AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

2024, Vol. 11, No. 4, 493–510
https://doi.org/10.1037/dec0000250

Boosting Human Competences With Interpretable and Explainable Artificial Intelligence

Stefan M. Herzog[1] and Matija Franklin[2]

[1] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[2] Causal Cognition Lab, Division of Psychology and Language Sciences, University College London

Artificial intelligence (AI) is becoming integral to many areas of life, yet many—if not most—AI systems are opaque black boxes. This lack of transparency is a major source of concern, especially in high-stakes settings (e.g., medicine or criminal justice). The field of explainable AI (XAI) addresses this issue by explaining the decisions of opaque AI systems. However, such post hoc explanations are troubling because they cannot be faithful to what the original model computes—otherwise, there would be no need to use that black box model. A promising alternative is simple, inherently interpretable models (e.g., simple decision trees), which can match the performance of opaque AI systems. Because interpretable models represent—by design—faithful explanations of themselves, they empower informed decisions about whether to trust them. We connect research on XAI and inherently interpretable AI with that on behavioral science and boosts for competences. This perspective suggests that both interpretable AI and XAI could boost people's competences to critically evaluate AI systems and their ability to make accurate judgments (e.g., medical diagnoses) in the absence of any AI support. Furthermore, we propose how to empirically assess whether and how AI support fosters such competences. Our theoretical analysis suggests that interpretable AI models are particularly promising and—because of XAI's drawbacks—preferable. Finally, we argue that explaining large language models (LLMs) faces similar challenges as XAI for supervised machine learning and that the gist of our conjectures also holds for LLMs.

Keywords: explainable artificial intelligence, interpretable artificial intelligence, boosting, competences, knowledge

Artificial intelligence (AI) is becoming integral to many areas of life, including health care (e.g., predictive diagnostics), management (e.g., personnel selection), and entertainment (e.g., recommender systems). AI refers to the simulation of human intelligence in machines designed to perform tasks that typically require human intelligence (M. Mitchell, 2020; Russell & Norvig, 2021). Recently, generative AI in the form of large language models (LLMs) has gained widespread attention because of chatbots (e.g., ChatGPT) and other applications built on LLMs that are applicable to a wide range of use cases in daily life (Liao & Wortman Vaughan, 2024). However, many—if not most—AI systems are opaque black boxes (e.g., random forests, deep-learning neural networks, or transformer-based LLMs). That is, it is not clear exactly how and why they arrive at their outputs (Rudin, 2019; Rudin et al., 2022; Zhao et al., 2024). This is especially troubling for

Stefan M. Herzog https://orcid.org/0000-0003-2329-6433

Matija Franklin https://orcid.org/0000-0003-1846-8907

The authors have no competing interests to declare. The authors thank Mirjam Jenny for comments on an earlier version of the article and Deb Ain for editing the article.

Stefan M. Herzog contributed to conceptualization, project administration, visualization, writing–original draft, and writing–review and editing. Matija Franklin contributed to conceptualization, project administration, writing–original draft, and writing–review and editing (Holcombe et al., 2020).

Correspondence concerning this article should be addressed to Stefan M. Herzog, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzealle 94, 14195 Berlin, Germany. Email: herzog@mpib-berlin.mpg.de

high-stakes domains (e.g., medicine or criminal justice) because, for example, AI models trained on data that reflects social biases will discriminate (Gallegos et al., 2024; Wachter et al., 2020; Yeom et al., 2018) against, say, minorities (see Obermeyer et al., 2019, for an example in medicine). Without knowing how an AI system decides, it is unclear whether it should be trusted to advise human decision-makers or even make decisions on its own.

In this article, we ask how interpretable AI and explainable AI (XAI) could boost two human competences: critically evaluating AI systems and people's ability to make accurate judgments (e.g., medical diagnoses) in the absence of any AI support. To this end, we first introduce the distinction between XAI (Dwivedi et al., 2023; Miller, 2019; Speith, 2022; Wu et al., 2024; Zhao et al., 2024) and inherently interpretable AI (Rudin, 2019; Rudin et al., 2022) and review some key concepts in XAI. We then connect research on XAI and inherently interpretable AI with research on behavioral science and boosts for competences in decision making and beyond (Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press). Last, we lay out our perspective on how interpretable AI and XAI could boost human competences and how to empirically assess whether and how AI support fosters such competences.

We focus on two qualitatively different forms of competences. The first is what we refer to as people's *critical AI competences*: their ability to critically evaluate AI systems (Almatrafi et al., 2024; Pinski & Benlian, 2024), illustrated by the two challenges of detecting the use of objectionable features (i.e., assessing whether a model is biased, e.g., against marginalized social groups; Obermeyer et al., 2019) and detecting flawed reasoning—that is, whether a model learned nonsensical relationships (e.g., by capitalizing on spurious correlations; Banerjee et al., 2023; Bihani & Rayz, 2024; Geirhos et al., 2020; Sun et al., 2024). The second is what we refer to as people's *task competences*: people's ability to make accurate judgments and predictions based on fallible cues (Harvey, 2012; Karelaia & Hogarth, 2008; e.g., predicting a patient's future health status given their medical test results) in the absence of any AI support.

Our discussion focuses on supervised machine learning—that is, learning models using data sets with a labeled ground truth or gold standard, such as predicting the future health status of a patient. We close by arguing that explaining LLMs (Zhao et al., 2024) faces similar challenges as XAI for supervised machine learning and that the gist of our conjectures also holds for LLMs.

XAI Versus Inherently Interpretable AI

The emerging field of XAI is focused on developing explainability methods that enhance the understandability of opaque AI systems (for reviews and taxonomies of XAI approaches in supervised machine learning and LLMs, see, e.g., Dwivedi et al., 2023; Liao & Wortman Vaughan, 2024; Miller, 2019; Speith, 2022; Wu et al., 2024; Zhao et al., 2024); see also Rudin (2019) for a critical discussion of the concept of "explanation" in XAI. However, such post hoc methods have severe inherent limitations (Rudin, 2019; Rudin et al., 2022; Turpin et al., 2023; Ye & Durrett, 2022), since they cannot, in principle, faithfully represent the original, opaque black box model's decision-making process. As Rudin (2019) noted:

> Explainable ML [machine learning] methods provide explanations that are not faithful to what the original model computes. Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable.) This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space.
>
> An inaccurate (low-fidelity) explanation model limits trust in the explanation, and by extension, trust in the black box that it is trying to explain. An explainable model that has a 90% agreement with the original model indeed explains the original model most of the time. However, an explanation model that is correct 90% of the time is wrong 10% of the time. If a tenth of the explanations are incorrect, one cannot trust the explanations, and thus one cannot trust the original black box. If we cannot know for certain whether our explanation is correct, we cannot know whether to trust either the explanation or the original model. (p. 207)

For supervised machine learning tasks, a promising alternative to XAI is simple, inherently interpretable models (e.g., simple decision trees or checklists), which can match the performance of opaque AI systems (Katsikopoulos et al., 2020; Rudin, 2019; Rudin et al., 2022). An interpretable

model can be both accurate and understandable because it "obeys a domain-specific set of constraints [that] allow it (or its predictions, or the data) to be more easily understood by humans" (Rudin et al., 2022, pp. 3–4). Interpretable models represent—by design—faithful explanations of themselves. Because there is no uncertainty about the faithfulness of their explanation, interpretable models empower informed decisions about whether to trust them (Rudin et al., 2022). Unfortunately, inherent interpretability may not be an option for LLMs (and other forms of generative AI). Given LLMs' immense, inherent complexity (Liao & Wortman Vaughan, 2024; Wu et al., 2024; Zhao et al., 2024), inherently interpretable LLMs would be an oxymoron.

Importantly, transparency is not sufficient for interpretability. For example, a large, deep decision tree with many decision nodes or a regression model with many predictors are both transparent models in that every component can in principle be inspected, but the typical human will lack the cognitive bandwidth to grasp the models in their entirety. They will therefore be unable to simulate the outputs of the model given the inputs of a case (Lipton, 2018) unless they tediously note down the interim steps or calculations using external aids such as pen and paper or a computer, which defeats the point of using an interpretable model. One empirical approach for assessing the interpretability of a model is to check the extent to which a human can simulate the outputs of an interpretable model when given the case inputs (Poursabzi-Sangdeh et al., 2021; e.g., patient information), a concept known as *simulatability* (Lipton, 2018).

There is a common—mostly unquestioned—assumption that a trade-off must be made between an AI model's accuracy and its interpretability. However, some authors argue that with enough time and expertise, it is possible for many supervised tasks to design an interpretable model that performs well because in many domains, qualitatively different models perform similarly well, so it is likely that some of these models are simple and inherently interpretable (Rudin, 2019; Rudin et al., 2022; Semenova et al., 2022). These authors further argue that the presumed trade-off between accuracy and interpretability is the result of allocating a considerable amount of effort to devising ever more complicated opaque models or post hoc explainability methods instead of investing effort into using domain knowledge to make it possible to create accurate inherently interpretable models. This point resonates with research on the importance of human expertise in designing effective and robust algorithms and AI systems more generally (Dawes, 1979; Gigerenzer, 2023; Mosqueira-Rey et al., 2023; Raisch & Fomina, 2024; Ribeiro et al., 2016; Simchon & Gilead, 2024; Wang et al., 2018). For example, humans' domain knowledge can narrow down the space of possible data, examples, features, or models to consider (e.g., Bengio et al., 2009; Gennatas et al., 2020; Simchon & Gilead, 2024). Consider Gennatas et al. (2020), who showed that a rule-based machine learning model for clinical predictions that derived its decision rules from empirical data could be trained with less data and performed better on new, unseen data when only the rules that both the data and the domain experts agreed on were included. Incorporating domain knowledge during model development can make models both simpler and more robust.

Explainability Methods

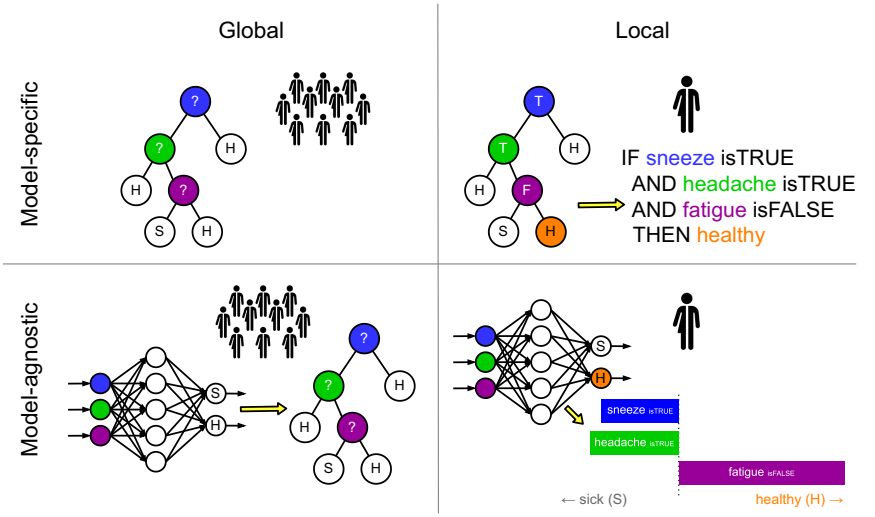
There are many explainability methods and many ways to categorize them into different types of approaches (see, e.g., Speith, 2022; Zhao et al., 2024). Here we highlight two conceptual distinctions along which to categorize explainability methods that are often mentioned in the literature: the scope of an explanation (i.e., global explanation of the whole model vs. local explanation for an individual data point) and whether the explainability method is model-specific or model-agnostic (see Figure 1). Although interpretable AI is not a subset of XAI (see Rudin et al., 2022, for a discussion on how interpretable AI predates XAI by decades), we argue that it is nevertheless useful to consider how interpretable models fit into frameworks from XAI.

Scope of Explanations: Global Versus Local

Explanations can be global or local (Speith, 2022; Zhao et al., 2024). *Global explanations* aim to offer a holistic, wide-angle view of the model's overall decision-making behavior (e.g., how a population of patients is diagnosed by an AI model; see Figure 1, left column). In contrast, *local explanations* aim to explain the predictions made by AI models for a specific, individual data point (e.g., an individual patient diagnosed with an AI model; see Figure 1, right column).

Figure 1

Illustration of the Scope and Model Agnosticity of Explainable Artificial Intelligence Methods Using a Toy Scenario of Diagnosing a Hypothetical Medical Condition Based on Three Symptoms (Sneezing, Headache, and Fatigue)



Note. *Global, model-specific explanation* (top left): An inherently interpretable decision tree that is simple enough to be its own explanation. Users can inspect the decision nodes (i.e., the exact features used) and can simulate (i.e., predict) how the tree will categorize any new patient. *Local, model-specific explanation* (top right): A simple, inherently interpretable decision tree applied to a particular patient. The Boolean decision rule (on the right) shows the features used for the categorization and thus explains the path the patient took through the decision tree (indicated by colored circles). In this example, the local explanation is faithful by design. In other cases, however, model-specific, local explanations can also be post hoc and thus not faithful by design. *Global, model-agnostic explanation* (bottom left): Categorizations of an opaque neural network are approximated using an inherently interpretable model (here, a simple decision tree) as a surrogate model. This approach is model-agnostic, as it could be applied to other classification models (e.g., an ensemble of decision trees). *Local, model-agnostic explanation* (bottom right): Categorizations of an opaque neural network for a particular patient are explained using a feature attribution method (e.g., local interpretable model-agnostic explanations or SHapley Additive exPlanations) that can be applied to other classification models. "T" = true (i.e., indicating that the respective feature is present in a patient). Adapted from Figure 1 in "Interpretability of Machine Learning-Based Prediction Models in Healthcare," by G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, 2020, *WIREs Data Mining and Knowledge Discovery*, *10*(5), p. 7 (https://doi.org/10.1002/widm.1379). Copyright 2020 by Wiley Periodicals LLC. Adapted with permission. Unisex icon made by Sudowoodo from https://www.flaticon.com. See the online article for the color version of this figure.

Model-Specific Versus Model-Agnostic Explainability Methods

Explainability methods can be *model-specific* or *model-agnostic* (Speith, 2022). Model-specific explainability methods leverage the architecture of an AI method to derive explanations. For example, there are explainability methods for the random forest algorithm (Biau & Scornet, 2016; Breiman, 2001) that capitalize on the fact that this algorithm is an ensemble of many decision trees (e.g., model-specific variable importance metrics or methods to derive a surrogate model in the form of short lists of simple rules; Fokkema & Strobl, 2020; Strobl et al., 2008). As another example, very simple decision trees (Katsikopoulos et al., 2020; Z. J. Wang et al., 2022; e.g., see Figure 1, upper left panel) are inherently interpretable and thus self-explanatory by design. This property is, however, specific to the architecture of a simple decision tree and cannot be applied to other AI models.

Model-agnostic explainability methods, in contrast, provide explanations for AI predictions that are not dependent on the model's internal architecture, making them applicable to a variety

of models. They examine the input–output relationship to determine how different features influence predictions without requiring knowledge of the model's internal workings. Prominent examples of model-agnostic methods include local interpretable model-agnostic explanations (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP; Lundberg et al., 2020).

In the next two subsections, we review two specific types of explainability methods: feature attribution methods and counterfactual XAI.

Global and Local Feature Attribution Methods

Feature attribution methods are techniques used to determine the contribution of each input variable to the output of a predictive model. One example of a feature attribution method is SHAP, which assigns an importance value to each feature based on its contribution to the prediction (Bordt & von Luxburg, 2023). For instance, SHAP may be used to identify which clinical variables most significantly influence a model's prediction about whether or not a patient has a particular disease.

Global feature attribution methods offer a comprehensive and aggregated view of the importance of features across an entire population of cases. For example, global SHAP values assess the average impact of each feature across a model's predictions across all cases (Bordt & von Luxburg, 2023; Lundberg & Lee, 2017). In contrast, local feature attribution methods pinpoint the specific elements that significantly influence a model's prediction for a particular case. For example, saliency maps express the presumed importance of pixels in an image in determining a classification (Müller, 2024; e.g., cancer detection on X-rays). By design, inherently interpretable models (e.g., simple decision trees) represent their own faithful explanation at both the global and local levels (see Figure 1, top left and top right panels).

Counterfactual Explanations

Counterfactual explanations (Cheng et al., 2024; Guidotti, 2024; Wachter et al., 2018) explain an AI model's predictions by finding the smallest input changes that would lead to a different prediction (i.e., the "closest possible world"; Wachter et al., 2018, p. 845). Consider the following counterfactual explanation:

> You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan. (Wachter et al., 2018, p. 844)

Here a stated decision is explained by a counterfactual statement clarifying how the world would have needed to be different for the opposite decision to occur.

Boosting Human Competences With Interpretable AI and XAI

There is a vast literature on AI- and algorithm-assisted hybrid decision making—spread across fields such as human–computer interaction, management, judgment and decision making, and psychology—investigating when and why people are willing or reluctant to use algorithms (e.g., Burton et al., 2020; Dawes et al., 1989; Glikson & Woolley, 2020) and how to harness the complementarity between humans and AI systems (e.g., Lai et al., 2023; Nauta et al., 2023; Pescetelli, 2021; Raisch & Fomina, 2024; Steyvers & Kumar, 2023; Zöller et al., 2024). Here we argue that there is a fruitful but underexplored connection between interpretable AI and XAI and research on behavioral science and boosting people's competences in decision making and beyond (see also Franklin, 2022). Before unpacking this connection, we turn to the concept of boosting human competences (Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press).

Boosting Human Competences

Boosts are interventions informed by behavioral science that leverage human cognition or the environment to help people improve their existing competences or develop new ones (Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press), such as understanding statistical information (e.g., what a positive medical diagnostic test actually means; Gigerenzer et al., 2007) or successfully dealing with the challenges of online environments (Kozyreva et al., 2020, 2024). For example, boosts can help people learn to use techniques employed by fact-checkers to detect misinformation and assess the trustworthiness of online sources (e.g., lateral reading, i.e., searching online to find what others say about the trustworthiness of a source; McGrew, 2024; e.g., detecting whether an organization is merely a front for, say, a fossil fuel lobby organization) or to detect personality-based

microtargeting using simple self-awareness interventions (Lorenz-Spreen et al., 2021).

Boosts can have temporary or lasting effects (Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press). Short-term boosts foster a competence that is bound to a particular setting, such as a boost that helps people interpret medical tests by presenting the diagnostic information—prevalence, sensitivity, and specificity—in a format that makes the correct answer intuitively clear (i.e., natural frequencies, which already incorporate the prevalence and thus simplify the calculation of the Bayesian posterior probability; McDowell & Jacobs, 2017). Once a short-term boost is removed (e.g., diagnostic information is presented in a less intuitive format, i.e., again just as prevalence, sensitivity, and specificity), people's level of competence will likely return to baseline.

Long-term boosts, in contrast, aim at producing long-lasting competences that persist even after the boost has ended or been removed. For example, a long-term boost may teach people how to convert information expressed in a nonintuitive format into a more intuitive format by themselves. For instance, a training in how to convert, say, the diagnostic reliability of a medical test—expressed as prevalence, sensitivity, and specificity—into natural frequencies increased people's performance even weeks after the original training intervention (Sedlmeier & Gigerenzer, 2001). Long-term boosting thus strongly differs from the nudging approach (Thaler & Sunstein, 2021), where changes to the choice architecture ("nudges," e.g., changing the default in a privacy form to opt-out or opt-in) aim to directly influence behavior without targeting competences and are thus unlikely to affect people's decisions whenever the nudge is not or no longer present (Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press).

In the next two sections, we raise conjectures about when and how human competences can be boosted using interpretable AI and XAI, drawing on the concepts introduced so far. We argue that there is a conceptual parallel to be exploited between short- and long-term boosts and the effects of explainability methods and interpretable AI on short- and long-term changes in the competences and performance of people using AI models. First, we focus on people's critical AI competences—that is, their ability to critically evaluate AI systems (Almatrafi et al., 2024; Pinski & Benlian, 2024). Second, we ask when and how interpretable AI and XAI can boost people's task competences—that is, how well people can make judgments and predictions based on fallible cues (Harvey, 2012; Karelaia & Hogarth, 2008; e.g., predicting a patient's future health status given their medical test results) in the absence of AI support.

Critical AI Competences

AI models pose several challenges for individuals and society (see, e.g., Gallegos et al., 2024; Ji et al., 2023; Liao & Wortman Vaughan, 2024; Mehrabi et al., 2021; S. Mitchell et al., 2021; van Giffen et al., 2022, for reviews on errors, biases, transparency, and issues of fairness in supervised machine learning and LLMs). Being able to critically evaluate AI systems is thus a crucial part of AI literacy (Almatrafi et al., 2024; Pinski & Benlian, 2024). Here we present a theoretical analysis on how inherently interpretable models and three classes of XAI—global feature attribution methods, local feature attribution methods, and counterfactual explanations—might boost critical AI competences. To do so, we use the two AI challenges of detecting the use of objectionable features (i.e., assessing whether a model is biased) and detecting flawed reasoning (i.e., whether a model learned nonsensical relationships).

Detecting the Use of Objectionable Features

Algorithmic decision making can produce unfair and discriminating decisions and can perpetuate racial and socioeconomic inequalities because of biases in AI systems (Bai et al., 2024; Gallegos et al., 2024; S. Mitchell et al., 2021; Wachter et al., 2020). A key critical AI competence is therefore to assess whether a model is biased, for example, against marginalized social groups (e.g., racialized groups; Obermeyer et al., 2019). In the simplest case, this means being able to ascertain whether a model uses objectionable features (e.g., protected attributes including ethnicity or gender) or proxies thereof (Hertwig et al., 2023; Yeom et al., 2018).

Both inherently interpretable models and the three classes of XAI we focus on—global feature attribution methods, local feature attribution methods, and counterfactual explanations—provide sufficient information to raise a red

flag. If objectionable features or proxies for them appear in a model or an explanation, there is reason to suspect that the model might be biased, and further scrutiny is warranted. Even local XAI methods (e.g., local feature attribution methods or counterfactual explanations), which present information on just one case, provide sufficient information. Seeing an objectionable feature (or a proxy thereof) used in a single case is enough to know that it is used in the model as a whole and thus that the entire model is suspect. However, interpretable models and models explained using global feature attribution methods provide a stronger signal to assess the risk of bias because they summarize how a model uses features across a whole population of cases and thus represent an aggregated and less noisy signal.

Detecting Flawed Reasoning

Another important challenge posed by AI is that AI models lack common sense (M. Mitchell, 2020; Williams & Huckle, 2024) and can learn nonsensical relationships, for example, by capitalizing on spurious correlations ("shortcut learning"; Banerjee et al., 2023; Bihani & Rayz, 2024; Geirhos et al., 2020; Sun et al., 2024) or neglecting the setting's causal texture (e.g., Subbaswamy & Saria, 2020). Such models might perform well in settings similar to the one on which they were trained but may not generalize well to other settings. Subbaswamy and Saria (2020) offered the following example of flawed logic:

> Consider the mortality risk prediction model trained by Caruana and others (2015) on a dataset of hospitalized pneumonia patients, using information such as lab measurements, vital signs, and comorbidities. While the model had high predictive accuracy on one dataset, it was unstable to shifts in the choices driving which patients get admitted to the ICU [(intensive care unit)] versus the floor. As a result, when they evaluated it for triaging pneumonia patients upon ED [(emergency department)] presentation, they found that their model incorrectly predicted lower risk for patients with pneumonia and asthma versus those with only pneumonia. This shift in ICU admission policy, while subtle, had big implications: had the model been deployed for triage, it would have greatly endangered asthmatic pneumonia patients by suggesting they should be sent home. (p. 346)

It was an interpretable model that led Caruana et al. (2015) to detect the problem with the data set and naïvely fitting supervised machine learning models:

> One of the methods being evaluated was rule-based learning. Although models based on rules were not as accurate as the neural net models, they were intelligible, i.e., interpretable by humans. On one of the pneumonia datasets, the rule-based system learned the rule "HasAsthama(x) [sic] ⇒ LowerRisk(x)," i.e., that patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia than the general population. Needless to say, this rule is counterintuitive. But it reflected a true pattern in the training data: patients with a history of asthma who presented with pneumonia usually were admitted not only to the hospital but directly to the ICU (Intensive Care Unit). The good news is that the aggressive care received by asthmatic pneumonia patients was so effective that it lowered their risk of dying from pneumonia compared to the general population. The bad news is that because the prognosis for these patients is better than average, models trained on the data incorrectly learn that asthma lowers risk, when in fact asthmatics have much higher risk (if not hospitalized). (p. 1721)

Detecting flawed reasoning in an AI system is a key critical AI competence that opens up opportunities to improve a model based on humans' knowledge about a domain (Gennatas et al., 2020; Ribeiro et al., 2016; Simchon & Gilead, 2024; Subbaswamy & Saria, 2020; e.g., medicine). Based on domain knowledge, the feature in question (asthma) was a reasonable feature to use, but it was used in a nonsensical way (Caruana et al., 2015). In other settings, a model might use a feature that is clearly nonsensical or at least suspicious (e.g., patients' handedness) or a feature that is otherwise problematic (e.g., one that is trivially related to the prediction target and would not be available in practice; Kapoor & Narayanan, 2023). Thus, there is an asymmetry: Knowing that a model uses a nonsensical feature—even in just one case—is sufficient to cast doubt on the reasonableness of the whole model, whereas the use of a sensical feature is not enough to confirm a model's reasonableness; an analyst also needs to know how the feature impacts the decision.

Interpretable models such as the one used by Caruana et al. (2015) enable analysts to detect flawed reasoning in AI by allowing for a straightforward comparison of the workings of the model with experts' domain knowledge (Gennatas et al., 2020; Rudin et al., 2022; Z. J. Wang et al., 2022). In contrast, the extent to which XAI methods make such comparisons possible depends on whether they clarify how the feature impacts the decision. For example, counterfactual explanations and some feature attribution methods (e.g., SHAP dependency plots) do clarify this

relationship, while other forms of XAI highlight the importance of a feature but not how it influences the decision (e.g., undirected feature importance measures, including mean absolute SHAP values). Relatedly, global explanations and interpretable models (which represent their own global and local explanations) are more informative than local explanations for individual cases. Because global explanations and interpretable models summarize the influence of a feature across the whole population of cases, they allow for a more reliable assessment of possible problems.

In sum, interpretable models enable people to address challenges in AI such as the use of objectionable features and flawed reasoning. The extent to which XAI does so varies depending on the specifics of the explainability methods. Generally speaking, XAI methods that are global and clarify the directionality of features seem most promising to address the two AI challenges discussed here.

The way in which interpretable AI and some forms of XAI might foster critical AI competences can be likened to short-term boosts (Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press): They foster a competence that is bound to a particular setting in which a human decision-maker has access to an interpretable model or a post hoc explanation of an opaque model (XAI). It is unclear what long-lasting critical AI competences those AI decision aids might instill in decision-makers that would still be at their disposal once the decision aids were no longer available.

Task Competences

In this section, we examine when and how interpretable AI and XAI can boost people's task competences—that is, how well people can make judgments and predictions based on fallible cues (Harvey, 2012; Karelaia & Hogarth, 2008; Stewart, 2001) in the absence of AI support. There is evidence that AI and XAI decision aids improve people's task performance (Schemmer et al., 2022). But why should interpretable AI and certain forms of XAI boost people's competences to make accurate judgments in the long term, that is, even once AI support has been removed?

This conjecture builds on two streams of research. One is research on AI-assisted decision making and problem solving, which recently started investigating how AI support can improve human learning, for example, in visual (e.g., Alipour et al., 2021; Goyal et al., 2019; Schemmer, Bartos, et al., 2023) or planning tasks (Karny et al., 2024). The other is decades-old research on multiple-cue learning (Harvey, 2012), which studies how and how well people make judgments and predictions about a target variable (e.g., a patient's future health outcome) based on multiple fallible cues (e.g., medical tests). Multiple-cue learning research has compared various interventions aimed at improving people's performance, including outcome feedback (i.e., telling people the correct answer after each judgment), cognitive information (i.e., telling people how they are using the cues in their judgments), and task information (i.e., giving people information about the task environment; Balzer et al., 1989; Karelaia & Hogarth, 2008). For example, Balzer et al. (1992) asked participants to predict a baseball team's number of wins in a season based on five statistics per team (earned run average, batting average, double plays, stolen bases, and errors). In the task-information condition, 1 week before the main study session, participants received information about the prediction task, including visual information on how predictable the criterion was by each of the statistics, how to best weigh those statistics (i.e., in which direction and by how much), and the actual average number of games won per team. This study, along with other multiple-cue learning studies (Harvey, 2012; Karelaia & Hogarth, 2008), found that providing task information (i.e., informing the learner about the key relations between cues and outcomes in the task environment) improved performance and was superior to all other types of interventions tested.

There is a clear conceptual overlap between the kinds of information conveyed in task-information interventions in multiple-cue learning studies (Balzer et al., 1989; Harvey, 2012) and the kinds of information provided by interpretable models and certain forms of XAI (Speith, 2022; i.e., counterfactual explanations and global explanations of how features relate to the model's predictions). For example, the study by Balzer et al. (1992) gave participants summaries of the predictive relationships between the cues and the criterion provided (e.g., how to best weigh cues), which is similar to the way interpretable models and certain forms of XAI methods show how to use various features to mimic an AI model's decisions (see also the concept of simulatability

in XAI; Lipton, 2018; Poursabzi-Sangdeh et al., 2021). The key difference here is that while multiple-cue learning studies typically summarize the actual relationship between the cues and the criterion (Balzer et al., 1989; Harvey, 2012; Karelaia & Hogarth, 2008), interpretable models and XAI methods summarize the relationship between the cues and the AI model's judgments.

Assuming that AI decision aids perform reasonably well at their task, they or their explanation de facto convey summarized information about the task environment in much the same way that task-information interventions in multiple-cue learning paradigms do. Consequently, it is reasonable to expect that AI decision aids would improve people's task competences in making judgments and predictions for the same reasons that task-information interventions do in multiple-cue learning tasks (Balzer et al., 1989; Harvey, 2012). Crucially, those task competences should be long-lasting, that is, people should continue to make better judgments and predictions (compared to before they were exposed to an AI decision aid) even once those decision aids are removed (analogous to how task information can improve judgment performance a week later; Balzer et al., 1992).

There are, however, several caveats to consider when using AI decision aids to boost task competences (see also Simkute et al., 2022). First, whenever AI models are problematic (e.g., are biased or learn nonsensical relationships; see the Critical AI Competences section), those problematic tendencies might be instilled into the human decision-makers as well. Second, decision-makers may misinterpret what they can learn from AI models. For example, counterfactual explanations (Celar & Byrne, 2023; Warren et al., 2023) can make people misinterpret correlational associations as causal effects (Tešić & Hahn, 2022). Because there is no guarantee that counterfactuals align with the actual causal relationships between features and the target outcome, care must be taken that counterfactual explanations do not invite unjustified inferences and induce people to learn brittle relationships that will not generalize well. Last, people may be—for better or worse—reluctant to adopt insights from AI models if those insights disagree with people's assumptions about the task environment (see also Chen et al., 2023). If people's assumptions are off and the AI model is sensible, there is a missed opportunity to boost people's task competences. However, if the model is off, people's reluctance to incorporate its insights protects them against the sabotaging effect of subpar AI models—highlighting again the complementarity of human and AI insights (see Dawes, 1979; Gennatas et al., 2020; Raisch & Fomina, 2024; Simchon & Gilead, 2024; Steyvers et al., 2022; Zöller et al., 2024).

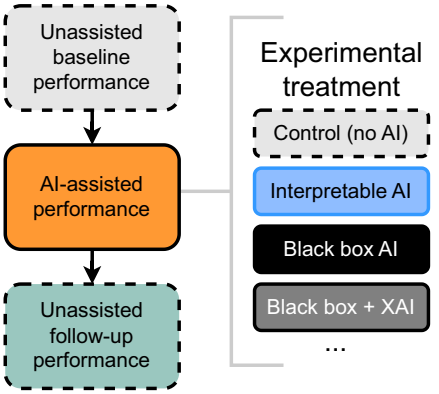
An open question is how well different forms of AI support (interpretable AI, counterfactual explanations, global explanations of how features relate to what the model predicts, or other forms of XAI) would boost human task competences. We conjecture that this will depend on the factors influencing both when and why task-information interventions improve performance in multiple-cue learning tasks (Balzer et al., 1989; Karelaia & Hogarth, 2008) and when and why decision support systems lead to automation bias and deskilling (Karny et al., 2024; Schemmer et al., 2021), as well as on the trade-off—if any—between a model's interpretability and its performance (Semenova et al., 2022). For example, effective interpretable models in a task domain are likely promising tools to boost people's task competences because they use a small number of features and thus do not overwhelm people with information (see also Stewart, 2001; Steyvers & Kumar, 2023).

Empirical Framework for Studying How Interpretable AI and XAI Affect Competences

In the previous section, we presented two classes of conjectures. First, with respect to critical AI competences, we conjectured that interpretable models and those XAI systems that are global and clarify the directionality of features enable people to better critically evaluate AI systems (e.g., detecting the use of objectionable features and flawed reasoning) compared to other XAI systems (e.g., local explanations or nondirectional feature attribution methods) or opaque AI systems. Second, with respect to task competences, we conjectured that accurate interpretable models and accurate XAI systems that are global and clarify the directionality of features improve people's task performance more than other XAI systems (e.g., local explanations or nondirectional feature attribution methods) or opaque AI systems once the AI support is no longer available.

How could conjectures such as those be empirically tested? Figure 2 shows our proposed

Figure 2
Empirical Paradigm for Studying the Short-Term and Long-Term Boosting Effects of Interpretable AI and XAI on Human Competences



Note. See also Franklin and Lagnado (2022). AI = artificial intelligence; XAI = explainable AI. See the online article for the color version of this figure.

experimental paradigm (see also Franklin, 2022; Franklin & Lagnado, 2022). At its core, the paradigm uses the commonly employed between-participants comparisons in research on AI-assisted decision making, comparing the task performance (or other variables of interest, e.g., trust, understanding, or usability) of different AI support conditions (e.g., participants with XAI support vs. opaque AI support vs. no AI support; Lai et al., 2023; Nauta et al., 2023; Rong et al., 2024; Schemmer, Bartos, et al., 2023). The proposed paradigm naturally makes it possible to study our and other conjectures about interpretable AI and XAI as short-term boosts for critical AI competences (e.g., to detect AI bias or flawed reasoning) by contrasting suitable outcome variables beyond task performance (see Rong et al., 2024, for a review of measures used in XAI research).

However, there are three aspects in our proposed research framework that are less commonly used in empirical research on AI-assisted decision making (Lai et al., 2023; Nauta et al., 2023; Rong et al., 2024) and are crucial for effectively studying the short-term and long-term boosting effects of interpretable AI and XAI on human competences. First, and most importantly, after the main, intervention part of a study, participants' unassisted follow-up task performance is assessed (and ideally again after several days, weeks, or months). This makes it possible to study when and how interpretable AI and XAI can boost task competences in the absence of AI support (see also Karny et al., 2024). To what extent do the effects of support by interpretable AI and XAI persist even when the support is removed? If the effect persists for a particular intervention, the AI support used in that intervention boosted competences. In contrast, if the behavior (largely) returns to baseline, the AI support did not foster long-lasting competences (see also Hertwig & Grüne-Yanoff, 2017; Herzog & Hertwig, in press).

Second, whenever feasible, we encourage the use of mixed designs (i.e., studying the same participants in different conditions) and the corresponding mixed-level statistical analyses, which efficiently account for heterogeneity in participants and items (Baayen et al., 2008; Judd et al., 2012; Veenman et al., 2024). Foremost, we propose establishing participants' baseline task performance in an unassisted condition prior to the main, intervention part of a study. This yields greater statistical power in detecting intervention effects in the main task and in the unassisted follow-up task (when analyzing performance relative to the unassisted baseline performance). Furthermore, particularly when the study population is difficult to recruit (e.g., domain experts), we suggest manipulating the different AI support conditions within participants (see also Rong et al., 2024).

Last, we recommend using cognitive-modeling approaches (Farrell & Lewandowsky, 2018) to better understand how AI support influences humans' decisions and the accuracy of the final decisions made by a joint human–AI system (Steyvers & Kumar, 2023), as well how AI support shapes the development of long-lasting competences. For example, there is a debate in research on AI-assisted decision making on how to measure appropriate reliance on AI support (Fok & Weld, 2024; Guo et al., 2024; Schemmer, Kuehl, et al., 2023; Schoeffer et al., 2023). We conjecture that framing AI support as a signal detection task (see Hautus et al., 2021; Langer et al., 2023; Sheridan, 2019) could bring clarity to some of the issues discussed. Signal detection tasks make it possible to break down participants' decisions into their ability to, for example, distinguish between correct and incorrect AI advice, as well as their tendency to adopt AI

advice more or less leniently. This then allows for an examination of how different AI support settings can influence people's detection ability and/or advice-taking tendency.

What About LLMs?

In the sections above, we proposed two classes of conjectures about how interpretable models and certain forms of XAI systems boost people's competences to critically evaluate AI systems and make accurate judgments in the absence of AI support. When developing those conjectures, we focused on "classic" supervised machine learning (e.g., triaging patients in an emergency department). But to what extent do those conjectures also hold for LLMs? Here we briefly discuss three key ideas.

First, although classic supervised machine learning and generative AI in the form of LLMs are qualitatively different, LLMs, as decision aids or autonomous decision agents face, in many ways, the same—and also additional, unique (Liao & Wortman Vaughan, 2024)—fundamental problems. In particular, LLMs also make errors, for example, because they hallucinate (Ji et al., 2023) or learn spurious correlations that do not generalize well (Bihani & Rayz, 2024; Sun et al., 2024). LLMs can also be biased and unfair (Bai et al., 2024; Echterhoff et al., 2024; Gallegos et al., 2024). Thus, there is a clear need to understand why LLMs output what they do. But given that post hoc explanations of LLMs' outputs are not faithful (Chuang et al., 2024; Turpin et al., 2023; Ye & Durrett, 2022), these explanations cannot be trusted for the same reasons as those discussed in the context of classic supervised machine learning (Rudin, 2019; Rudin et al., 2022). We conjectured that inherently interpretable AI should generally be preferred over post hoc explanations of black box AI in supervised machine learning tasks because inherently interpretable AI circumvents many of the problems of post hoc explanations. Unfortunately, inherent interpretability may not be an option for LLMs. Given LLMs' immense, inherent complexity (Liao & Wortman Vaughan, 2024; Wu et al., 2024; Zhao et al., 2024), inherently interpretable LLMs would be an oxymoron. Thus, we need to assume that LLMs can only be explained post hoc, which is worrisome for broadly the same reasons (Rudin, 2019; Rudin et al., 2022) we discussed in the XAI Versus Inherently Interpretable AI section, leaving us between a rock and a hard place. Either users or policymakers accept the risks of explaining LLM outputs post hoc or they need to accept that they simply cannot—or should not—use LLMs in particular settings (e.g., high-stakes settings, such as medicine or criminal justice).

Second, let us consider situations in which LLMs that are explained post hoc are considered acceptable. We do not see any principled reasons why the conjectures derived for supervised machine learning should not also hold for generative AI in the form of LLMs. For example, global feature attribution methods for LLMs (Zhao et al., 2024) applied to an LLM processing many cases (e.g., patients) should make it easier to detect whether or not an LLM uses objectionable features (e.g., race) compared to local explanations of how the LLM judges an individual case. However, LLMs can handle many more types of tasks than the typically task-specific, classic methods of supervised machine learning, including summarizing text (Nguyen et al., 2024; D. Zhang, Yu, et al., 2024), generating text (Li et al., 2024), measuring mental representations from textual data (Bhatia & Aka, 2022; Bhatia & Walasek, 2023; Demszky et al., 2023), and using multimodal information (T. Zhang, Ladhak, et al., 2024, e.g., text and images). There is a vast space of situations where our conjectures do not apply or are simply mute, and additional or different forms of evaluation (Chang et al., 2024) and post hoc explanations for LLMs will be required (Steyvers et al., 2024; Wu et al., 2024; Zhao et al., 2024)—which might offer different opportunities and challenges for fostering or stifling human competences.

Last, given the exciting developments in LLMs, it is tempting to think that classic supervised machine learning models and the need for them to be explained will soon be obsolete because LLMs will be used everywhere instead. However, there are reasons to believe that this will not be the case. First, LLMs struggle with tabular data (van Breugel & van der Schaar, 2024), the most common form of data in most decision-making settings (e.g., rows of patients and columns with outcome and predictor variables). More generally, deep-learning neural network models have yet to match the performance of other supervised machine learning models when using tabular data (Grinsztajn et al., 2022). Second, LLMs exhibit

stronger social biases than do classic machine learning models (Liu et al., 2024), which make them generally less preferable whenever classic machine learning models are available. Third, multipurpose generative AI systems such as LLMs are computationally orders of magnitude more costly, and thus environmentally less sustainable, than task-specific systems (Luccioni et al., 2024), let alone simple, classic machine learning models. Thus, even with an increased focus on efficient training and deploying LLMs (Wan et al., 2024), there will likely be increased pressure to use, wherever possible, simple and more efficient machine learning models. Last, regulatory reasons (Goodman & Flaxman, 2017; Wachter et al., 2024) might restrict or even prevent the use of LLMs in certain domains and thus necessitate the use of simpler models that are more explicitly grounded in and validated by experts' domain knowledge.

Taken together, these reasons suggest that, at least for the medium-term future, it is likely that classic machine learning, LLM-based, and hybrid AI systems will all be deployed. There might even be ways in which LLMs could support the use of more classic models by helping explain their outputs (Kroeger et al., 2024; Zytek et al., 2024)—although also here the lack of faithfulness of LLM-generated explanations (Chuang et al., 2024; Turpin et al., 2023; Ye & Durrett, 2022) is an issue and would need to be monitored.

Conclusions

It may seem ironic that there is such intense scrutiny of the transparency and accountability of AI systems (Goodman & Flaxman, 2017; Liao & Wortman Vaughan, 2024; Wachter et al., 2024; Wachter & Mittelstadt, 2019) given that human decision-makers are the original opaque black boxes (Farrell & Lewandowsky, 2018; Hammond, 1996; Hoffman, 1960). However, given the increasing sway of AI systems worldwide, we do not believe that humanity has the luxury to entertain such whataboutisms and should welcome this increased scrutiny. Taking both human and artificial black boxes into focus, here we strive to connect research on inherently interpretable AI and XAI with research on behavioral science and boosting. We present a novel perspective on how interpretable AI and XAI could boost human competences to critically evaluate AI systems and make accurate judgments in the absence of AI support. Our theoretical analysis suggests that inherently interpretable AI models (Katsikopoulos et al., 2020; Rudin, 2019; Rudin et al., 2022) are particularly promising for boosting human competences and should thus be considered and prioritized wherever feasible.

References

Alipour, K., Ray, A., Lin, X., Cogswell, M., Schulze, J. P., Yao, Y., & Burachas, G. T. (2021). Improving users' mental model with attention-directed counterfactual edits. *Applied AI Letters*, *2*(4), Article e47. https://doi.org/10.1002/ail2.47

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open*, *6*, Article 100173. https://doi.org/10.1016/j.caeo.2024.100173

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). *Measuring implicit bias in explicitly unbiased large language models*. arXiv. https://doi.org/10.48550/arXiv.2402.04105

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*(3), 410–433. https://doi.org/10.1037/0033-2909.106.3.410

Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes*, *53*(1), 35–54. https://doi.org/10.1016/0749-5978(92)90053-A

Banerjee, I., Bhattacharjee, K., Burns, J. L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B. N., Shiradkar, R., & Gichoya, J. (2023). "Shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology*, *20*(9), 842–851. https://doi.org/10.1016/j.jacr.2023.06.025

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). *Curriculum learning*. ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, New York, NY, United States. https://doi.org/10.1145/1553374.1553380

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214. https://doi.org/10.1177/09637214211068113

Bhatia, S., & Walasek, L. (2023). Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences*, *120*(25), Article e2220726120. https://doi.org/10.1073/pnas.2220726120

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bihani, G., & Rayz, J. T. (2024). *Learning shortcuts: On the misleading promise of NLU in language models*. arXiv. https://doi.org/10.48550/arXiv.2401.09615

Bordt, S., & von Luxburg, U. (2023). *From Shapley values to generalized additive models and back*. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. Retrieved March 27, 2024, from https://proceedings.mlr.press/v206/bordt23a.html

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission*. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, United States. https://doi.org/10.1145/2783258.2788613

Celar, L., & Byrne, R. M. J. (2023). How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, *51*(7), 1481–1496. https://doi.org/10.3758/s13421-023-01407-5

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), Article 39. https://doi.org/10.1145/3641289

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human–AI decision-making with explanations. *Proceedings of the ACM on Human–Computer Interaction*, *7*(CSCW2), Article 370. https://doi.org/10.1145/3610219

Cheng, F., Zouhar, V., Chan, R. S. M., Fürst, D., Strobelt, H., & El-Assady, M. (2024). *Interactive analysis of LLMs using meaningful counterfactuals*. arXiv. https://doi.org/10.48550/arXiv.2405.00708

Chuang, Y.-N., Wang, G., Chang, C.-Y., Tang, R., Zhong, S., Yang, F., Du, M., Cai, X., & Hu, X. (2024). *FaithLM: Towards faithful explanations for large language models*. arXiv. https://doi.org/10.48550/arXiv.2402.04678

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), Article 194. https://doi.org/10.1145/3561048

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). *Cognitive bias in high-stakes decision-making with LLMs*. arXiv. https://doi.org/10.48550/arXiv.2403.00811

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

Fok, R., & Weld, D. S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*. Advance online publication. https://doi.org/10.1002/aaai.12182

Fokkema, M., & Strobl, C. (2020). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods*, *25*(5), 636–652. https://doi.org/10.1037/met0000256

Franklin, M. (2022). The influence of explainable artificial intelligence: Nudging behavior or boosting capability? *ICML-22 Workshop on Human-Machine Collaboration and Teaming held at the Thirty-ninth International Conference on Machine Learning (ICML-22)*. arXiv. https://doi.org/10.48550/arXiv.2210.02407

Franklin, M., & Lagnado, D. (2022). Human-AI interaction paradigm for evaluating explainable artificial intelligence. In C. Stephanidis, M. Antona, & S. Ntoa (Eds.), *HCI international 2022 posters* (pp. 404–411). Springer International Publishing. https://doi.org/10.1007/978-3-031-06417-3_54

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational*

Linguistics, *50*(3), 1–79. https://doi.org/10.1162/coli_a_00524

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., Interian, Y., Luna, J. M., Simone, C. B., II, Auerbach, A., Delgado, E., van der Laan, M. J., Solberg, T. D., & Valdes, G. (2020). Expert-augmented machine learning. *Proceedings of the National Academy of Sciences*, *117*(9), 4571–4577. https://doi.org/10.1073/pnas.1906831117

Gigerenzer, G. (2023). Psychological AI: Designing algorithms informed by human psychology. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231180597

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). *Counterfactual visual explanations*. Proceedings of the 36th International Conference on Machine Learning. Retrieved July 11, 2024, from https://proceedings.mlr.press/v97/goyal19a.html

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 507–520). Curran Associates, Inc. Retrieved July 9, 2024, from https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html

Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, *38*, 2770–2824. https://doi.org/10.1007/s10618-022-00831-6

Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). *A decision theoretic framework for measuring AI reliance*. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, United States. https://doi.org/10.1145/3630106.3658901

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press. https://doi.org/10.1093/oso/9780195097344.001.0001

Harvey, N. (2012). Learning judgment and decision making from feedback. In M. K. Dhami, A. Schlottmann, & M. R. Waldmann (Eds.), *Judgment and decision making as a skill: Learning, development and evolution* (pp. 199–226). Cambridge University Press.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide* (3rd ed.). Routledge. https://doi.org/10.4324/9781003203636

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Hertwig, R., Herzog, S. M., & Kozyreva, A. (2023). Blinding to circumvent human biases: Deliberate ignorance in humans, institutions, and machines. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231188052

Herzog, S. M., & Hertwig, R. (in press). Boosting: Empowering citizens with behavioral science. *Annual Review of Psychology*.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*(2), 116–131. https://doi.org/10.1037/h0047807

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLOS ONE*, *15*(12), Article e0244611. https://doi.org/10.1371/journal.pone.0244611

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Karny, S., Mayer, L. W., Ayoub, J., Song, M., Su, H., Tian, D., Moradi-Pari, E., & Steyvers, M. (2024). *Learning with AI assistance: A path to better task performance or dependence?* Proceedings of the ACM Collective Intelligence Conference, New York, NY, United States. https://doi.org/10.1145/3643562.3672610

Katsikopoulos, K. V., Şimşek, Ö., Buckmann, M., & Gigerenzer, G. (2020). *Classification in the wild: The science and art of transparent decision making*. MIT Press. https://doi.org/10.7551/mitpress/11790.001.0001

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Kroeger, N., Ley, D., Krishna, S., Agarwal, C., & Lakkaraju, H. (2024). *In-context explainers: Harnessing LLMs for explaining black box models*. arXiv. https://doi.org/10.48550/arXiv.2310.05797

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). *Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies*. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, United States. https://doi.org/10.1145/3593013.3594087

Langer, M., Baum, K., & Schlicker, N. (2023). *A signal detection perspective on error and unfairness detection as a critical aspect of human oversight of AI-based systems*. PsyArXiv. https://doi.org/10.31234/osf.io/ke256

Li, J., Tang, T., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2024). Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, *56*(9), Article 230. https://doi.org/10.1145/3649449

Liao, Q. V., & Wortman Vaughan, J. (2024). *AI transparency in the age of LLMs: A human-centered research roadmap*. Harvard Data Science Review. https://doi.org/10.1162/99608f92.8036d03b

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Liu, Y., Gautam, S., Ma, J., & Lakkaraju, H. (2024). Confronting LLMs with traditional ML: Rethinking the fairness of large language models in tabular classifications. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: Human language technologies (Volume 1: Long Papers)* (pp. 3603–3620). Association for Computational Linguistics. Retrieved July 9, 2024, from https://aclanthology.org/2024.naacl-long.198

Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people's ability to detect microtargeted advertising. *Scientific Reports*, *11*(1), Article 15541. https://doi.org/10.1038/s41598-021-94796-z

Luccioni, S., Jernite, Y., & Strubell, E. (2024). *Power hungry processing: Watts driving the cost of AI deployment?* Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, United States. https://doi.org/10.1145/3630106.3658542

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, United States. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, *143*(12), 1273–1312. https://doi.org/10.1037/bul0000126

McGrew, S. (2024). Teaching lateral reading: Interventions to help people read like fact checkers. *Current Opinion in Psychology*, *55*, Article 101737. https://doi.org/10.1016/j.copsyc.2023.101737

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), Article 115. https://doi.org/10.1145/3457607

Miller, T. (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

Mitchell, M. (2020). *Artificial intelligence: A guide for thinking humans*. Pelican Books.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*,

56(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Müller, R. (2024). *How explainable AI affects human performance: A systematic review of the behavioural consequences of saliency maps*. arXiv. https://doi.org/10.48550/arXiv.2404.16042

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, *55*(13s), Article 295. https://doi.org/10.1145/3583558

Nguyen, H., Chen, H., Pobbathi, L., & Ding, J. (2024). *A comparative study of quality evaluation methods for text summarization*. arXiv. https://doi.org/10.48550/arXiv.2407.00747

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pescetelli, N. (2021). A brief taxonomy of hybrid intelligence. *Forecasting*, *3*(3), 633–643. https://doi.org/10.3390/forecast3030039

Pinski, M., & Benlian, A. (2024). AI literacy for users—A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans*, *2*(1), Article 100062. https://doi.org/10.1016/j.chbah.2024.100062

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). *Manipulating and measuring model interpretability*. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, New York, NY, United States. https://doi.org/10.1145/3411764.3445315

Raisch, S., & Fomina, K. (2024). Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*. Advance online publication. https://doi.org/10.5465/amr.2021.0421

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, United States. https://doi.org/10.1145/2939672.2939778

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(4), 2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85. https://doi.org/10.1214/21-SS133

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Kühl, N., Liebschner, J., & Satzger, G. (2023). *Towards effective human-AI decision-making: The role of human learning in appropriate reliance on AI advice*. ICIS 2023 Proceedings, 14. https://doi.org/10.48550/arXiv.2310.02108

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). *A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making*. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, United States. https://doi.org/10.1145/3514094.3534128

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). *Appropriate reliance on AI advice: Conceptualization and the effect of explanations*. Proceedings of the 28th International Conference on Intelligent User Interfaces, New York, NY, United States. https://doi.org/10.1145/3581641.3584066

Schemmer, M., Kühl, N., & Satzger, G. (2021). *Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence*. arXiv. https://doi.org/10.48550/arXiv.2109.13827

Schoeffer, J., Jakubik, J., Voessing, M., Kuehl, N., & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. In P. Lukowicz, S. Mayer, J. Koch, J. Shawe-Taylor, & I. Tiddi (Eds.), *HHAI 2023: Augmenting human intellect* (pp. 46–59). IOS Press. https://doi.org/10.3233/FAIA230074

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400. https://doi.org/10.1037/0096-3445.130.3.380

Semenova, L., Rudin, C., & Parr, R. (2022). *On the existence of simpler machine learning models*. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, United States. https://doi.org/10.1145/3531146.3533232

Sheridan, T. B. (2019). Extending three existing models to analysis of trust in automation: Signal detection, statistical parameter estimation, and model-based control. *Human Factors*, *61*(7), 1162–1170. https://doi.org/10.1177/0018720819829951

Simchon, A., & Gilead, M. (2024). A psychologically informed approach to "actuarial" decision making. *Decision*, *11*(4), 700–707. https://doi.org/10.1037/dec0000232

Simkute, A., Surana, A., Luger, E., Evans, M., & Jones, R. (2022). *XAI for learning: Narrowing down the digital divide between "new" and "old" experts*. Adjunct Proceedings of the 2022 Nordic Human–Computer Interaction Conference, New York, NY, United States. https://doi.org/10.1145/3547522.3547678

Speith, T. (2022). *A review of taxonomies of explainable artificial intelligence (XAI) methods*. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, United States. https://doi.org/10.1145/3531146.3534639

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81–106). Springer. https://doi.org/10.1007/978-0-306-47630-3_5

Steyvers, M., & Kumar, A. (2023). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231181102

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), Article e2111547119. https://doi.org/10.1073/pnas.2111547119

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L., & Smyth, P. (2024). *The calibration gap between model and human confidence in large language models*. arXiv. https://doi.org/10.48550/arXiv.2401.13835

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, *10*(5), Article e1379. https://doi.org/10.1002/widm.1379

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), Article 307. https://doi.org/10.1186/1471-2105-9-307

Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, *21*(2), 345–352. https://doi.org/10.1093/biostatistics/kxz041

Sun, Z., Xiao, Y., Li, J., Ji, Y., Chen, W., & Zhang, M. (2024). Exploring and mitigating shortcut learning for generative large language models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 6883–6893). ELRA; ICCL. Retrieved June 28, 2024, from https://aclanthology.org/2024.lrec-main.602

Tešić, M., & Hahn, U. (2022). Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that? *Patterns*, *3*(12), Article 100635. https://doi.org/10.1016/j.patter.2022.100635

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition*. Yale University Press.

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 74952–74965). Curran Associates, Inc. Retrieved June 28, 2024, from https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

van Breugel, B., & van der Schaar, M. (2024). *Position: Why tabular foundation models should be a research priority*. 41st International Conference on Machine Learning. Retrieved July 9, 2024, from https://openreview.net/forum?id=amRSBdZlw9

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Veenman, M., Stefan, A. M., & Haaf, J. M. (2024). Bayesian hierarchical modeling: An introduction and reassessment. *Behavior Research Methods*, *56*, 4600–4631. https://doi.org/10.3758/s13428-023-02204-3

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of Big Data and AI. *Columbia Business Law Review*, *2019*(2), 494–620. https://doi.org/10.7916/cblr.v2019i2.3424

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–889. https://heinonline.org/HOL/LandingPage?handle=hein.journals/hjlt31&div=29&id=&page=

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review*, *123*, 735–790. https://doi.org/10.2139/ssrn.3792772

Wachter, S., Mittelstadt, B., & Russell, C. (2024). *Do large language models have a legal duty to tell the truth?* SSRN. https://doi.org/10.2139/ssrn.4771884

Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Liu, J., Qu, Z., Yan, S., Zhu, Y., Zhang, Q., Chowdhury,

M., & Zhang, M. (2024). *Efficient large language models: A survey*. arXiv. https://doi.org/10.48550/arXiv.2312.03863

Wang, J., Oh, J., Wang, H., & Wiens, J. (2018). *Learning credible models*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, United States. https://doi.org/10.1145/3219819.3220070

Wang, Z. J., Zhong, C., Xin, R., Takagi, T., Chen, Z., Chau, D. H., Rudin, C., & Seltzer, M. (2022). *TimberTrek: Exploring and curating sparse decision trees with interactive visualization*. 2022 IEEE Visualization and Visual Analytics (VIS), Oklahoma City, OK, United States. https://doi.org/10.1109/VIS54862.2022.00021

Warren, G., Byrne, R. M. J., & Keane, M. T. (2023). *Categorical and continuous features in counterfactual explanations of AI systems*. Proceedings of the 28th International Conference on Intelligent User Interfaces, New York, NY, United States. https://doi.org/10.1145/3581641.3584090

Williams, S., & Huckle, J. (2024). *Easy problems that LLMs get wrong*. arXiv. https://doi.org/10.48550/arXiv.2405.19616

Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Liu, T., Zhai, X., Yao, W., Li, J., Du, M., & Liu, N. (2024). *Usable XAI: 10 strategies towards exploiting explainability in the LLM era*. arXiv. https://doi.org/10.48550/arXiv.2403.08946

Ye, X., & Durrett, G. (2022). The unreliability of explanations in few-shot prompting for textual reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 30378–30392). Curran Associates, Inc. Retrieved June 28, 2024, from https://proceedings.neurips.cc/paper_files/paper/2022/hash/c402501846f9fe03e2cac015b3f0e6b1-Abstract-Conference.html

Yeom, S., Datta, A., & Fredrikson, M. (2018). Hunting for discriminatory proxies in linear regression models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 1–11). Curran Associates, Inc. Retrieved April 5, 2023, from https://proceedings.neurips.cc/paper/2018/hash/6cd9313ed34ef58bad3fdd504355e72c-Abstract.html

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). *MM-LLMs: Recent advances in multimodal large language models*. arXiv. https://doi.org/10.48550/arXiv.2401.13601

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39–57. https://doi.org/10.1162/tacl_a_00632

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), Article 20. https://doi.org/10.1145/3639372

Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., Kurvers, R. H. J. M., & Herzog, S. M. (2024). *Human–AI collectives produce the most accurate differential diagnoses*. arXiv. https://doi.org/10.48550/arXiv.2406.14981

Zytek, A., Pidò, S., & Veeramachaneni, K. (2024). *LLMs for XAI: Future directions for explaining explanations*. arXiv. https://doi.org/10.48550/arXiv.2405.06064

Received July 21, 2023
Revision received July 22, 2024
Accepted July 25, 2024 ■