jain, Saahil, Jurafsky, Dan, Kalluri, Pratyusha, Karamcheti, Siddharth, Keeling, Geoff, Khani, Fereshte, Khattab, Omar, Koh, Pang Wei, Krass, Mark, Krishna, Ranjay, Kuditipudi, Rohith, Kumar, Ananya, Ladhak, Faisal, Lee, Mina, Lee, Tony, Leskovec, Jure, Levent, Isabelle, Li, Xiang Lisa, Li, Xuechen, Ma, Tengyu, Malik, Ali, Manning, Christopher D., Mirchandani, Suvir, Mitchell, Eric, Munyikwa, Zanele, Nair, Suraj, Narayan, Avanika, Narayanan, Deepak, Newman, Ben, Nie, Allen, Niebles, Juan Carlos, Nilforoshan, Hamed, Nyarko, Julian, Ogut, Giray, Orr, Laurel, Papadimitriou, Isabel, Park, Joon Sung, Piech, Chris, Portelance, Eva, Potts, Christopher, Raghunathan, Aditi, Reich, Rob, Ren, Hongyu, Rong, Frieda, Roohani, Yusuf, Ruiz, Camilo, Ryan, Jack, Ré, Christopher, Sadigh, Dorsa, Sagawa, Shiori, Santhanam, Keshav, Shih, Andy, Srinivasan, Krishnan, Tamkin, Alex, Taori, Rohan, Thomas, Armin W., Tramèr, Florian, Wang, Rose E., Wang, William, Wu, Bohan, Wu, Jiajun, Wu, Yuhuai, Xie, Sang Michael, Yasunaga, Michihiro, You, Jiaxuan, Zaharia, Matei, Zhang, Michael, Zhang, Tianyi, Zhang, Xikun, Zhang, Yuhui, Zheng, Lucia, Zhou, Kaitlyn, Liang, Percy, 2021. On the opportunities and risks of foundation models. https://doi.org/10.48550/ARXIV.2108.07258.

Bommasani, Rishi, Klyman, Kevin, Longpre, Shayne, Kapoor, Sayash, Maslej, Nestor, Xiong, Betty, Zhang, Daniel, Liang, Percy, 2023. The foundation model transparency index. arXiv:2310.12941 [cs.LG].

Bordt, Sebastian, Finck, Michèle, Raidl, Eric, von Luxburg, Ulrike, 2022. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Seoul, Republic of Korea. Association for Computing Machinery, New York, NY, USA, pp. 891–905. https://doi.org/10.1145/3531146.3533153.

Breen, Gerald-Mark, Wan, Thomas T.H., Zhang, Ning Jackie, Marathe, Shriram S., Seblega, Binyam K., Paek, Seung Chun, 2009. Improving doctor–patient communication: examining innovative modalities vis-à-vis effective patient-centric care management technology. Journal of Medical Systems 33 (2), 155–162. https://doi.org/10.1007/s10916-008-9175-3.

Buchanan, Bruce G., Shortliffe, Edward H., 1984. Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project. The Addison-Wesley Series in Artificial Intelligence. Addison-Wesley Longman Publishing Co., Inc.

Buijsman, Stefan, 2022. Defining explanation and explanatory depth in XAI. Minds and Machines 32 (3), 563–584. https://doi.org/10.1007/s11023-022-09607-9.

Butlin, Patrick, Long, Robert, Elmoznino, Eric, Bengio, Yoshua, Birch, Jonathan, Constant, Axel, Deane, George, Fleming, Stephen M., Frith, Chris, Ji, Xu, Kanai, Ryota, Klein, Colin, Lindsay, Grace, Michel, Matthias, Mudrik, Liad, Peters, Megan A.K., Schwitzgebel, Eric, Simon, Jonathan, VanRullen, Rufin, 2023. Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708 [cs.AI].

Cai, Carrie J., Jongejan, Jonas, Holbrook, Jess, 2019a. The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). Marina del Ray, California. Association for Computing Machinery, New York, NY, USA, pp. 258–262. https://doi.org/10.1145/3301275.3302289.

Cai, Carrie J., Reif, Emily, Hegde, Narayan, Hipp, Jason, Kim, Been, Smilkov, Daniel, Wattenberg, Martin, Viegas, Fernanda, Corrado, Greg S., Stumpe, Martin C.,

Terry, Michael, 2019b. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Glasgow, Scotland, UK. Association for Computing Machinery, New York, NY, USA, pp. 1–14. https://doi.org/10.1145/3290605.3300234.

Cai, Carrie J., Winter, Samantha, Steiner, David, Wilcox, Lauren, Terry, Michael, 2019c. "Hello AI": uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. Proceedings of the ACM on Human–Computer Interaction 3 (CSCW), Article 104, 24 pages. https://doi.org/10.1145/3359206.

Carvalho, Diogo V., Pereira, Eduardo M., Cardoso, Jaime S., 2019. Machine learning interpretability: a survey on methods and metrics. Electronics 8 (8), 832.

Chang, Joseph Chee, Kittur, Aniket, Hahn, Nathan, 2016. Alloy: clustering with crowds and computation. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). San Jose, California, USA. Association for Computing Machinery, New York, NY, USA, pp. 3180–3191. https://doi.org/10.1145/2858036.2858411.

Chen, Chaofan, Li, Oscar, Tao, Daniel, Barnett, Alina, Rudin, Cynthia, Su, Jonathan K., 2019. This looks like that: deep learning for interpretable image recognition. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

Chen, Jianbo, Song, Le, Wainwright, Martin, Jordan, Michael, 2018. Learning to explain: an information-theoretic perspective on model interpretation. In: Dy, Jennifer, Krause, Andreas (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80. PMLR, pp. 883–892. https://proceedings.mlr.press/v80/chen18j.html.

Chen, Xiang 'Anthony', 2024. HCI papers cite HCI papers, increasingly so. In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA, Article 556, 6 pages. https://doi.org/10.1145/3613905.3644070.

Chromik, Michael, Eiband, Malin, Völkel, Sarah Theres, Buschek, Daniel, 2019. Dark patterns of explainability, transparency, and user control for intelligent systems. In: Trattner, Christoph, Parra, Denis, Riche, Nathalie (Eds.), Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), CEUR Workshop Proceedings, vol. 2327. Los Angeles, USA, March 20, 2019. CEUR-WS.org. https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-7.pdf.

Committee on Quality of Health Care in America, 2002. Crossing the quality chasm: a new health system for the 21st century. Journal for Healthcare Quality 24 (5), 52.

Cooley, Mike, 1989. Human-Centred Systems. Springer, London, pp. 133–143. https://doi.org/10.1007/978-1-4471-1717-9_10.

Corti, Lorenzo, Oltmans, Rembrandt, Jung, Jiwon, Balayn, Agathe, Wijsenbeek, Marlies, Yang, Jie, 2024. "It is a moving process": understanding the evolution of explainability needs of clinicians in pulmonary medicine. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642551.

Corti, Lorenzo, Yang, Jie, 2023. ARTIST: ARTificial intelligence for simplified text. arXiv:2308.13458.

Croskerry, Pat, 2013. From mindless to mindful practice—cognitive bias and clinical decision making. The New England Journal of Medicine 368 (26), 2445–2448.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li, 2009. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

di Giovanni, Marco, Brambilla, Marco, 2021. EFSG: evolutionary fooling sentences generator. In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), pp. 171–178. https://doi.org/10.1109/ICSC50631.2021.00037.

Dietvorst, Berkeley J., Simmons, Joseph P., Massey, Cade, 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology. General 144 (1), 114.

Dorst, Kees, Cross, Nigel, 2001. Creativity in the design process: co-evolution of problem–solution. Design Studies 22 (5), 425–437. https://doi.org/10.1016/S0142-694X(01)00009-6.

Durieux, Pierre, Nizard, Rémy, Ravaud, Philippe, Mounier, Nicolas, Lepage, Eric, 2000. A clinical decision support system for prevention of venous thromboembolism: effect on physician behavior. JAMA 283 (21), 2816–2821. https://doi.org/10.1001/jama.283.21.2816. https://jamanetwork.com/journals/jama/articlepdf/192759/joc00041.pdf.

Ehsan, Upol, Liao, Q. Vera, Muller, Michael, Riedl, Mark O., Weisz, Justin D., 2021a. Expanding explainability: towards social transparency in AI systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Yokohama, Japan. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. https://doi.org/10.1145/3411764.3445188.

Ehsan, Upol, Liao, Q. Vera, Passi, Samir, Riedl, Mark O., Daumé, Hal, 2024a. Seamful XAI: operationalizing seamful design in explainable AI. Proceedings of the ACM on Human–Computer Interaction 8 (CSCW1), Article 119, 29 pages. https://doi.org/10.1145/3637396.

Ehsan, Upol, Riedl, Mark O., 2020. Human-centered explainable AI: towards a reflective sociotechnical approach. In: Stephanidis, Constantine, Kurosu, Masaaki, Degen, Helmut, Reinerman-Jones, Lauren (Eds.), HCI International 2020 – Late Breaking Papers: Multimodality and Intelligence. Springer International Publishing, Cham, pp. 449–466.

Ehsan, Upol, Riedl, Mark O., 2024. Social construction of XAI: do we need one definition to rule them all? Patterns 5 (2).

Ehsan, Upol, Saha, Koustuv, De Choudhury, Munmun, Riedl, Mark O., 2023a. 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), Article 34, 32 pages. https://doi.org/10.1145/3579467.

Ehsan, Upol, Tambwekar, Pradyumna, Chan, Larry, Harrison, Brent, Riedl, Mark O., 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In: Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). Marina del Ray, California.

Association for Computing Machinery, New York, NY, USA, pp. 263–274. https://doi.org/10.1145/3301275.3302316.

Ehsan, Upol, Watkins, Elizabeth A., Wintersberger, Philipp, Manger, Carina, Kim, Sunnie S.Y., Van Berkel, Niels, Riener, Andreas, Riedl, Mark O., 2024b. Human-centered explainable AI (HCXAI): reloading explainability in the era of large language models (LLMs). In: Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA, Article 477, 6 pages. https://doi.org/10.1145/3613905.3636311.

Ehsan, Upol, Wintersberger, Philipp, Liao, Q. Vera, Mara, Martina, Streit, Marc, Wachter, Sandra, Riener, Andreas, Riedl, Mark O., 2021b. Operationalizing human-centered perspectives in explainable AI. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21). Yokohama, Japan. Association for Computing Machinery, New York, NY, USA, Article 94, 6 pages. https://doi.org/10.1145/3411763.3441342.

Ehsan, Upol, Wintersberger, Philipp, Liao, Q. Vera, Watkins, Elizabeth Anne, Manger, Carina, Daumé III, Hal, Riener, Andreas, Riedl, Mark O., 2022. Human-centered explainable AI (HCXAI): beyond opening the black-box of AI. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). New Orleans, LA, USA. Association for Computing Machinery, New York, NY, USA, Article 109, 7 pages. https://doi.org/10.1145/3491101.3503727.

Ehsan, Upol, Wintersberger, Philipp, Watkins, Elizabeth A., Manger, Carina, Ramos, Gonzalo, Weisz, Justin D., Daumé III, Hal, Riener, Andreas, Riedl, Mark O., 2023b. Human-centered explainable AI (HCXAI): coming of age. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 353, 7 pages. https://doi.org/10.1145/3544549.3573832.

Ekstrand, Michael D., Pera, Maria Soledad, Wright, Katherine Landau, 2023. Seeking information with a more knowledgeable other. Interactions 30 (1), 70–73. https://doi.org/10.1145/3573364.

Feldhus, Nils, Schwarzenberg, Robert, Möller, Sebastian, 2021. Thermostat: a large collection of NLP model explanations and analysis tools. In: Adel, Heike, Shi, Shuming (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 87–95. https://doi.org/10.18653/v1/2021.emnlp-demo.11.

Feldhus, Nils, Wang, Qianli, Anikina, Tatiana, Chopra, Sahil, Oguz, Cennet, Möller, Sebastian, 2023. InterroLang: exploring NLP models and datasets through dialogue-based explanations. In: Bouamor, Houda, Pino, Juan, Bali, Kalika (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp. 5399–5421. https://doi.org/10.18653/v1/2023.findings-emnlp.359.

Ferri, Fred F., Facp, M.D., 2023. Ferri's Clinical Advisor 2024, E-Book. Elsevier Health Sciences, pp. 775–777. Chapter Idiopathic Pulmonary Fibrosis.

Fogliato, Riccardo, Chappidi, Shreya, Lungren, Matthew, Fisher, Paul, Wilson, Diane, Fitzke, Michael, Parkinson, Mark, Horvitz, Eric, Inkpen, Kori, Nushi, Besmira, 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Seoul, Republic of Korea. Association for Computing Machinery, New York, NY, USA, pp. 1362–1374. https://doi.org/10.1145/3531146.3533193.

Freiesleben, Timo, König, Gunnar, 2023. Dear XAI community, we need to talk! In: Longo, Luca (Ed.), Explainable Artificial Intelligence. Springer Nature Switzerland, Cham, pp. 48–65.

Furtado, Adabriand, Andrade, Nazareno, Oliveira, Nigini, Brasileiro, Francisco, 2013. Contributor profiles, their dynamics, and their importance in five Q&A sites. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13). San Antonio, Texas, USA. Association for Computing Machinery, New York, NY, USA, pp. 1237–1252. https://doi.org/10.1145/2441776.2441916.

Gaver, Bill, Martin, Heather, 2000. Alternatives: exploring information appliances through conceptual design proposals. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00). The Hague, The Netherlands. Association for Computing Machinery, New York, NY, USA, pp. 209–216. https://doi.org/10.1145/332040.332433.

Gebru, Timnit, Morgenstern, Jamie, Vecchione, Briana, Wortman Vaughan, Jennifer, Wallach, Hanna, Daumé III, Hal, Crawford, Kate, 2021. Datasheets for datasets. Communications of the ACM 64 (12), 86–92. https://doi.org/10.1145/3458723.

Geva, Mor, Caciularu, Avi, Dar, Guy, Roit, Paul, Sadde, Shoval, Shlain, Micah, Tamir, Bar, Goldberg, Yoav, 2022. LM-debugger: an interactive tool for inspection and intervention in transformer-based language models. In: Che, Wanxiang, Shutova, Ekaterina (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, Abu Dhabi, UAE, pp. 12–21. https://doi.org/10.18653/v1/2022.emnlp-demos.2.

Ghorbani, Amirata, Wexler, James, Zou, James Y., Kim, Been, 2019. Towards automatic concept-based explanations. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

Glavaš, Goran, Štajner, Sanja, 2015. Simplifying lexical simplification: do we need simplified corpora? In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 63–68.

Goyal, Yash, Wu, Ziyan, Ernst, Jan, Batra, Dhruv, Parikh, Devi, Lee, Stefan, 2019. Counterfactual visual explanations. In: Chaudhuri, Kamalika, Salakhutdinov, Ruslan (Eds.), Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97. PMLR, pp. 2376–2384. https://proceedings.mlr.press/v97/goyal19a.html.

Gray, Colin M., Kou, Yubo, Battles, Bryan, Hoggatt, Joseph, Toombs, Austin L., 2018. The dark (patterns) side of UX design. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Montreal QC, Canada. Association for Computing Machinery, New York, NY, USA, pp. 1–14. https://doi.org/10.1145/3173574.3174108.

Green, Ben, Viljoen, Salomé, 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Barcelona, Spain. Association for Computing Machinery, New York, NY, USA, pp. 19–31. https://doi.org/10.1145/3351095.3372840.

Guidotti, Riccardo, 2022. Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery 2022. https://doi.org/10.1007/s10618-022-00831-6.

Guidotti, Riccardo, Monreale, Anna, Ruggieri, Salvatore, Pedreschi, Dino, Turini, Franco, Giannotti, Fosca, 2018a. Local rule-based explanations of black box decision systems. arXiv:1805.10820 [cs.AI].

Guidotti, Riccardo, Monreale, Anna, Ruggieri, Salvatore, Turini, Franco, Giannotti, Fosca, Pedreschi, Dino, 2018b. A survey of methods for explaining black box models. ACM Computing Surveys 51 (5), Article 93, 42 pages. https://doi.org/10.1145/3236009.

Guo, Han, Rajani, Nazneen, Hase, Peter, Bansal, Mohit, Xiong, Caiming, 2021. FastIF: scalable influence functions for efficient model interpretation and debugging. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 10333–10350. https://doi.org/10.18653/v1/2021.emnlp-main.808.

Gururaja, Sireesh, Bertsch, Amanda, Na, Clara, Widder, David, Strubell, Emma, 2023. To build our future, we must know our past: contextualizing paradigm shifts in natural language processing. In: Bouamor, Houda, Pino, Juan, Bali, Kalika (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp. 13310–13325. https://doi.org/10.18653/v1/2023.emnlp-main.822.

He, Gaole, Kuiper, Lucie, Gadiraju, Ujwal, 2023. Knowing about knowing: an illusion of human competence can hinder appropriate reliance on AI systems. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. https://doi.org/10.1145/3544548.3581025.

Hind, Michael, 2019. Explaining explainable AI. XRDS 25 (3), 16–19. https://doi.org/10.1145/3313096.

Holder, Eric, Huang, Lixiao, Chiou, Erin, Jeon, Myounghoon, Lyons, Joseph B., 2021. Designing for bi-directional transparency in human–AI–robot-teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 65 (1), 57–61. https://doi.org/10.1177/1071181321651052.

Holzinger, Andreas, Langs, Georg, Denk, Helmut, Zatloukal, Kurt, Müller, Heimo, 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (4), e1312.

Hong, Matthew K., Fourney, Adam, DeBellis, Derek, Amershi, Saleema, 2021. Planning for natural language failures with the AI playbook. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Yokohama, Japan. Association for Computing Machinery, New York, NY, USA, Article 386, 11 pages. https://doi.org/10.1145/3411764.3445735.

Hong, Sungsoo Ray, Hullman, Jessica, Bertini, Enrico, 2020. Human factors in model interpretability: industry practices, challenges, and needs. Proceedings of the ACM on Human–Computer Interaction 4 (CSCW1), Article 68, 26 pages. https://doi.org/10.1145/3392878.

Hoover, Benjamin, Strobelt, Hendrik, Gehrmann, Sebastian, 2020. exBERT: a visual analysis tool to explore learned representations in transformer models. In: Celikyilmaz, Asli, Wen, Tsung-Hsien (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, pp. 187–196. https://doi.org/10.18653/v1/2020.acl-demos.22.

Horn, Colby, Manduca, Cathryn, Kauchak, David, 2014. Learning a lexical simplifier using Wikipedia. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 458–463.

Jameson, Anthony, Gabrielli, Silvia, Oulasvirta, Antti, 2009. Users' preferences regarding intelligent user interfaces: differences among users and changes over time. In: Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09). Sanibel Island, Florida, USA. Association for Computing Machinery, New York, NY, USA, pp. 497–498. https://doi.org/10.1145/1502650.1502734.

Jansen, Bernard J., Booth, Danielle L., Spink, Amanda, 2008. Determining the informational, navigational, and transactional intent of Web queries. Information Processing & Management 44 (3), 1251–1266. https://doi.org/10.1016/j.ipm.2007.07.015.

Jesus, Sérgio, Belém, Catarina, Balayan, Vladimir, Bento, João, Saleiro, Pedro, Bizarro, Pedro, Gama, João, 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Virtual Event, Canada. Association for Computing Machinery, New York, NY, USA, pp. 805–815. https://doi.org/10.1145/3442188.3445941.

Ji, Ziwei, Lee, Nayeon, Frieske, Rita, Yu, Tiezheng, Su, Dan, Xu, Yan, Ishii, Etsuko, Bang, Ye Jin, Madotto, Andrea, Fung, Pascale, 2023. Survey of hallucination in natural language generation. ACM Computing Surveys 55 (12), Article 248, 38 pages. https://doi.org/10.1145/3571730.

Jiang, Albert Q., Sablayrolles, Alexandre, Mensch, Arthur, Bamford, Chris, Chaplot, Devendra Singh, de las Casas, Diego, Bressand, Florian, Lengyel, Gianna, Lample, Guillaume, Saulnier, Lucile, Lavaud, Lélio Renard, Lachaux, Marie-Anne, Stock, Pierre, Le Scao, Teven, Lavril, Thibaut, Wang, Thomas, Lacroix, Timothée, El Sayed, William, 2023. Mistral 7B. arXiv:2310.06825 [cs.CL].

Jo, Eunkyung, Ryu, Myeonghan, Kenderova, Georgia, So, Samuel, Shapiro, Bryan, Papoutsaki, Alexandra, Epstein, Daniel A., 2022. Designing flexible longitudinal regimens: supporting clinician planning for discontinuation of psychiatric drugs. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). New Orleans, LA, USA. Association for Computing Machinery, New York, NY, USA, Article 352, 17 pages. https://doi.org/10.1145/3491102.3502206.

Karaca, Ozan, Çalışkan, S. Ayhan, Demir, Kadir, 2021. Medical artificial intelligence readiness scale for medical students (MAIRS-MS) – development, validity and reliability study. BMC Medical Education 21 (1), 112. https://doi.org/10.1186/s12909-021-02546-6.

Kaur, Harmanpreet, Nori, Harsha, Jenkins, Samuel, Caruana, Rich, Wallach, Hanna, Wortman Vaughan, Jennifer, 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA, pp. 1–14. https://doi.org/10.1145/3313831.3376219.

Kim, Sunnie S.Y., Watkins, Elizabeth Anne, Russakovsky, Olga, Fong, Ruth, Monroy-Hernández, Andrés, 2023. "Help me help the AI": understanding how explainability can support human-AI interaction. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages. https://doi.org/10.1145/3544548.3581001.

Koh, Pang Wei, Liang, Percy, 2017. Understanding black-box predictions via influence functions. In: Precup, Doina, Teh, Yee Whye (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70. PMLR, pp. 1885–1894. https://proceedings.mlr.press/v70/koh17a.html.

Krishna, Satyapriya, Han, Tessa, Gu, Alex, Pombra, Javin, Jabbari, Shahin, Wu, Steven, Lakkaraju, Himabindu, 2022. The disagreement problem in explainable machine learning: a practitioner's perspective. CoRR arXiv:2202.01602. https://arxiv.org/abs/2202.01602.

Kulesza, Todd, Burnett, Margaret, Wong, Weng-Keen, Stumpf, Simone, 2015. Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15). Atlanta, Georgia, USA. Association for Computing Machinery, New York, NY, USA, pp. 126–137. https://doi.org/10.1145/2678025.2701399.

Kulesza, Todd, Stumpf, Simone, Burnett, Margaret, Wong, Weng-Keen, Riche, Yann, Moore, Travis, Oberst, Ian, Shinsel, Amber, McIntosh, Kevin, 2010. Explanatory debugging: supporting end-user debugging of machine-learned programs. In: 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 41–48. https://doi.org/10.1109/VLHCC.2010.15.

Königstorfer, Florian, Thalmann, Stefan, 2020. Applications of artificial intelligence in commercial banks – a research agenda for behavioral finance. Journal of Behavioral and Experimental Finance 27, 100352. https://doi.org/10.1016/j.jbef.2020.100352.

Lai, Vivian, Zhang, Yiming, Chen, Chacha, Liao, Q. Vera, Tan, Chenhao, 2023. Selective explanations: leveraging human input to align explainable AI. arXiv:2301.09656 [cs.AI].

Langer, Markus, Oster, Daniel, Speith, Timo, Hermanns, Holger, Kästner, Lena, Schmidt, Eva, Sesing, Andreas, Baum, Kevin, 2021. What do we want from explainable artificial intelligence (XAI)? – a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296, 103473. https://doi.org/10.1016/j.artint.2021.103473.

Laupichler, Matthias Carl, Aster, Alexandra, Schirch, Jana, Raupach, Tobias, 2022. Artificial intelligence literacy in higher and adult education: a scoping literature review. Computers and Education: Artificial Intelligence 3, 100101. https://doi.org/10.1016/j.caeai.2022.100101.

Lertvittayakumjorn, Piyawat, Specia, Lucia, Toni, Francesca, 2020. FIND: human-in-the-loop debugging deep text classifiers. In: Webber, Bonnie, Cohn, Trevor, He, Yulan, Liu, Yang (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp. 332–348. https://doi.org/10.18653/v1/2020.emnlp-main.24.

Lertvittayakumjorn, Piyawat, Toni, Francesca, 2021. Explanation-based human debugging of NLP models: a survey. Transactions of the Association for Computational Linguistics 9, 1508–1528. https://doi.org/10.1162/tacl_a_00440. https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00440/1983435/tacl_a_00440.pdf.

Liang, Paul Pu, Lyu, Yiwei, Chhablani, Gunjan, Jain, Nihal, Deng, Zihao, Wang, Xingbo, Morency, Louis-Philippe, Salakhutdinov, Ruslan, 2023. MultiViz: towards user-centric visualizations and interpretations of multimodal models. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 214, 21 pages. https://doi.org/10.1145/3544549.3585604.

Liao, Q. Vera, Gruen, Daniel, Miller, Sarah, 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 (CHI '20). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA, pp. 1–15. https://doi.org/10.1145/3313831.3376590.

Liao, Q. Vera, Zhang, Yunfeng, Luss, Ronny, Doshi-Velez, Finale, Dhurandhar, Amit, 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 10 (1), 147–159. https://doi.org/10.1609/hcomp.v10i1.21995.

Liesenfeld, Andreas, Dingemanse, Mark, 2024. Rethinking open source generative AI: open washing and the EU AI act. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Rio de Janeiro, Brazil. Association for Computing Machinery, New York, NY, USA, pp. 1774–1787. https://doi.org/10.1145/3630106.3659005.

Liquin, Emily G., Lombrozo, Tania, 2020. A functional approach to explanation-seeking curiosity. Cognitive Psychology 119, 101276. https://doi.org/10.1016/j.cogpsych.2020.101276.

Lombrozo, Tania, 2010. Causal–explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions. Cognitive Psychology 61 (4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002.

Lombrozo, Tania, 2012. Explanation and abductive inference. In: Oxford Handbook of Thinking and Reasoning, pp. 260–276.

Lombrozo, Tania, Liquin, Emily G., 2023. Explanation is effective because it is selective. Current Directions in Psychological Science 32 (3), 212–219. https://doi.org/10.1177/09637214231156106.

Long, Duri, Magerko, Brian, 2020. What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA, pp. 1–16. https://doi.org/10.1145/3313831.3376727.

Lourenço, Raoni, Freire, Juliana, Shasha, Dennis, 2019. Debugging machine learning pipelines. In: Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning (DEEM'19). Amsterdam, Netherlands. Association for Computing Machinery, New York, NY, USA, Article 3, 10 pages. https://doi.org/10.1145/3329486.3329489.

Lötsch, Jörn, Kringel, Dario, Ultsch, Alfred, 2022. Explainable artificial intelligence (XAI) in biomedicine: making AI decisions trustworthy for physicians and patients. BioMedInformatics 2 (1), 1–17. https://doi.org/10.3390/biomedinformatics2010001.

Madsen, Andreas, Chandar, Sarath, Reddy, Siva, 2024. Are self-explanations from large language models faithful? arXiv:2401.07927 [cs.CL].

Maharjan, Raju, Rohani, Darius Adam, Bækgaard, Per, Bardram, Jakob, Doherty, Kevin, 2021. Can we talk? Design implications for the questionnaire-driven self-report of health and wellbeing via conversational agent. In: Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21). Bilbao (online), Spain. Association for Computing Machinery, New York, NY, USA, Article 5, 11 pages. https://doi.org/10.1145/3469595.3469600.

Markus, Aniek F., Kors, Jan A., Rijnbeek, Peter R., 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics 113, 103655. https://doi.org/10.1016/j.jbi.2020.103655.

Mazo, Claudia, Kearns, Cathriona, Mooney, Catherine, Gallagher, William M., 2020. Clinical decision support systems in breast cancer: a systematic review. Cancers 12 (2). https://doi.org/10.3390/cancers12020369.

Mehrabi, Ninareh, Morstatter, Fred, Saxena, Nripsuta, Lerman, Kristina, Galstyan, Aram, 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys 54 (6), Article 115, 35 pages. https://doi.org/10.1145/3457607.

Meta, 2023. Llama 2: open foundation and fine-tuned chat models. arXiv:2307.09288 [cs.CL].

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., Dean, Jeff, 2013. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119.

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.

Miller, Tim, 2023. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Chicago, IL, USA. Association for Computing Machinery, New York, NY, USA, pp. 333–342. https://doi.org/10.1145/3593013.3594001.

Mitchell, Margaret, Wu, Simone, Zaldivar, Andrew, Barnes, Parker, Vasserman, Lucy, Hutchinson, Ben, Spitzer, Elena, Raji, Inioluwa Deborah, Gebru, Timnit, 2019. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Atlanta, GA, USA. Association for Computing Machinery, New York, NY, USA, pp. 220–229. https://doi.org/10.1145/3287560.3287596.

Mohseni, Sina, Zarei, Niloofar, Ragan, Eric D., 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. The ACM Transactions on Interactive Intelligent Systems 11 (3–4), Article 24, 45 pages. https://doi.org/10.1145/3387166.

Nanda, Neel, Bloom, Joseph, 2022. TransformerLens. https://github.com/neelnanda-io/TransformerLens.

Nauta, Meike, van Bree, Ron, Seifert, Christin, 2021. Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14933–14943.

Nori, Harsha, Jenkins, Samuel, Koch, Paul, Caruana, Rich, 2019. InterpretML: a unified framework for machine learning interpretability. arXiv preprint. arXiv:1909.09223.

OpenAI, 2023. ChatGPT. https://chat.openai.com/chat. (Accessed 19 June 2024).

Osheroff, Jerome A., Teich, Jonathan M., Middleton, Blackford, Steen, Elaine B., Wright, Adam, Detmer, Don E., 2007. A roadmap for national action on clinical decision support. Journal of the American Medical Informatics Association 14 (2), 141–145. https://doi.org/10.1197/jamia.M2334. https://academic.oup.com/jamia/article-pdf/14/2/141/2563507/14-2-141.pdf.

Park, Joon Sung, O'Brien, Joseph, Cai, Carrie Jun, Morris, Meredith Ringel, Liang, Percy, Bernstein, Michael S., 2023. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). San Francisco, CA, USA. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763.

Park, Soya, Wang, April Yi, Kawas, Ban, Liao, Q. Vera, Piorkowski, David, Danilevsky, Marina, 2021. Facilitating knowledge sharing from domain experts to data scientists for building NLP models. In: 26th International Conference on Intelligent User Interfaces (IUI '21). College Station, TX, USA. Association for Computing Machinery, New York, NY, USA, pp. 585–596. https://doi.org/10.1145/3397481.3450637.

Pekarek Doehler, Simona, Berger, Evelyne, 2019. On the reflexive relation between developing L2 interactional competence and evolving social relationships: a longitudinal study of word-searches in the 'wild'. In: Conversation Analytic

Research on Learning-in-Action. Springer International Publishing, Cham, pp. 51–75. https://doi.org/10.1007/978-3-030-22165-2_3.

Pocevičiūtė, Milda, Eilertsen, Gabriel, Lundström, Claes, 2020. Survey of XAI in Digital Pathology. In: Artificial Intelligence and Machine Learning for Digital Pathology. Springer International Publishing, Cham, pp. 56–88. https://doi.org/10.1007/978-3-030-50402-1_4.

Pradhan, Alisha, Lazar, Amanda, Findlater, Leah, 2020. Use of intelligent voice assistants by older adults with low technology use. ACM Transactions on Computer-Human Interaction 27 (4), Article 31, 27 pages. https://doi.org/10.1145/3373759.

Preece, Alun, Harborne, Dan, Braines, Dave, Tomsett, Richard, Chakraborty, Supriyo, 2018. Stakeholders in explainable AI. arXiv:1810.00184 [cs.AI].

Rahwan, Iyad, Cebrian, Manuel, Obradovich, Nick, Bongard, Josh, Bonnefon, Jean-François, Breazeal, Cynthia, Crandall, Jacob W., Christakis, Nicholas A., Couzin, Iain D., Jackson, Matthew O., Jennings, Nicholas R., Kamar, Ece, Kloumann, Isabel M., Larochelle, Hugo, Lazer, David, McElreath, Richard, Mislove, Alan, Parkes, David C., Pentland, Alex 'Sandy', Roberts, Margaret E., Shariff, Azim, Tenenbaum, Joshua B., Wellman, Michael, 2019. Machine behaviour. Nature 568 (7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2016. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). San Francisco, California, USA. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2018. Anchors: high-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence 32 (1). https://doi.org/10.1609/aaai.v32i1.11491.

Ribeiro, Marco Tulio, Wu, Tongshuang, Guestrin, Carlos, Singh, Sameer, 2020. Beyond accuracy: behavioral testing of NLP models with CheckList. In: Jurafsky, Dan, Chai, Joyce, Schluter, Natalie, Tetreault, Joel (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442.

Rostamzadeh, Negar, Mincu, Diana, Roy, Subhrajit, Smart, Andrew, Wilcox, Lauren, Pushkarna, Mahima, Schrouff, Jessica, Amironesei, Razvan, Moorosi, Nyalleng, Heller, Katherine, 20022. Healthsheet: development of a transparency artifact for health datasets. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Seoul, Republic of Korea. Association for Computing Machinery, New York, NY, USA, pp. 1943–1961. https://doi.org/10.1145/3531146.3533239.

Sarti, Gabriele, Feldhus, Nils, Sickert, Ludwig, van der Wal, Oskar, 2023. Inseq: an interpretability toolkit for sequence generation models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-demo.40.

Schaaf, Jannik, Sedlmayr, Martin, Schaefer, Johanna, Storf, Holger, 2020. Diagnosis of rare diseases: a scoping review of clinical decision support systems. Orphanet Journal of Rare Diseases 15 (1), 263. https://doi.org/10.1186/s13023-020-01536-z.

Sculley, D., Holt, Gary, Golovin, Daniel, Davydov, Eugene, Phillips, Todd, Ebner, Dietmar, Chaudhary, Vinay, Young, Michael, Crespo, Jean-François, Dennison,

Dan, 2015. Hidden technical debt in machine learning systems. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

Selsam, Daniel, Liang, Percy, Dill, David L., 2017. Developing bug-free machine learning systems with formal mathematics. In: Precup, Doina, Teh, Yee Whye (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70. PMLR, pp. 3047–3056. https://proceedings.mlr.press/v70/selsam17a.html.

Selvaraju, Ramprasaath R., Cogswell, Michael, Das, Abhishek, Vedantam, Ramakrishna, Parikh, Devi, Batra, Dhruv, 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128 (2), 336–359. https://doi.org/10.1007/s11263-019-01228-7.

Shah, Chirag, Bender, Emily M., 2024. Envisioning information access systems: what makes for good tools and a healthy web? ACM Transactions on the Web 18 (3), Article 33, 24 pages. https://doi.org/10.1145/3649468.

Sharifi Noorian, Shahin, Qiu, Sihang, Gadiraju, Ujwal, Yang, Jie, Bozzon, Alessandro, 2022. What should you know? A human-in-the-loop approach to unknown unknowns characterization in image recognition. In: Proceedings of the ACM Web Conference 2022 (WWW '22). Virtual Event, Lyon, France. Association for Computing Machinery, New York, NY, USA, pp. 882–892. https://doi.org/10.1145/3485447.3512040.

Singh, Amitojdeep, Sengupta, Sourya, Lakshminarayanan, Vasudevan, 2020. Explainable deep learning models in medical image analysis. Journal of Imaging 6 (6). https://doi.org/10.3390/jimaging6060052.

Sivaraman, Venkatesh, Bukowski, Leigh A., Levin, Joel, Kahn, Jeremy M., Perer, Adam, 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075.

Slack, Dylan, Hilgard, Sophie, Jia, Emily, Singh, Sameer, Lakkaraju, Himabindu, 2020. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). New York, NY, USA. Association for Computing Machinery, New York, NY, USA, pp. 180–186. https://doi.org/10.1145/3375627.3375830.

Sliwa, Julia, 2021. Toward collective animal neuroscience. Science 374 (6566), 397–398. https://doi.org/10.1126/science.abm3060.

Smith, Helen, 2021. Clinical AI: opacity, accountability, responsibility and liability. AI & Society 36 (2), 535–545. https://doi.org/10.1007/s00146-020-01019-6.

Sokol, Kacper, Flach, Peter, 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Barcelona, Spain. Association for Computing Machinery, New York, NY, USA, pp. 56–67. https://doi.org/10.1145/3351095.3372870.

Song, Da, Wang, Zhijie, Huang, Yuheng, Ma, Lei, Zhang, Tianyi, 2023. DeepLens: interactive out-of-distribution data detection in NLP models. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 739, 17 pages. https://doi.org/10.1145/3544548.3580741.

Specia, Lucia, 2010. Translating from complex to simplified sentences. In: International Conference on Computational Processing of the Portuguese

Language. Springer, pp. 30–39.

Speer, Robyn, Chin, Joshua, Havasi, Catherine, 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. Proceedings of the AAAI Conference on Artificial Intelligence 31 (1). https://doi.org/10.1609/aaai.v31i1.11164.

Subramonyam, Hariharan, Im, Jane, Seifert, Colleen, Adar, Eytan, 2022. Solving separation-of-concerns problems in collaborative design of human-AI systems through leaky abstractions. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). New Orleans, LA, USA. Association for Computing Machinery, New York, NY, USA, Article 481, 21 pages. https://doi.org/10.1145/3491102.3517537.

Sundararajan, Mukund, Taly, Ankur, Yan, Qiqi, 2017. Axiomatic attribution for deep networks. In: Precup, Doina, Teh, Yee Whye (Eds.), Proceedings of the 34th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 70. PMLR, pp. 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html.

Tabard, Aurélien, Mackay, Wendy E., Eastmond, Evelyn, 2008. From individual to collaborative: the evolution of prism, a hybrid laboratory notebook. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08). San Diego, CA, USA. Association for Computing Machinery, New York, NY, USA, pp. 569–578. https://doi.org/10.1145/1460563.1460653.

Teevan, Jaime, Dumais, Susan T., Liebling, Daniel J., 2010. A longitudinal study of how highlighting web content change affects people's web interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10). Atlanta, Georgia, USA. Association for Computing Machinery, New York, NY, USA, pp. 1353–1356. https://doi.org/10.1145/1753326.1753530.

Tenney, Ian, Wexler, James, Bastings, Jasmijn, Bolukbasi, Tolga, Coenen, Andy, Gehrmann, Sebastian, Jiang, Ellen, Pushkarna, Mahima, Radebaugh, Carey, Reif, Emily, Yuan, Ann, 2020. The language interpretability tool: extensible, interactive visualizations and analysis for NLP models. In: Liu, Qun, Schlangen, David (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, pp. 107–118. https://doi.org/10.18653/v1/2020.emnlp-demos.15.

Thieme, Anja, Nori, Aditya, Ghassemi, Marzyeh, Bommasani, Rishi, Andersen, Tariq Osman, Luger, Ewa, 2023. Foundation models in healthcare: opportunities, risks & strategies forward. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 512, 4 pages. https://doi.org/10.1145/3544549.3583177.

Tocchetti, Andrea, Corti, Lorenzo, Balayn, Agathe, Yurrita, Mireia, Lippmann, Philip, Brambilla, Marco, Yang, Jie, 2025. A.I. robustness: a human-centered perspective on technological challenges and opportunities. ACM Computing Surveys 57 (6), 141. https://doi.org/10.1145/3665926.

Tocchetti, Andrea, Yang, Jie, Brambilla, Marco, et al., 2023. Rationale trees: towards a formalization of human knowledge for explainable natural language processing. In: Proceedings of the 4th Italian Workshop on Explainable Artificial Intelligence co-located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), vol. 3518. CEUR-WS.org, pp. 29–46.

Tomsett, Richard, Braines, Dave, Harborne, Dan, Preece, Alun, Chakraborty, Supriyo, 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv:1806.07552 [cs.AI].

Tonekaboni, Sana, Joshi, Shalmali, McCradden, Melissa D., Goldenberg, Anna, 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Doshi-Velez, Finale, Fackler, Jim, Jung, Ken, Kale, David, Ranganath, Rajesh, Wallace, Byron, Wiens, Jenna (Eds.), Proceedings of the 4th Machine Learning for Healthcare Conference. In: Proceedings of Machine Learning Research, vol. 106. PMLR, pp. 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html.

Topol, Eric J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 25 (1), 44–56. https://doi.org/10.1038/s41591-018-0300-7.

Tversky, Amos, Kahneman, Daniel, 1974. Judgment under uncertainty: heuristics and biases. Science 185 (4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124. https://www.science.org/doi/pdf/10.1126/science.185.4157.1124.

UK Government, 2022. Accessible communication formats. https://www.gov.uk/government/publications/inclusive-communication/accessible-communication-formats.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. In: Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vaughan, Misha, Courage, Catherine, 2007. SIG: capturing longitudinal usability: what really affects user performance over time? In: CHI '07 Extended Abstracts on Human Factors in Computing Systems (CHI EA '07). San Jose, CA, USA. Association for Computing Machinery, New York, NY, USA, pp. 2149–2152. https://doi.org/10.1145/1240866.1240970.

Vaughan, Misha, Courage, Catherine, Rosenbaum, Stephanie, Jain, Jhilmil, Hammontree, Monty, Beale, Russell, Welsh, Dan, 2008. Longitudinal usability data collection: art versus science? In: CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08). Florence, Italy. Association for Computing Machinery, New York, NY, USA, pp. 2261–2264. https://doi.org/10.1145/1358628.1358664.

Velickovski, Filip, Ceccaroni, Luigi, Roca, Josep, Burgos, Felip, Galdiz, Juan B., Marina, Nuria, Lluch-Ariet, Magí, 2014. Clinical decision support systems (CDSS) for preventive management of COPD patients. Journal of Translational Medicine 12 (2), S9. https://doi.org/10.1186/1479-5876-12-S2-S9.

Vygotsky, Lev S., 2012. Thought and Language. MIT Press.

Wachter, Sandra, Mittelstadt, Brent, Russell, Chris, 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard Journal of Law & Technology (Harvard JOLT) 31 (2), 841–888. https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=859.

Wallace, Eric, Tuyls, Jens, Wang, Junlin, Subramanian, Sanjay, Gardner, Matt, Singh, Sameer, 2019. AllenNLP interpret: a framework for explaining predictions of NLP models. In: Padó, Sebastian, Huang, Ruihong (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Association for Computational Linguistics, Hong Kong, China, pp. 7–12. https://doi.org/10.18653/v1/D19-3002.

Walsh, Seán, de Jong, Evelyn E.C., van Timmeren, Janna E., Ibrahim, Abdalla, Compter, Inge, Peerlings, Jurgen, Sanduleanu, Sebastian, Refaee, Turkey, Keek,

Simon, Larue, Ruben T.H.M., van Wijk, Yvonka, Even, Aniek J.G., Jochems, Arthur, Barakat, Mohamed S., Leijenaar, Ralph T.H., Lambin, Philippe, 2019. Decision support systems in oncology. JCO Clinical Cancer Informatics 3, 1–9. https://doi.org/10.1200/CCI.18.00001. PMID: 30730766.

Wang, Jindong, Xixu, Hu, Hou, Wenxin, Chen, Hao, Zheng, Runkai, Wang, Yidong, Yang, Linyi, Ye, Wei, Huang, Haojun, Geng, Xiubo, et al., 2023. On the robustness of ChatGPT: an adversarial and out-of-distribution perspective. In: ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models.

Wang, Tong, Chen, Ping, Amaral, Kevin, Qiang, Jipeng, 2016. An experimental study of LSTM encoder-decoder model for text simplification. arXiv preprint. arXiv:1609.03663.

Wood, Elena A., Ange, Brittany L., Miller, D. Douglas, 2021. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. Journal of Medical Education and Curricular Development 8, 23821205211024078. https://doi.org/10.1177/23821205211024078. PMID: 34250242.

Wu, Tongshuang, Zhu, Haiyi, Albayrak, Maya, Axon, Alexis, Bertsch, Amanda, Deng, Wenxing, Ding, Ziqi, Guo, Bill, Gururaja, Sireesh, Kuo, Tzu-Sheng, Liang, Jenny T., Liu, Ryan, Mandal, Ihita, Milbauer, Jeremiah, Ni, Xiaolin, Padmanabhan, Namrata, Ramkumar, Subhashini, Sudjianto, Alexis, Taylor, Jordan, Tseng, Ying-Jui, Vaidos, Patricia, Wu, Zhijin, Wu, Wei, Yang, Chenyang, 2023. LLMs as workers in human-computational algorithms? Replicating crowdsourcing pipelines with LLMs. arXiv:2307.10168 [cs.CL].

Xie, Yao, Chen, Melody, Kao, David, Gao, Ge, Chen, Xiang 'Anthony', 2020. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Honolulu, HI, USA. Association for Computing Machinery, New York, NY, USA, pp. 1–13. https://doi.org/10.1145/3313831.3376807.

Xu, Xuhai, Yu, Anna, Jonker, Tanya R., Todi, Kashyap, Lu, Feiyu, Qian, Xun, Evangelista Belo, João Marcelo, Wang, Tianyi, Li, Michelle, Mun, Aran, Wu, Te-Yen, Shen, Junxiao, Zhang, Ting, Kokhlikyan, Narine, Wang, Fulton, Sorenson, Paul, Kim, Sophie, Benko, Hrvoje, 2023. XAIR: a framework of explainable AI in augmented reality. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 202, 30 pages. https://doi.org/10.1145/3544548.3581500.

Yang, Qian, Steinfeld, Aaron, Zimmerman, John, 2019. Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Glasgow, Scotland, UK. Association for Computing Machinery, New York, NY, USA, pp. 1–11. https://doi.org/10.1145/3290605.3300468.

Yildirim, Nur, Kass, Alex, Tung, Teresa, Upton, Connor, Costello, Donnacha, Giusti, Robert, Lacin, Sinem, Lovic, Sara, O'Neill, James M., Meehan, Rudi O'Reilly, Loideáin, Eoin Ó, Pini, Azzurra, Corcoran, Medb, Hayes, Jeremiah, Cahalane, Diarmuid J., Shivhare, Gaurav, Castoro, Luigi, Caruso, Giovanni, Oh, Changhoon, McCann, James, Forlizzi, Jodi, Zimmerman, John, 2022. How experienced designers of enterprise applications engage AI as a design material. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). New Orleans, LA, USA. Association for Computing Machinery, New York, NY, USA, Article 483, 13 pages. https://doi.org/10.1145/3491102.3517491.

Yuan, Jun, Vig, Jesse, Rajani, Nazneen, 2022. ISEA: an interactive pipeline for semantic error analysis of NLP models. In: 27th International Conference on Intelligent User Interfaces (IUI '22). Helsinki, Finland. Association for Computing Machinery, New York, NY, USA, pp. 878–888. https://doi.org/10.1145/3490099.3511146.

Yurrita, Mireia, Draws, Tim, Balayn, Agathe, Murray-Rust, Dave, Tintarev, Nava, Bozzon, Alessandro, 2023. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Hamburg, Germany. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. https://doi.org/10.1145/3544548.3581161.

Zhang, Xingxing, Lapata, Mirella, 2017. Sentence simplification with deep reinforcement learning. arXiv preprint. arXiv:1703.10931.

Zhang, Zijian, Rudra, Koustav, Anand, Avishek, 2021. Explain and predict, and then predict again. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21). Virtual Event, Israel. Association for Computing Machinery, New York, NY, USA, pp. 418–426. https://doi.org/10.1145/3437963.3441758.

Zhang, Zijian, Singh, Jaspreet, Gadiraju, Ujwal, Anand, Avishek, 2019. Dissonance between human and machine understanding. Proceedings of the ACM on Human–Computer Interaction 3 (CSCW), Article 56, 23 pages. https://doi.org/10.1145/3359158.

Zimmerman, John, Forlizzi, Jodi, 2014. Research Through Design in HCI. Springer, New York, NY, pp. 167–189. https://doi.org/10.1007/978-1-4939-0378-8_8.