Who's Thinking? : A Push for Human-Centered Evaluation of LLMs using the XAI Playbook

Anonymous Authors

ABSTRACT

Deployed artificial intelligence (AI) often impacts humans, and there is no one-size-fits-all metric to evaluate these tools. Human-centered evaluation of AI-based systems combines quantitative and qualitative analysis and human input. It has been explored to some depth in the explainable AI (XAI) and human-computer interaction (HCI) communities. Gaps remain, but the basic understanding that humans interact with AI and accompanying explanations, and that humans' needs—complete with their cognitive biases and quirks—should be held front and center, is accepted by the community. In this paper, we draw parallels between the relatively mature field of XAI and the rapidly evolving research boom around large language models (LLMs). Accepted evaluative metrics for LLMs are not human-centered. We argue that many of the same paths tread by the XAI community over the past decade will be retread when discussing LLMs. Specifically, we argue that humans' tendencies—again, complete with their cognitive biases and quirks—should rest front and center when evaluating deployed LLMs. We outline three developed focus areas of human-centered evaluation of XAI: mental models, use case utility, and cognitive engagement, and we highlight the importance of exploring each of these concepts for LLMs. Our goal is to jumpstart human-centered LLM evaluation.

CCS CONCEPTS

• **Human-centered computing** → **User models**; *HCI theory, concepts and models*; • **Computing methodologies** → Philosophical/theoretical foundations of artificial intelligence; **Artificial intelligence**.

KEYWORDS

large language models, cognition, human-centered evaluation

1 INTRODUCTION

In this paper, we first provide an overview of Explainable AI (XAI) and its evaluation strategies. We then provide an overview of Large Language Models (LLMs) and their current evaluation frameworks.

We draw parallels between XAI and the rapidly booming field of Generative AI and LLMs, and we argue that three focus areas of human-centered evaluation of XAI must also be applied to LLM evaluation.

We encourage work that:

- seeks to understand the **mental models** of LLMs that everyday users hold
- evaluates the utility of LLMs via **use case** based user studies
- documents how **cognitively engaged** users are with LLM outputs and how this affects user and system behavior

These are vital for evaluating how LLMs meet real human needs (with their cognitive quirks) and work within real human environments. We provide a brief primer on XAI and LLMs in the succeeding sections, then support our case that the lessons learned in XAI research may largely transfer to those in the more nascent LLM space.

1.1 Explainable AI

Defining explainability and the related notion of interpretability for machine learning (ML) models is not black and white, and is a subject of ongoing discussion.[1] For the purposes of our discussion, we focus on the most common type of model transparency seen in industry: post-hoc explanations of black box models. Post-hoc explainers often rely on only a trained model's inputs and outputs to identify patterns in how the model makes decisions. These methods aim to unlock transparency in order to allow stakeholders to understand the decision-making process of models to improve trust and mitigate downstream harms [16, 27].

One key consideration for XAI is that there is no such thing as a ground truth for explanations. While classic supervised ML problems have ground truth labels, there is no exactly-correct explanation for how a black box produces outputs, for all but the simplest models. This lack of ground truth means that classic accuracy-based metrics cannot be measured for XAI. Instead, a variety of measurable, quantitative properties have been developed to evaluate how "good" an XAI method is. These properties [13] include: *fidelity/faithfulness* (how consistent are explanations to the underlying black box), *robustness* (how consistent are explanations to infinitesimally small changes in the input), *compactness* (how measurably complex are the explanations), and *homogeneity* (how do the qualities of the explanations differ across subgroups). These different properties target different notions of what a practitioner might want from an explanation, and thus can sometimes be in

---

[1]We make no normative judgments with respect to the ongoing debate regarding explainability versus interpretability, and redirect the interested reader to the work of [16, 27, 48] for a primer on that discussion. Rather, we acknowledge that XAI—for all of its potential pitfalls and flaws—has found a home in production use of models in industry, thus spurring on a wide variety of research programs covering usability, efficacy, and validity. Indeed, the existence of that increasingly-mature decade-long body of research motivates this paper, where we predict a similar pattern of disagreeing lines of analysis maturing in parallel with respect to LLMs.

trade-off with one another [13]. Further, these measurable properties largely only explore the technical validity of the produced explanations. It is vital to acknowledge that there are other areas of consideration beyond these mathematical and technical formulations.

In practice, practitioners often use XAI as an assistive tool for accomplishing some downstream task—model debugging, generating hypotheses, ensuring compliance, etc [see., e.g. 36]. This means that the context of XAI use involves some human using an explainer as a piece of evidence to make some decision. Thus, we need to consider how practitioners use, receive, and comprehend outputted explanations. This is especially true because data scientists are humans, and thus susceptible to cognitive biases such as confirmation bias. Scenarios have been identified where explanations are found to be misleading and lead to misplaced trust in faulty models [12, 31]. In human-AI teams in the medical domain, studies have shown clinicians making more errors than usual when the AI system produces incorrect predictions [26]. **Human-centered evaluation** of XAI focuses on defining the goals of XAI to meet real human needs and to work within real human environments. This means acknowledging the lack of perfect rationality in human decision-making.

1.2 Large Language Models

In the last year, Large Language Models (LLMs) have exploded in popularity—from research, to industry, to public use and, perhaps more importantly, public awareness and accessibility. These unsupervised autoregressive models predict next tokens (character, word, or string) based on past context [29]. They can generate text for a variety of traditional generative tasks including summarization [53], dialogue [43, 44], and more. Language models have grown increasingly *larger* under pressures to increase performance. Two common identifiers of how "large" a language model is are its number of parameters and the size of the dataset it was trained on [see, e.g., 6, for in-depth coverage]. Using two then-state-of-the-art (SotA) models as examples, in 2019 BERT [15] featured $O(10^8)$ parameters and was trained on a dataset size of 16 GB. In 2022, only 3 years later, GPT-3 [43] was released, featuring $O(10^{11})$ parameters and a training dataset size of 570 GB. We, again, make no normative statements regarding the "scale is all you need" camp; rather, we note these are broadly acknowledged metrics in the community.

Current LLM evaluation mechanisms include quantitative metrics measuring notions of *accuracy* (how similar are the generated outputs to the expected outputs), *robustness* (how resilient is the model to transformations of the input), *calibration* (how meaningful are the generated probabilities in respect to uncertainty), *efficiency* (what are the energy, carbon, and time costs for training and inference) and more [35]. Some also go beyond singular training-time loss objectives and implement reinforcement learning using human feedback in the loop [33].

A variety of potential harms and failure modes for LLMs have also been identified. There are substantial environmental costs associated with the volume of computational power required for training and inference [6, 49]. There are also concerns of LLMs propagating unfairness or bias. Counterfactual fairness [24] examines how perturbing the demographic signals of existing test examples

can change the performance of the model (e.g. "He worked at the local hospital" versus "She worked at the local hospital"). Fairness can also be evaluated via performance disparities between demographic subgroups. There are also other concerning forms of biases such as stereotypical associations, erasure, and over-representation in the semantics of its output [35, 38]. Finally, LLMs have been shown to produce toxic outputs. Toxicity in this context refers to hateful, violent, or offensive text [46], and has been shown to result even when the text prompt input is not itself toxic [19].

LLMs often suffer from factual errors—they can "hallucinate" information [50] by providing very confident-sounding but entirely false responses. Chatbot LLMs have also been found to engage in disturbingly emotional personal conversations when session lengths are not limited [47]. There are also privacy concerns- work has shown that LLMs are susceptible to training data leakage under adversarial attack [11].

Although many concerning issues of LLMs have been identified, these largely only reflect distributional or instance-wise properties of their outputs. Zooming out on a human-centered paradigm, the ML community has not yet devoted efforts to consider how the presence (and potential ubiquity) of this new technology may fundamentally affect society and the ways that individual humans make decisions. We hope this work jumpstarts efforts to address this gap.

2 PARALLELS BETWEEN XAI AND LLMS

When examining LLMs within a human-centered paradigm, we notice that like XAI outputs, LLM outputs are often meant for some downstream decision or task—what email to send to your client, what quick summary of an important document you will read, what answer is provided for a pertinent question. These use cases can range from the high-stakes to the everyday, but align in that the outputs of the AI system—the XAI attachment to a traditional machine-learning-based system, or an LLM) are not the human's end goal. Rather, they are a tool to help someone accomplish some other task, and so we must evaluate how these AI outputs help or hurt human interests. Another key similarity between XAI and LLMs is that they are both inherently open-ended systems, so there is typically no ground truth. Although recent work has argued for moving past an accuracy-first ML paradigm [25], XAI and LLMs force this to be the case with the lack of ground truth labels.

For these reasons, frameworks of human-centered qualitative evaluation must be employed. For XAI, this is a developed field of study with an emphasis on a multitude of aspects of cognition. For LLMs, this is not the case and these qualitative aspects of evaluation have not yet been explored. **We argue that the areas of human interaction and cognition that have been dominant in XAI and HCI communities need to be held front and center for LLM research.** We need to understand how users make decisions about whether to utilize the outputs of LLMs, the mental models that users have of these technologies, whether LLM outputs are actually helpful in downstream tasks, and how much users cognitively engage with the outputs to verify their correctness and lack of harm. We dive into these three areas of concern in the following section.

3 FRAMEWORK OF QUALITATIVE EVALUATION

3.1 Mental Model Matching

A user's mental model [42] of a technology is their internal understanding of how a technology works. For instance, maybe your mental model of a crosswalk is that pushing the crosswalk button will cause the walk signal to appear more quickly. However, for many cities, that button does not actually do anything [39]. This is an example of a user's mental model not aligning with a technology's true model. People rely heavily on their mental models of technology to make decisions. It has been found that XAI stakeholders use their mental models of XAI to decide when to use the technology [10], to evaluate how much to trust the outputted explanations [10, 20, 22], and to make sense of any results [22, 30].

These personal mental models are formulated from a user's perceptions and interactions with the technology and how they believe the system works. While ML practitioners may have had access to specialized training on how LLMs work, this is decidedly *not* the case for the vast majority of the general population. We cannot assume that everybody will have the same understanding of how a technology works as we do. To our knowledge, no work has explored the mental models the general public holds for any foundation model, let alone LLMs. How a general user believes an LLM to work may be very different from how it actually works, and this mismatch can be dangerous. It is not difficult to imagine frightening scenarios where users anthropomorphize or deify an LLM chatbot, understanding it to be a "magical" source of ground truth. This could very quickly lead to conspiracy theories and the legitimization of disinformation campaigns [see, e.g., 23]. It is important to consider if this is an issue of messaging and education—informing the public via AI literacy—or of regulators—to implement policies that force the algorithm providers to provide accurate, comprehensible warning labels on the limitations of their technology.

3.2 Use Case Utility

As previously discussed, XAI and LLMs are often tools for accomplishing some other goal. The term *use case* in the XAI literature [36] refers to a specific usage scenario and its associated downstream task or end goal. It has been found in the XAI literature that although it might be easy to assume that an explanation will be helpful for a user accomplishing a task like model debugging or model understanding, this is not necessarily the case [31]. When the performance of that downstream task is measured, the presence of explanations can sometimes have no effect, or can even have a negative effect on performance, especially if the XAI is faulty [4, 26]. Very limited work has explored the utility of LLMs in use-case–specified user studies, but a user study on Microsoft/Github's Copilot [1], an LLM-based code generation tool, found that it "did not necessarily improve the task completion time or success rate" [52].

In response to these findings, the XAI community has pushed for a context-based approach when building explanation methods [5, 36]. Advocates push for first identifying a specific use case, then understanding the types of transparency useful and relevant for each stakeholder in that context, then designing the explanation

method with these learnings held front and center, and finally evaluating how helpful the explanation was for specific tasks through practitioner user studies. In scenarios where the users are domain experts (e.g. medical treatment), this may involve long term collaboration to understand domain requirements and preferences.

There are also parallels with validity and alignment theory. A system is *valid* if it does what it purports to do. If it doesn't, it is likely because there are issues of *alignment* between the intent of the system builder and the property the algorithm optimizes [40]. Some work has flagged that it is irresponsible to publicly market LLMs for tasks they are not optimized for and were not designed for [6, 28]. For instance, dialogue LLM tools generate probabilistic outputs, not search results. Future works in this space could examine context-based design approaches and target alignment.

3.3 Cognitive Engagement

Cognitive effort is a form of labor, and unsurprisingly, people tend to favor less demanding forms of cognition and other mental shortcuts [18, 51]. As an example, when asked to "agree" to a user agreement when signing up for a new platform, you are probably more likely to check the box than to cognitively engage with the language of the agreement.

Unfortunately, this human tendency can lead to unintended or dangerous outcomes because humans are susceptible to a wide variety of cognitive biases such as confirmation bias. Confirmation bias [41] refers to the interpreting of new evidence in ways that confirm one's existing beliefs. For XAI, this manifests as practitioners only superficially examining explanations instead of digging deeply, leading to over-trust, misuse, and a lack of accurate understanding of the outputs [31]. This can be dangerous when it results in the over-confident deployment of a faulty or biased model. Further, user's need to reach a certain level of cognitive engagement before being able to learn or gain knowledge from the AI interaction [17]. The XAI and HCI communities have deeply, albeit not exhaustively, explored these phenomena and the difficulties of measuring it accurately [8, 31]. Those communities have also identified mitigation strategies, primarily through cognitive forcing strategies [9, 14, 32]. Forcing users to cognitively engage through some small task *before* showing a system's output yielded the highest performance in a comparative study [21]. This is not new or specific to XAI. Train conductors in Japan famously point and call out important information on their journeys—a cognitive forcing method which has reduced human errors by nearly 85% [45].

Issues of cognitive engagement should be held front and center for researchers of LLMs. Because of their massive scale and public accessibility, LLMs may quickly become ubiquitous in all aspects of daily life. Realistically, how much will users actually cognitively engage with the magnitude of generated outputs to ensure that they are correct and aligned with their intentions? Consider an LLM-generated email: how often and how deeply will a user review that generated email before sending it? What if it's not just one email, what if it's every email? Will they always catch when the generated output says something incorrect, or worse, inappropriate. Furthermore, our attention spans have decreased dramatically with the increase in digital stimulation[37]. The potential of cognitive

forcing strategies as a potential route for responsible LLM use should be explored.

Another aspect of concern is that LLM outputs often *sound* very confident, even if what they are saying is hallucinated [50]. When the user inquires about the incorrectness, they also have a documented tendency to argue that the user is wrong and that their response is correct. In fact, some have called LLMs "mansplaining as a service" [34]. This can make it more difficult for humans to implement cognitive checks on LLM outputs. While some recent LLM work has outlined categories of failure modes for LLMs based on the types of cognitive biases use [29], we push for greater work in this field. We believe that there are unknown consequences of offloading our cognitive load onto these agents.

4 WHY IS THIS IMPORTANT

While there are similarities between XAI and LLMs, there is at least one key difference—the potential scale of LLM usage is massive as a fundamentally lower-level ML-based building block than XAI. Their levels of performance have resulted in widespread public interest in practically all sectors of society. ChatGPT [43] set historic records for its customer growth, with over 100 million users in its first 2 months [2]. Unlike XAI, whose users are largely technical practitioners [7] focusing on one pipeline, LLMs are often designed for public use and are meant to help with a wide variety of tasks. People can repeatedly integrate LLMs into their daily lives in many ways—to decide what to eat for breakfast, to write the responses to emails left unanswered from yesterday, to develop the sales pitch they have to present mid-morning, to generate a funny poem during a break from work, etc. The ability to influence or make decisions is a form of inherent power, and offloading cognition onto AI agents must first be met with caution. Other works foreshadow potential existential threats that larger language models and the trajectory towards Artificial General Intelligence (AGI) can cause [3].

The consequences of not having a qualitative understanding of how humans interact with LLM outputs is grave. We urge for immediate efforts to address these gaps, and make a few suggestions for future works, taking inspiration from previous XAI research routes. We encourage work that seeks to understand the mental models of LLMs that everyday users hold, that evaluates the utility of LLMs via use case based user studies, and that documents how cognitive engagement affects user behavior. It is dangerous to continue developing and making available larger and larger language models without a proper understanding of how humans will (or will not) cognitively engage with its outputs.

5 CONCLUSION

In this paper, we push to jumpstart the human-centered evaluation of Large Language Models. We outline similarities between the more mature field of XAI and the nascent research boom of LLMs. We outline three focus areas of consideration- Mental Model Matching, Use Case Utility, and Cognitive Engagement. These all center on understanding real human needs and their cognitive quirks, the human-AI interaction effects on decision making tasks, and the potential dangers of offloading cognitive labor to these agents.

REFERENCES

[1] 2022. GitHub copilot · your AI pair programmer. https://github.com/features/copilot

[2] 2023. CHATGPT reaches 100 million users two months after launch. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[3] 2023. SATML 2023 - Timnit Gebru - eugenics and the promise of utopia through agi. https://www.youtube.com/watch?v=P7XT4TWLzJw

[4] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429* (2020).

[5] Katharina Beckh, Sebastian Müller, Matthias Jakobs, Vanessa Toborek, Hanxiao Tan, Raphael Fischer, Pascal Welke, Sebastian Houben, and Laura von Rueden. 2021. Explainable machine learning with prior knowledge: An overview. *arXiv preprint arXiv:2105.10172* (2021).

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[7] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces.* ACM. https://doi.org/10.1145/3377325.3377498

[9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[10] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. 2023. Improving Human-AI Collaboration with Descriptions of AI Behavior. *arXiv preprint arXiv:2301.06937* (2023).

[11] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models.. In *USENIX Security Symposium*, Vol. 6.

[12] Lok Chan, Kenzie Doyle, Duncan McElfresh, Vincent Conitzer, John P Dickerson, Jana Schaich Borg, and Walter Sinnott-Armstrong. 2020. Artificial Artificial Intelligence: Measuring Influence of AI 'Assessments' on Moral Decision-Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 214–220.

[13] Zixi Chen, Varshini Subhash, Marton Havasi, Weiwei Pan, and Finale Doshi-Velez. 2022. What Makes a Good Explanation?: A Harmonized View of Properties of Explanations. https://doi.org/10.48550/ARXIV.2211.05667

[14] Pat Croskerry. 2003. Cognitive forcing strategies in clinical decisionmaking. *Annals of emergency medicine* 41, 1 (2003), 110–120.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[16] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://doi.org/10.48550/ARXIV.1702.08608

[17] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces.* 794–806.

[18] Ellen C. Garbarino and Julie A. Edell. 1997. Cognitive Effort, Affect, and Choice. *Journal of Consumer Research* 24, 2 (Sept. 1997), 147–158. https://doi.org/10.1086/209500

[19] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. https://doi.org/10.48550/ARXIV.2009.11462

[20] Lisa Graichen, Matthias Graichen, and Mareike Petrosjan. 2022. How to Facilitate Mental Model Building and Mitigate Overtrust Using HCXAI.

[21] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[22] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference*

on Human Factors in Computing Systems (CHI'22).

[23] Tiffany Hsu and Stuart A. Thompson. 2023. Disinformation researchers raise alarms about A.I. Chatbots. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html

[24] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. https://doi.org/10.48550/ARXIV.1911.03064

[25] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[26] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry* 11, 1 (Feb. 2021). https://doi.org/10.1038/s41398-021-01224-x

[27] Alon Jacovi. 2023. Trends in Explainable AI (XAI) Literature. https://doi.org/10.48550/ARXIV.2301.05433

[28] Steven Johnson and Nikita Iziev. 2022. A.I. is mastering language. should we trust what it says? https://www.nytimes.com/2022/04/15/magazine/ai-language.html

[29] Erik Jones and Jacob Steinhardt. 2022. Capturing Failures of Large Language Models via Human Cognitive Biases. https://doi.org/10.48550/ARXIV.2202.12299

[30] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining interpretability and explainability using sensemaking theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 702–714.

[31] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI 2020*. https://www.microsoft.com/en-us/research/publication/interpreting-interpretability-understanding-data-scientists-use-of-interpretability-tools-for-machine-learning/ CHI 2020 Honorable Mention Award.

[32] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ quality & safety* 25, 10 (2016), 808–820.

[33] Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating Reinforcement Learning from Human Feedback (RLHF). *Hugging Face Blog* (2022). https://huggingface.co/blog/rlhf.

[34] Christine Lemmer-Webber. 2022. Mansplaining as a Service. https://phpc.social/@andrewfeeney/109466122845775778

[35] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[36] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[37] GLORIA MARK. 2023. *Attention span: A groundbreaking way to restore balance, happiness and productivity.* HANOVER SQUARE PRESS.

[38] Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing. *arXiv preprint arXiv:2212.10678* (2022).

[39] Christopher Mele. 2016. Pushing that crosswalk button may make you feel better, but ... https://www.nytimes.com/2016/10/28/us/placebo-buttons-elevators-crosswalks.html?src=twr&_r=1&mtrref=www.dailymail.co.uk&gwh=4C4B6CDBECBC508615DD599A3BB87FD1&gwt=pay

[40] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[41] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (June 1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[42] Donald A. Norman. 2002. *The Design of Everyday Things.* Basic Books, Inc., USA.

[43] OpenAI. 2022. GPT-3 powers the next generation of apps. https://openai.com/blog/gpt-3-apps/

[44] Sundar Pichai. 2023. An important next step on our ai journey. https://blog.google/technology/ai/bard-google-ai-search-updates/

[45] Christopher R. 2022. How the ritual of pointing and calling - Shisa Kanko - embeds us in the world. https://www.christopherroosen.com/blog/2020/4/20/how-the-ritual-of-pointing-and-calling-shisa-kanko-embeds-us-in-the-world

[46] Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. 2022. Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models. https://doi.org/10.48550/ARXIV.2206.08325

[47] Kevin Roose. 2023. A conversation with Bing's chatbot left me deeply unsettled. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[48] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[49] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM (CACM)* 63, 12 (2020), 54–63.

[50] Jacob Steinhardt. 2023. Aligning ML Systems with Human Intent. (2023). https://jsteinhardt.stat.berkeley.edu/talks/satml/tutorial.html SaTML 2023.

[51] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.

[52] Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*. 1–7.

[53] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. *arXiv preprint arXiv:2301.13848* (2023).