Trust Junk and Evil Knobs: Calibrating Trust in AI Visualization

Emily Wall
Emory University

Laura Matzen
Sandia National Laboratories

Mennatallah El-Assady
ETH Zürich

Peta Masters
King's College London

Helia Hosseinpour
University of California Merced

Alex Endert
Georgia Tech

Rita Borgo
King's College London

Polo Chau
Georgia Tech

Adam Perer
Carnegie Mellon University

Harald Schupp
University of Konstanz

Hendrik Strobelt
IBM Research AI, Cambridge, Mass.

Lace Padilla
Northeastern University, Boston

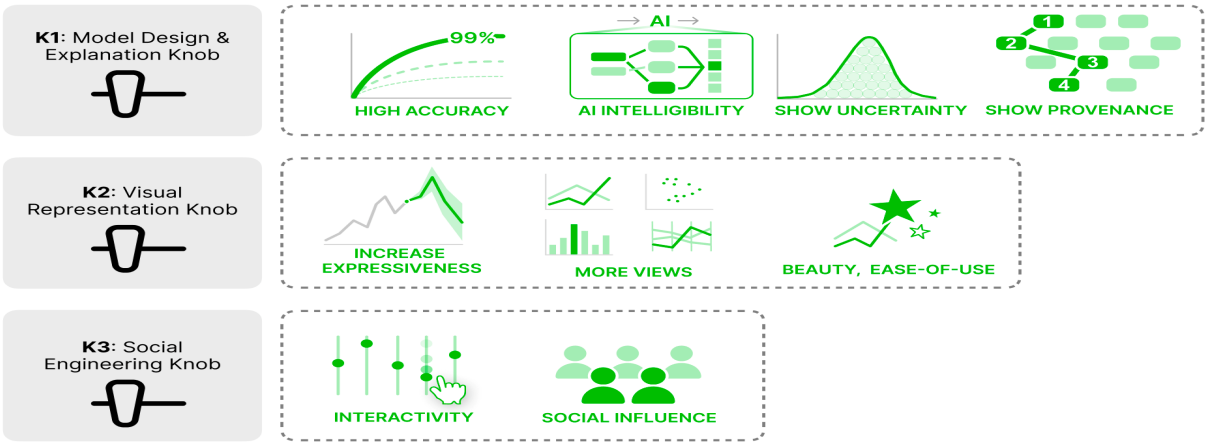


Figure 1: Evil knobs? Turned up too far, trust-enhancing designs may lead to information overload, misrepresentation and deceit.

Abstract

Many papers make claims about specific visualization techniques that are said to enhance or calibrate trust in AI systems. But a design choice that enhances trust in some cases appears to damage it in others. In this paper, we explore this inherent duality through an analogy with "knobs". Turning a knob too far in one direction may result in under-trust, too far in the other, over-trust or, turned up further still, in a confusing distortion. While the designs or so-called "knobs" are not inherently evil, they can be misused or used in an adversarial context and thereby manipulated to mislead users or promote unwarranted levels of trust in AI systems. When a visualization that has no meaningful connection with the underlying model or data is employed to enhance trust, we refer to the result as "trust junk." From a review of 65 papers, we identify nine commonly made claims about trust calibration. We synthesize them into a framework of knobs that can be used for good or "evil," and distill our findings into observed pitfalls for the responsible design of human-AI systems.

Index Terms: Computing methodologies—Artificial intelligence; Human-centered computing—Human computer interaction (HCI)

1 Introduction

Many tools and techniques have been designed to increase or calibrate trust in AI [3, 38, 56, 76, 98], leading to the derivation of design spaces [24], formalized models [42] and frameworks [61]. Users are encouraged to place their trust in entities ranging from training data, models and explanations, through visual mappings and quality metrics, to the creator of the system or regulatory bodies that affect it.

Designers and developers of AI tools have a responsibility for the choices they make during the design process when deciding which components to include that foster trust, and the degree to which they will utilize them. Should they ensure that their visualizations show uncertainty in the AI outcome, for example? What explanations might be provided to increase transparency? Each choice they make can have downstream impact on user trust [68].

Design choices may be thought of as **knobs** that can be selected for inclusion and adjusted during the design process to support appropriate levels of trust in the underlying model. While such choices are usually made with good intentions (to improve usability and promote warranted trust [53]), misuse of the same choices may mislead or manipulate users [23] (see Section 4).

In this paper, we refer to trust-related design choices capable of misleading or manipulating users—whether adversarially (i.e., with intent) or as a result of using them thoughtlessly or turning them up too far—as **evil knobs**. Moreover, when design choices lead to a disconnect between information intended to improve a user's trust in the model and the model itself, we refer to the result as **trust junk**, analogous to the negative effects that visualization embellishment has on graphical integrity, often referred to as "chart junk" [89]. And just as "chart junk" has been shown to have benefits (e.g., improving memorability [8]), so-called "trust junk" may also play a positive role in human-machine teaming.

In what follows, we identify nine common **claims**—where authors have explicitly attributed a change in user trust to a design choice—drawn from examination of 65 papers (Section 3) in relation to the collection and transformation of data, the statistical modeling or AI method used, the user interface and visual presentation of results. We organize these claims into a framework of knobs that can be 'dialed up' to amplify not only warranted but also *un*warranted trust and, at the highest settings, may simply confuse or mislead (Section 4). We conclude with a discussion of guidelines, guardrails, and best practices for the use of researchers, designers and developers of tools that foster trust in AI (Section 5).

2 Background

We contextualize our framework in relation to relevant literature on trust in AI, AI design guidelines, and dark patterns in UI design.

Trust in AI. Historically, research on trust in AI has been focused on one overriding problem: how to ensure that people trust a system enough to use it without simultaneously encouraging them to over-trust it to the extent that they become vulnerable to its inadequacies or imperfections. This quest to achieve **appropriate trust**—that is, trust that increases when it is warranted but decreases when it is not—is articulated in seminal work from Lee and See, describing trust in relation to automation in general [48] and numerous contemporary works continue to explore the notion, particularly in the context of social robotics [22] and autonomous vehicles [67].

The focus on calibrating human trust in AI has fallen largely to HCI practitioners, building on decades of experience encouraging human trust in e-commerce [59], online reviews, and recommendations [82]. These endeavors have led to the development of user-friendly aids and artifacts such as the "online trust signals" investigated by Casado-Aranda et al. [14]. Though well-intentioned, the results may increase user trust while having little or nothing to do with the quality of the products themselves – what we refer to as "trust junk." Trust junk elicits unwarranted trust in AI systems [42] and measures taken to contain the related problem of over-trust by dialing up the volume on AI's trust*worthiness* have spawned whole new sub-disciplines such as explainable AI (XAI) to increase transparency [1, 60] and ethical AI in an attempt to make systems more responsible [17, 26].

The picture is complicated in that good-faith efforts to improve explainability or increase AI transparency sometimes has little or no impact on user trust [98]. At the same time, AI technology may be used in bad faith or, as seen with the popularization of large language models such as ChatGPT, used to generate plausible but misleading information, so-called 'hallucinations' [69] and downright untruths [20]. And just as adversarial machine learning [7] has compromised our ability to depend on ML classifiers, adversarial explanations like those explored by Schneider et al. [77] can compromise XAI, systematically undermining our best attempts at achieving appropriate trust in AI.

AI Guidelines. Guidelines about AI systems and ethics aim to control these troublesome developments [35]. For instance, recommended guidelines about human-centered AI (HCAI) include implementing systems that (1) are reliable, (2) promote safety culture, and (3) are viewed as trustworthy [80]. Other approaches have categorized dimensions of importance in AI guidelines related to concepts such as explainability, transparency, data privacy, and model fairness [51]. Crucially, for these factors to support appropriate user trust, they must be calibrated to fit the user, the AI tool, and the use case. Novice users may over-rely on incorrect advice while experienced users dismiss recommendations regardless of their quality [3, 32, 64]. Similarly, explanations of model behavior and estimations of uncertainty can help users understand context and evaluate reliability [2, 56, 88], but may also make users overconfident [41, 98]. Good initial performance from an AI system can lead to over-reliance on later outputs [64] whereas early errors have an out-sized negative impact on trust [64,65]. Interaction also impacts trust, as human-in-the-loop feedback has been shown to reduce the trust of the participant and their perception of accuracy, even if the system accuracy improves based on their interactions [38].

In summary, a literal interpretation of AI guidelines, absent of context, can be misconstrued, unfairly extrapolated, or inappropriately applied. We attempt to fill a gap in the application of these guidelines by drawing specific attention to ways that even well-intentioned choices may lead to distortion, trust junk, or "evil" misuse.

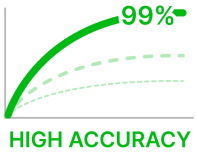
Dark Patterns in UI Design. The term Dark Patterns was coined by Brignull in 2010 following the explosion of web-based e-commerce [36] to describe deceptive or misleading UI/UX design decisions which take advantage of human psychology to manipulate and enforce patterns of behavior in users. Dark Patterns are created with the purpose of limiting freedom of choice, understanding, and agency. The literature highlights five main categories: [33, 39, 54, 57]: *pressure* to take or not take a certain action, *operational constraint* where no decision-making option is provided, *obstruction* which limits user agency inserting obstacles in the user path, *sneaking* which forces unwanted actions on users playing with distraction or confusion, and *misleading* where distractors are used to divert attention. Luguri et al. [54] provide comprehensive empirical evidence of the manipulative power of Dark Patterns, no matter their nuances in terms of complexity or aggressiveness. The authors demonstrate how the strength of the manipulation, rooted in a solid understanding of human psychology [36], is independent of the background, education, and demographic factors of users. This concept inspires our lens of **evil knobs**, where designers may engage design patterns that intentionally leverage psychological constructs to promote over-trust or undue mistrust.

3 Trustworthy Design Claims

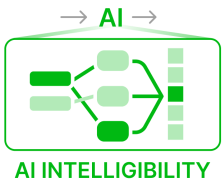
Numerous papers have made claims about features of an AI, data analytic, or visualization system that enhance the user's trust in the tool, as discussed elsewhere [49, 58, 76]. In many cases, these claims are supported by experiments or user studies that compare different designs and evaluate their impact on users' trust [21, 68, 95]. However, there are often contradictory findings: a particular design choice may enhance trust in one application or domain [25] but not in another [38, 94]. In this section, we summarize some of the claims made about design choices used to increase trust in AI systems.

Claims **C1**-**C9** (below) have been synthesized from 40 papers. Building on the corpus provided by a recent survey of human-centered evaluations of AI systems [83], two researchers searched abstracts from conference proceedings 2019-22 of CHI, Vis, EuroVis, TVCG, and IUI using the keywords "trust" and ("design" or "visualization"). Manual search of citations in the 55 identified papers led to 10 further sources. Papers were screened for relevance (i.e., that they related strictly to visualization) by two *other* researchers and further culled to include only those that make explicit claims supported by that paper's findings that some aspect of visualization has an impact on perceived trust. Each of the following claims was encountered in multiple papers. (See `http://tinyurl.com/5ym75dxt` for a complete table of sources.)

[C1] High accuracy models foster trust. If users believe that the system is producing good results, they are more likely to use and trust it. For instance, Hohman et al. motivated the design of their system Gamut by striving for models with a *"high level of accuracy so that users of the probe would trust its predictions were accurate and realistic,"* consequently finding in their study that *"participants were using the model to confirm prior beliefs about the data, slowly building trust that the model was producing accurate and believable predictions"* [37]. Other researchers have found that people have higher trust in models that have higher stated accuracy, even before seeing the results of the model in action [70].

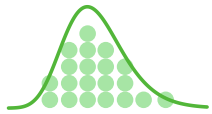


[C2] AI transparency, intelligibility or explainability fosters trust. To achieve appropriate calibration of trust, designers often try to make their models more transparent, intelligible, or explainable, so that users can develop an understanding of how the model works or why it produced a particular result. Transparency has been identified as one of the key factors that impacts users' trust in AI and machine learning models [83] and has subsequently been used to motivate design choices in a number of systems by allowing examination of results from black box models (e.g., [21,46]). For example, Sultanum et al. motivate trustworthy design by *"selectively exposing internal aspects of the automation*

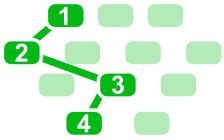


and providing extensive linkage to the original text" [84]. Several prior studies have found that *"[allowing] users to verify whether the system behavior is sound and to judge the appropriateness of the results"* [18] leads to the perception that systems are more reliable and trusted (e.g., as in [5, 27, 87]). Greater intelligibility or interpretability of model outputs is often achieved by providing explanations of why the model produced a specific result [48, 72]. Tintarev and Masthoff posit that *"good explanations could help inspire user trust"* [87].

[C3] Communicating uncertainty fosters trust. Many systems attempt to show as much information as possible to foster trust, especially uncertainty in data or model outputs. Sacha et al. claim a user's trust in a model *"depends on the extent of user's awareness of the underlying uncertainties generated on the system side"* [76]. Similarly, work has found that when a model indicates high confidence in a result, people are more likely to believe the result is accurate, and may change their own assessment to match that of the model [70], even when the model's output is incorrect [85].

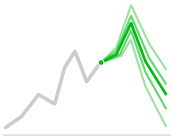

SHOW UNCERTAINTY

[C4] Showing provenance fosters trust. Provenance refers to information about the origin or history of some piece of information or analysis. Lee and See recommend showing a system's past performance, its purpose and design basis, as well as considering the context, cultural and organizational issues that influenced the system's development [48]. These are all factors that relate to the provenance of the tool itself. This kind of information can influence trust because people are more likely to trust a tool or model if they trust its source and the development process that produced it [47, 97]. Provenance can also refer to information about how a specific result was produced. This type of provenance can promote trust-building via "*track[ing] and show[ing] all user interactions*" [29]. Tracking user interactions can help to capture the analytic process and increase the reproducibility of an analysis.



SHOW PROVENANCE

[C5] Expressiveness fosters trust. A visualization is deemed expressive when it showcases, and presents solely, all the relationships within the data [55, 63]. The use of more expressive visualization techniques that convey granular distributional information may aid in understanding e.g., uncertainty in data and therefore increase user trust. These more expressive visualizations can fill in the subtle differences or properties in forecasts that are missed in intervals. For example, they may depict the distribution, the outliers, and a fuller view of the data [86]. However, users often have more trust in representations that are familiar to them [21]. That is, they may have higher levels of trust in information represented via a confidence interval than they would have in a more expressive, but less familiar, representation of the same information. That said, the more visualization creators show expressive uncertainty, the more people will grow accustomed to it and potentially trust it.


INCREASE EXPRESSIVENESS

[C6] Providing more views of the data fosters trust. Several papers have found that providing multiple views can increase users' understanding, which fosters higher levels of trust [73, 74, 96]. Roberts states: "*[Multiple views are] obviously useful in education as the learner may understand the information better through one presentation rather than another*" [74]. This claim also relates to the transparency (**C2**) and interactivity (**C8**) claims. Providing additional data or views can increase transparency and can allow users to select the views that best align with their own preferences or goals.


MORE VIEWS

[C7] Positive user experiences foster trust. Users often have higher trust in a system if it is easy to use or if their user experience is a positive one [19]. Thus, improving the usability of a system can improve user trust. In addition, people may be more trusting of tools that are visually appealing. The "beauty" of a visualization, characterized by vibrant colors and other features of hue and luminance, has been shown to causally affect trust [53]. This phenomenon extends beyond visualization to aspects of human-to-human trust. People tend to trust strangers they regard as more beautiful [93].


BEAUTY, EASE-OF-USE

[C8] Adding interactivity fosters trust. Several studies examine the relationship between trust and interactivity (e.g., with data, models, visualizations). Dietvorst et al. found that *"forecasters who have the ability to adjust an algorithm's forecasts believe it performs better than those who do not"* [25]. Similarly, Lekschas et al. found that interactive labeling was preferred over strictly automatic methods and increased user trust [50]. Across two studies with pathologists, Cai et al. found that *"[interactive] refinement tools increased the diagnostic utility of images found and increased user trust in the algorithm"* [11]. However, in some circumstances, allowing users to interact with a model and provide feedback on its results decreases trust, perhaps because it makes users more aware of the system's errors [38].


INTERACTIVITY

[C9] Social factors influence trust. Social factors that can influence a user's trust in a model or tool may be internal, incorporated into the user interface of the tool itself, or may arise externally. For example, providing users with a personification of the system, such as a virtual agent, can increase their trust in the system [92]. In this example, social factors are internal, incorporated into the design of a system. As an external example, the provision of social information [4] or framing that informs users how other people have used a particular tool may impact a user's trust. Users are more likely to incorporate a new tool into their workflow if it is used and recommended by colleagues. Similarly, as mentioned in the discussion of claim **C4**, people are more likely to trust a visualization if they trust the person or organization that developed it [97].


SOCIAL INFLUENCE

4 Good/Bad Knobs: A Framework of Design Choices

We now categorize the nine claims (**C1**-**C9**) according to that aspect of the design process within which they are most likely to be considered. Each cluster may be regarded as a *knob* that can be adjusted to different levels of intensity and fine-tuned through its constituent parts. As illustrated in Section 3, design choices are usually made with good intentions. However, improper combinations or tuning of the knobs has the potential to create trust junk and trust distortion. Moreover, the same knob that promotes an appropriate level of trust can also be used in an "evil" fashion to intentionally mislead or bias users or encourage them to trust a model that may not be trustworthy.

The goal of this framework is to help designers think about how their design choices might impact user trust, and to consider potential pitfalls that can arise from well-intentioned choices at various stages of the design process. To promote this type of evaluation, we consider the settings for each knob from the perspective of unintentional and intentional distortion, contrasting the positive and negative aspects of each design choice.

Related Claims	Good Outcomes	Trust Distortions	Evil Outcomes
Communicating model accuracy (**C1**)	Appropriate expectations of model performance under different circumstances	Users overlook model errors or second-guess their own judgments	Presenting inaccurate information about model accuracy
Increasing transparency and explainability (**C2**)	Improved understanding of performance and outputs leading to appropriate calibration of trust	Users overwhelmed with information or detail, causing inappropriate calibration of trust	Deliberately misleading explanations that cause inaccurate calibration of trust
Communicating uncertainty (**C3**)	Accurate understanding of uncertainty in model outputs and/or the strengths and limitations of model and data	Misunderstanding of uncertainty, confusion, information overload, incorrect decisions	Oversimplification of model performance, unsupported conclusions, incorrect decisions
Showing provenance (**C4**)	Better understanding of data and/or the analytic process, greater reproducibility	Confusion and information overload, unintentional biases related to user perceptions	

Table 1: The Model Design and Explanation Knob **[K1]**

[K1] Model Design and Explanation

Having identified the domain problem, perhaps the most important design decision to be made concerns the amount and type of information to provide about the subject of the visualization. The design choices that can be tuned at this stage relate to a model's performance, transparency, and explainability. Design claims **C1**, **C2**, **C3** and **C4** contribute to the Model Design and Explanation Knob (see Table 1).

Striving for a high-performing model **C1** is typically a good design choice (save for the often hidden expense of model fairness [16]) but different ways of communicating the model's performance to users can create trust distortions or trust junk. For example, when users are told that a model has high performance or that it has high confidence in a particular result, they are more likely to trust its output even if they have little understanding of how well it actually performs (cf. [70]) or when the model is incorrect (cf. [85]). Thus, telling users that a model has high performance could cause them to overlook model errors or override their own judgments. Information about model performance can also be used maliciously. At the "evil knob" extreme, a designer can give users inaccurate information about a model's performance in order to inappropriately inflate their trust in the model.

While there is plenty of evidence that explanations and other forms of model transparency have an impact on user trust (**C2**), it is not entirely clear when explanations are helpful to users or when they become detrimental. In some cases, efforts at transparency increase users' objective understanding of an algorithm but have no impact on their reported trust in it [15, 81]. In other cases, explanations have no impact or even a negative impact [10, 18, 45]. Some users may ignore explanations altogether [66]. Where explanations *do* influence user trust, it is sometimes in inappropriate ways. For example, researchers observed a placebo effect where explanations that provide no additional information improve trust just as much as explanations that do [28]. That is, users are influenced by the mere existence of an explanation, that is, the *appearance* of transparency, rather than the quality of the explanations. Moreover, providing explanations can persuade users to comply with an incorrect recommendation from a model, even if it clashes with their own assessment [85]. Thus, even well-intentioned explanations can have negative consequences, particularly if they cause users to second-guess their own judgments.

Negative effects on trust are observed for provision of both too little and too much information. Eslami et al. found "*vague and oversimplified language made many existing ad explanations uninterpretable and sometimes untrustworthy*" [31], while Kizilcec found that too much information likewise eroded trust [45]. Some studies also demonstrate a disconnect between transparency that supports high levels of *trust* and transparency that supports high levels of *task performance*. For example, in a study of map-based visualizations, Xiong et al. found that the visualizations that users trusted the most (i.e., those that hit a "sweet spot" of transparency) were *not* the visualizations they selected when asked to complete a decision-making task [95].

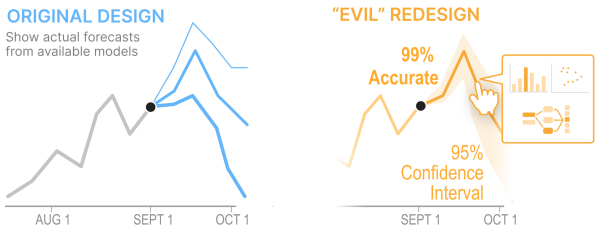


Figure 2: A "good" visualization of multiple forecasts [68] (left) and a potentially "evil" re-design (right). There-design adjusts knobs **K1** (C1 accuracy, C2 transparency/explainability, C3 uncertainty, **K2** (C6 more views) and **K3** (C8 interactivity). It encourages over-reliance and overwhelms the user while providing an illusion of control.

Thus, transparency and explainability make for a volatile setting and can create trust junk if used inappropriately. For example, some approaches to explainability provide users with information about how a model was built or trained. A designer might provide references to papers about deep learning techniques that are similar to the technique they used to develop their model in an effort to build trust in their approach. But if the information is not meaningful to the user or is disconnected from what they need to know it may become trust junk. That is, the information may increase their trust without increasing their understanding of the model. At the "evil knob" extreme, providing inaccurate or misleading explanations could falsely inflate users' trust in a system, particularly if explanations hit a sweet spot in terms of their level of detail, providing enough information to be convincing but not so much that the user feels overwhelmed or begins to question the model's output. In fact, since models are often complex and hard to explain, it may be easier to hit that sweet spot of transparency for fake explanations than for real ones. Difficult though it may be to predict when an explanation may result in trust junk or distortions of trust, it is important for designers to be aware of the possible consequences and attempt to reason with their designs in good faith.

Information about uncertainty (**C3**) can also have a nuanced impact on trust. Users typically report high levels of trust in models that claim high performance and results that claim high certainty. Therefore, truthful representations of uncertainty may have the unintended consequence of degrading user trust overall, rather than helping users to calibrate their trust appropriately. Information about uncertainty can be difficult to communicate so often increases the complexity of a representation, and users may ignore it or make simplifying assumptions that are incorrect [44]. Thus, information about uncertainty has the potential to create distortions of trust even when its presence is necessary to support appropriate understanding of the model. From the adversarial "evil knobs" perspective, uncertainty information could be omitted to falsely inflate user trust. More insidiously, a knowledgeable adversary could take advantage of cognitive biases and shortcuts that have been observed for uncertainty visualizations to nudge users

toward specific biases in reasoning.

Information about provenance (**C4**) can be used to track users' interactions with a model, reducing confusion and increasing trust. But provenance information has the potential to distort trust if it overloads users with information. As discussed in Section 3, 'provenance' may refer to information about the source of the data or the model. Users are more likely to trust a model if they trust its provenance [97], but this type of provenance information can easily become trust junk if it is not meaningfully connected to the functioning of the model. For example, ornamenting a visualization with the logos of respected organizations could unintentionally bias users or, at the extreme, could be used as an evil knob to mislead users into thinking the data or tool is from a reputable source, when actually it is not.

Taken together, 'Model Design and Explanation' settings have a significant impact on trust calibration. Consider an online 'AI Broker': a recommender system that provides the user with advice about which insurance policies they might like to take out. **Appropriate K1** settings will result in the user being aware of the system's success rate, perhaps measured as the level of satisfaction expressed by similar users **(C1)**; understanding how it arrived at its recommendation **(C2)** cognizant of all potential sources of uncertainty, such as the likelihood of the user ever having to make a claim **(C3)** and knowing the source or sources of its data—which may include personal information about the user as well as supplied comparative information about the range of policies **(C4)**. Even a list such as the above of the types of information available (never mind the actual information that would need to be supplied)—while certainly worth considering before deciding whether to trust the model—is more than most users will wish to consciously consider when consulting an online advisor. Dialed up to 10 on every setting, **K1** is likely to result in the **distortion** of information overload. When this occurs, users may (a) disregard much of what they are seeing and (b) rely instead on whichever one particular aspect seems to provide the simplest 'heuristic', such as the system's declared level of confidence **(C3)**, the reputation of its supplier **(C4)** or the experience of other users **(C1)**. It is easy to see how this can become an **evil knob**. Deliberately or inadvertently, a designer may dial up some aspects of the visualization—over-representation making those elements more likely to be ignored—while simplifying other aspects, e.g., by providing a simple 5 stars for the policy other similar users have preferred **(C1)**. A bad actor deliberately manipulates the system to provide 5 stars for the policy it is promoting; but a misguided actor may end up with the same 'evil' result, using the 5 star system because it is simple, clear and well-understood, without noticing that other **K1** design choices may be confusing to—-and ignored by—the user.

[K2] Visual Representations

The second knob in our framework relates to visual representation. Clearly, most of the design claims from Section 3 could be considered under this category, as they are all typically realized through visualization. Claims **C5**-**C7**, however, are *explicitly* related to the mode of presentation and have the greatest direct impact on how those visualizations take shape (see Table 2).

As with **K1**, the impact of **K2** settings is entirely predictable. More expressive (**C5**) visual representations can provide users with improved understanding of model uncertainty, for example, but with the unintended consequence of reducing user trust [68]. Efforts to increase the expressiveness of a visualization can also turn into trust junk. For example, if a designer conveys uncertainty by plotting results from multiple models or variations of the same model, the differences between them may become perceptually indistinguishable as the number of plots increases. Although the intention may be to support appropriate understanding and trust, the representation can become disconnected from any meaningful communication if it surpasses the perceptual and/or reasoning limits of the human viewer.

Similar issues come into play when increasing the number of views (**C6**). Provision of multiple visual components enables users to analyze the data and model outcomes from multiple perspectives. This can improve their understanding, as well as facilitate diverse analysis tasks. However, if not tailored to the target stakeholder group, the provided views may cause information overload or confusion, leading to a distortion of trust. This pitfall is often unintentional. However, if causing information overload is an intentional design goal to falsely provide the users with a sense of completeness, provision of multiple views becomes an **evil knob**.

A positive user experience (**C7**) is clearly desirable. Increases in both usability and beauty have been shown to give users a positive feeling about a tool, leading to increased trust [53]. However, if attractiveness and ease of use are intentionally manipulated to hide flaws in the data analysis or modeling, features meant to enhance the user experience can be used evilly to mislead users instead.

Let us revisit the 'AI Broker' example. The interface to interact with the AI Broker may contain many views **(C6)** that represent different ways of considering insurance policies, such as a bar chart counting the number of insurance products by different providers, distribution curves of the deductible of respective options, and even highly expressive visualizations **(C5)** such as a configurable scatterplot that can show e.g., deductible v. coverage amount. The interface may leverage aesthetic design choices **(C7)** including appealing color palettes, simple font choices, smooth scrolling interactions, and so on. It is again easy to imagine how this knob, when turned to an extreme, can lead to information overload with irrelevant views or may be used 'evilly' to hide poor substance behind a beautiful facade.

[K3] Social Engineering

Our third knob relates to design choices that draw on interpersonal trust and other social factors (see Table 3). Two design claims are particularly relevant here, **C8** and **C9**. However, design choices related to provenance (**C4**) and positive user experience (**C7**) may also be considered in this context. If the user trusts the model's provenance due to interpersonal relationships or their feelings about the organization that produced it, they are probably more likely to trust the model itself. Similarly, a tool that is good-looking and easy to use is likely to seem more polished and professional. It implies that the person or organization who designed the tool knows what they are doing, so users may feel that the tool is more trustworthy on that account.

Interactivity (**C8**) is particularly useful for increasing trust in a model if it allows users to test their hypotheses about how the model works or to refine the information they receive so that it better meets their needs. It can improve how well a model works by allowing for the incorporation of expert knowledge and can support appropriate levels of trust by allowing users to develop a mental model of how the computational model works and what kinds of results they can expect from it under different circumstances. However, if there are too many choices to make, users may not understand all options available to them, leading to suboptimal settings, inefficiency, or confusion. And with too much interactivity, users may also miss important information if they fail to select the optimal combination of settings that would reveal it. These problems may be reduced by including recommended settings or defaults, but dependence on defaults may defeat the object. At some point, adding more and more methods of interacting with a model will have diminishing returns or even negative effects on user comprehension, veering into trust distortion territory. In the "evil knob" context, overwhelming users with a multitude of options for interacting with a model could intentionally generate confusion or the illusion of sophistication. Interactivity can give users a false sense of agency, leading to misplaced trust in the model [94]. Moreover, "placebo buttons," like the non-functional buttons at some pedestrian crossings, fairly common in real-world interactions [62], give users the illusion of agency and are just another form of 'trust junk'.

Several aspects of social factors can be modulated (**C9**). Anthropomorphism—the tendency to imbue human-like characteristics to artificial agents—represents an important setting to affect

Related Claims	Good Outcomes	Trust Distortions	Evil Outcomes
Increasing expressiveness (**C5**)	Improved understanding of uncertainty or subtleties in the data	Decrease in trust due to focus on complexity, outliers	Intentional information overload that obscures relevant info
Providing more views of the data (**C6**)	Supports multiple viewpoints, allows users to select the representations that best support their understanding or task needs	Information overload, users must sift through too much information to find the relevant views	Intentional information overload to provide a false sense of completeness
Creating positive user experiences (**C7**)	Improved usability and beauty that make the interface more pleasant and efficient	Style over substance, frivolous additions that cause incorrect calibration of trust	Slick user experiences or graphics hide the flaws in ineffective tools, encouraging adoption of tools that don't meet user needs

Table 2: The Visual Representation Knob (**K2**)

Related Claims	Good Outcomes	Trust Distortions	Evil Outcomes
Adding interactivity (**C8**)	Users can explore and refine results as needed, have a sense of agency that enhances trust, can engage in hypothesis testing to support a deeper understanding	Arbitrary manipulation of data, too much burden on the user, the potential for confusion	False sense of agency, the illusion of control
Social factors influence trust (**C9**)	Anthropromorphism can create a more positive user experience that leads to higher trust	Anthropromorphism can be annoying or intrusive and is rarely connected to the underlying model meaningfully	Intentional manipulation of social factors that increase or decrease trust
	Social factors can increase adoption of trustworthy tools (see also **C4**)	Social and organizational factors can also impede the adoption of a trustworthy tool	Reliance on a perception of authority to mislead users or push the adoption of inadequate tools
	Social Framing can nudge people toward better decisions, increase social accountability	Creation of unintentional biases in decision making	Misleading info about others' usage or conclusions may inappropriately influence decisions

Table 3: The Social Engineering Knob (**K3**)

trust within the social engineering knob [30, 43]. Anthropomorphism has been proven to influence the perception, adoption, and continued use of a system [52]. Early research related to health, crime, and recruiting revealed that people with a higher tendency to anthropomorphize non-human agents reported higher trust in artificial agents to make important decisions [90]. Beyond hypothetical scenarios, subsequent research has revealed effects of anthropomorphizing across several domains including for instance autonomous driving [91], an XAI design for speech recognition [92], and the sharing of personal information with chatbots [79]. Interestingly, while there is evidence that human speech may be particularly useful for eliciting anthropomorphizing [78] responses, it has been speculated that the voice must sound human-like to be effective [79]. In most cases, anthropomorphism is disconnected from the model itself. That is, although there can be good reasons for anthropomorphism, it is trust junk in the context of our framework, and researchers need to be aware of individual differences in the degree to which its use moderates the amount of trust attributed to artificial agents [90]. More research is needed to better understand when this approach is beneficial to users' performance and when it supports appropriate calibration of trust.

Other factors related to interpersonal trust can be used to influence users' trust and decision-making include 'social framing', which can nudge people towards particular behaviors or beliefs. While typically used to promote pro-social behavior such as increased recycling (e.g., [34]), it is also frequently used to manipulate people's purchasing or donation decisions (e.g., [23]).

Returning to our running example, with **appropriate** social engineering settings, a user will find the AI Broker an engaging, interactive tool with which they can enjoy comparing policies and examining the relative impact of features such as premiums, excesses and deductibles; and one whose advice encourages them to make informed decisions in their own and others' best interests. Turned up to the level of **distortion**, however, **K3** may result in irritation and confusion. A user may find themselves changing values without knowing why or what difference they're making and dealing with a virtual agent as annoying as the notorious Microsoft Paperclip. There is considerable potential for **evil knob** abuse of **K3** settings. A user can be made to feel empowered by the ability to manipulate irrelevant settings such as colors used in scatterplots or their AI Broker's vocal characteristics. Moreover, encouraged to regard the AI Broker as a friend, they may be persuaded to accept whatever recommendation is proffered.

5 Towards Guidance on Trust v. Design Constraints

In this section, we discuss the trade-off between trust and performance in visual designs and describe the pitfalls we have observed. Building on our analysis in Section 4, underlined by findings from a recent paper on *multiple forecast visualizations (MFV)* [68], we consider how these observations inform the way we combat potential issues and mis-uses of trust-enhancing designs.

Lessons on Trust from MFV. Trust can influence the extent to which users rely on the information they see in a visualization [58]. However, even if users trust a visualization, it does not necessarily mean that their interpretations of it will be successful. Recent work on COVID-19 forecast visualizations manipulates all three 'evil knobs' to modulate user trust with mixed results, demonstrating a disconnect between trust and successful interpretation of visualizations [68].

In this study, in which online participants viewed current COVID-19 forecasts for the US, researchers found that the most trusted visualizations were those with *less* visual information, and they were also least likely to lead to correct predictions of the COVID-19 mortality trend. As shown in Figure 3, researchers showed participants line charts with several types of COVID-19 forecast visualizations, including multiple visualizations, depicting forecasts from different forecasting groups (**K2**). Participants rated the trustworthiness of the visualization (*trust*), predicted the COVID-19 mortality trend in the next two weeks (*decision support*), and read information in the visualization (*usability*) [68].

Forecasts depicted by 95% confidence intervals or single forecasts showing no uncertainty (**K1**) were most trusted. Although visualizations with confidence intervals were highly trusted, they produced some surprising outcomes. When researchers used a 95% interval but changed the label indicating either 25%, 50% or 99%, participants did not proportionally scale their trust to the interval size. Further, the confidence interval consistently led to poor future trend predictions.

The poor performance produced by both the 95% confidence interval and single forecast without uncertainty likely resulted from these visualizations having low expressiveness, limiting their ability to display critical features of the forecasts. But high trust sometimes results from a forecast design's simplicity, which can make participants consider it clearer and easier to understand. In an analysis of strategies, participants commonly reported that they trusted visualizations that seemed clearer (**K3**). One participant wrote, *"The graphs using differently colored lines to represent each prediction were overly busy...The graphs with only one line or with a shaded area representing a range were much more clear. To me, one averaged line or one shaded area conveys more confidence in the idea of the prediction models in general and better highlights the important info..."* [68].

Importantly, the authors found a trade-off point between trust and performance, where from 6-9 model forecasts, participants had relatively high trust and performance [68]. The majority of participants (39.5%) reported that the visualizations that included several forecasts were more trustworthy. For example, participants wrote, *"I didn't trust the ones that didn't have so many lines, because it made me think that maybe they didn't investigate enough to give a trustful forecast"* ([68], archived data). The results revealed that trust and decision performance increased as the number of forecasted models increased and then plateaued after 6-9 model forecasts. MFVs with more than nine forecasts had poor usability and provided no additional trust or decision-making benefits (even creating some adverse effects on trust). Hence, designs that enhance trust may come at the expense of other goals, such as accuracy in a user's perception.

Observed Pitfalls. Our framework of knobs, in combination with the lessons above, imply a number of measures that ought to be practiced or avoided. Referencing the underlying claims, we distill our observations into a preliminary set of six potential pitfalls for well-intentioned designers of AI systems to consider.

1. Given the duality of any given design choice (**C-*all***), consider an *adversarial perspective* and ask yourself: what might go wrong? How might a design be misinterpreted? How might these results be misused?
2. Less is *probably* more. Over-amplification of any setting is likely to overwhelm a user (e.g., **C2**-**C8**).
3. Weigh the importance of potentially competing design goals such as *trust* and *decision support* (e.g., more granular model information (**C2**) may harm user experience (**C7**) by causing information overload; however, it may be necessary to support informed decisions in some cases).
4. Don't throw away pedagogically "bad" practices. Sometimes, *ingrained norms* may foster trust (e.g., consider rainbow color maps, well-liked by norm (**C7**) but widely regarded as perceptually ineffective).
5. Keep in mind the two common "evil" uses for knobs and *do not use them*: (1) to take advantage of information overload, (2) rely on perception of authority (often through complexity (**C2**-**C9**)).
6. Substantial *uncertainty information* (**C3**) does not seem to be needed for evoking trust in a visualization [40, 49]. However, the visualization principle of expressiveness (**C5**) can foster trust as it dictates that visualizations show all relationships in the data and only the relationships in the data [55, 63].

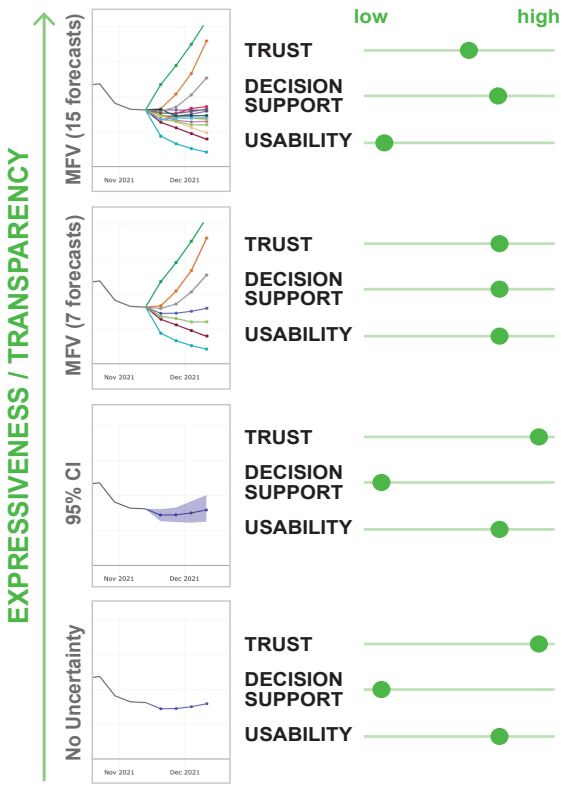


Figure 3: Illustration of trade-offs found in COVID-19 forecast visualizations [68]. The visualization techniques are sorted based on relative expressiveness.

6 Discussion

Evil or Uninformed? While we have framed some uses of design knobs as "evil" to emphasize the *potential* for misuse of trust measures, "evil" knobs are not always the result of ill-intent. In a more charitable interpretation of some misleading designs, it is often the case that individuals are uninformed or even that the design precedes research to suggest more optimal alternatives. For instance, the "cone of uncertainty" visualization commonly used to represent hurricane forecasts has prevailed in weather reporting for decades; however, research has recently shown that non-experts often misinterpret these visualizations compared to ensemble alternatives [9, 75]. These cutting-edge findings take time to propagate from lab theory to practice – this does not imply weather forecasters are "evil".

Similarly, there is controversy around a variety of US election maps that were circulated after the 2016 US presidential election. [13] demonstrates 32 different representations of the same data using choropleths v. cartograms, binary v. blended color maps, state v. county level aggregation, etc. Each gives a very different visceral pre-attentive impression of the 2016 US presidential election, with some maps containing overwhelming amounts of red, others more blue, while still others apparently more ambiguous; yet the underlying data remains the same. Creators of these maps, in some cases, may have "evil" intent; however, others may simply be uninformed about alternative representations. Not everyone who produces charts to analyze or communicate is an expert visualization researcher; not everyone can reasonably be expected to be aware of best practices or state-of-the-art advances in empirical knowledge of the field.

Confounding Guidance. While each design choice in making a user interface for trustworthy AI may independently increase trust, there

are likely situations where combining them in specific ways leads to confounding results. For instance, interactivity gives users the ability to explore and understand how changing parameters of the model result in different outcomes. However, showing this change across too many views may result in change blindness or general ambivalence to the resulting changes due to the number of visual changes in the interface. As a result, users may exhibit satisficing and "simply trust" the AI because it appears that their input is causing updates and lots of them. The interactions between knobs toward trust calibration thus warrants further exploration.

Ethical Implications. Our framework has the potential to enrich the ongoing debate on the role of trust in AI systems and the ethical considerations associated with entrusting critical decisions to such systems. While visualization principles offer numerous benefits for designing trustworthy AI systems, these require an ongoing commitment to ethical design practices and the incorporation of feedback from diverse stakeholders [6, 12]. Our "knobs" categorization highlights the necessity of taking into account differences between the decision-making processes of humans and AI systems [71], the importance of promoting human agency in the context of both direct and indirect interaction with data driven decision systems and the understanding of the determinants behind trust and acceptance of AI systems.

Limitations. We identify two primary limitations. First, a more expansive search for sources, covering earlier years, more venues and a broader range of keywords may capture additional concepts. However, the list of claims is intended to be representative, not comprehensive. Our aim is to provide sufficient research instances to demonstrate the wide range of claims that arise in contemporary visualization research. Second, because we contribute this framework as part of a vision to promote consideration of the duality of design choices, validation of the framework remains an important next step.

Research Opportunities. This framework suggests a number of important research opportunities. First, the framework suggests an agenda toward generating specific empirical support for the efficacy of specific knobs. In particular, as many of the knobs operate on a spectrum (e.g., level of detail in explanations, number of views, etc), future research could begin to titrate the level where these claims begin to break down. Furthermore, the notion of good and bad knobs affecting trust needs to be thought of in reference to the user group and the evolving needs and capabilities of users. Considering the example of COVID-19 forecasting, trust in AI system designs may systematically vary depending on whether the users are public health experts or lay people; i.e., what may appear as good design choice for one group might have the opposite effect on another. Thus, conceiving trust from the perspective of the interaction between the user and the system suggests an agenda in which interface design choices are aligned to particular user groups in order to capture how those choices may best be used to foster trust in AI systems.

7 Conclusion

Drawing on a review of 65 papers, we have identified nine design choices claimed to be capable of fostering or calibrating trust in AI systems. We conceptualize each claim as contributing to a knob which may be adjusted to increase trust but which, if turned up too far, may lead to a distortion such that the degree of trust becomes mismatched to the trustworthiness of the system. Our framework of knobs is based on three aspects of the design process during which their settings are most likely to be considered. We discussed the duality of many of the design claims: how in some cases a measure appears to enhance trust, in other cases detracts from it. We coined the expression "trust junk" for design choices that enhance trust, but have no relationship to the relevant model or data. Importantly, we warn of the potential danger that, in seeking to maximize users' trust in a system, designers may be tempted to employ "evil" knobs: design choices capable of increasing trust but realized in such a way that users are confused, deceived or manipulated though the intention may have been merely to inform. We conclude with observed pitfalls for fellow researchers and designers to consider when designing AI systems.

Acknowledgments

We thank our reviewers for helping to improve this paper.

Laura Matzen's contributions are supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Peta Masters' research is supported by the UKRI Trustworthy Autonomous Systems (TAS) Hub (EP/V00784X/1).

This work has benefitted from Dagstuhl Seminar 22351 "Interactive Visualization for Fostering Trust in ML."

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] J. Amann, D. Vetter, S. N. Blomberg, H. C. Christensen, M. Coffee, S. Gerke, T. K. Gilbert, T. Hagendorff, S. Holm, M. Livne, A. Spezzatti, I. Strümke, R. V. Zicari, and V. I. Madai. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, 1(2):e0000016, 2022.

[3] S. Bayer, H. Gimpel, and M. Markgraf. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, 00(00):1–29, 2021.

[4] E. Beauxis-Aussalet, M. Behrisch, R. Borgo, D. H. Chau, C. Collins, D. Ebert, M. El-Assady, A. Endert, D. A. Keim, J. Kohlhammer, et al. The role of interactive visualization in fostering trust in ai. *IEEE Computer Graphics and Applications*, 41(6):7–12, 2021.

[5] V. Bellotti and K. Edwards. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction*, 16(2-4):193–212, 2001.

[6] S. J. Bennett. Investigating the role of moral decision-making in emerging artificial intelligence technologies. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '19 Companion, p. 28–32. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3311957.3361858

[7] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[8] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE transactions on visualization and computer graphics*, 19(12):2306–2315, 2013.

[9] K. Broad, A. Leiserowitz, J. Weinkle, and M. Steketee. Misinterpretations of the "cone of uncertainty" in florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, 88(5):651–668, 2007.

[10] T. Bruzzese, I. Gao, G. Dietz, C. Ding, and A. Romanos. Effect of confidence indicators on trust in ai-generated profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020.

[11] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proc. Conference on Human Factors in Computing Systems*, CHI '19, pp. 4–14. ACM, 2019. doi: 10.1145/3290605.3300234

[12] F. M. Calisto, N. Nunes, and J. C. Nascimento. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies*, 168:102922, 2022.

[13] Cartogeek. Thematic maps of the 2016 presidential election (lower 48 states). `https://carto.maps.arcgis.com/apps/MinimalGallery/index.html?appid=b3d1fe0e8814480993ff5ad8d0c62c32`, 2020.

[14] L.-A. Casado-Aranda, A. Dimoka, and J. Sánchez-Fernández. Consumer processing of online trust signals: a neuroimaging study. *Journal of Interactive Marketing*, 47:159–180, 2019.

[15] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders. In *Proc. Conference on Human Factors in Computing Systems*, CHI '19, pp. 559:1–559:12. ACM, 2019. doi: 10.1145/3290605.3300789

[16] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[17] M. Coeckelbergh. *AI ethics*. Mit Press, 2020.

[18] S. Coppers, J. Van den Bergh, K. Luyten, K. Coninx, I. van der Lek-Ciudin, T. Vanallemeersch, and V. Vandeghinste. Intellingo: An Intelligible Translation Environment. In *Proc. Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, 2018. doi: 10.1145/3173574.3174098

[19] E. Costante, J. Den Hartog, and M. Petkovic. On-line trust perception: What really matters. In *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, pp. 52–59. IEEE, 2011.

[20] E. N. Crothers, N. Japkowicz, and H. L. Viktor. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002, 2023. doi: 10.1109/ACCESS.2023.3294090

[21] A. Dasgupta, J. Lee, R. Wilson, R. A. Lafrance, N. Cramer, K. Cook, and S. Payne. Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):271–280, 2017. doi: 10.1109/TVCG.2016.2598544

[22] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

[23] D. Defazio, C. Franzoni, and C. Rossi-Lamastra. How pro-social framing affects the success of crowdfunding projects: The role of emphasis and information crowdedness. *Journal of Business Ethics*, 171(2):357–378, 2021.

[24] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li. Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle. In *Designing Interactive Systems Conference 2021*, pp. 1591–1602, 2021.

[25] B. J. Dietvorst, J. P. Simmons, and C. Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

[26] V. Dignum. Responsibility and artificial intelligence. *The oxford handbook of ethics of AI*, 4698:215, 2020.

[27] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[28] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pp. 1–6, 2019.

[29] M. El-Assady, F. Sperrle, O. Deussen, D. Keim, and C. Collins. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):374–384, 2019. doi: 10.1109/TVCG.2018.2864769

[30] N. Epley, A. Waytz, and J. T. Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

[31] M. Eslami, S. R. Krishna Kumaran, C. Sandvig, and K. Karahalios. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proc. Conference on Human Factors in Computing Systems*, CHI '18, pp. 432:1–432:13, 2018. doi: 10.1145/3173574.3174006

[32] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1), 2021. doi: 10.1038/s41746-021-00385-9

[33] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs. The dark (patterns) side of ux design. CHI '18, p. 1–14. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174108

[34] L. Grazzini, P. Rodrigo, G. Aiello, and G. Viglia. Loss or gain? the role of message framing in hotel guests' recycling behaviour. *Journal of Sustainable Tourism*, 26(11):1944–1966, 2018.

[35] T. Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, 2020.

[36] Harry Brignul. Dark patterns, 2010. `https://www.darkpatterns.org`, Last accessed on 2022-10-13.

[37] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proc. Conference on Human Factors in Computing Systems*, CHI '19, pp. 579:1–579:13. ACM, 2019. doi: 10.1145/3290605.3300809

[38] D. Honeycutt, M. Nourani, and E. Ragan. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, pp. 63–72, 2020.

[39] L. Hornuf and S. Mangold. *Digital Dark Nudges*, pp. 89–104. Springer International Publishing, Cham, 2022. doi: 10.1007/978-3-031-04063-4_5

[40] J. Hullman. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics*, 26(1):130–139, 2019.

[41] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry*, 11(1), 2021. doi: 10.1038/s41398-021-01224-x

[42] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635, 2021.

[43] T. Jensen. Disentangling trust and anthropomorphism toward the design of human-centered ai systems. In H. Degen and S. Ntoa, eds., *Artificial Intelligence in HCI*, pp. 41–58. Springer International Publishing, Cham, 2021.

[44] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE transactions on visualization and computer graphics*, 27(2):272–282, 2020.

[45] R. F. Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 2390–2395, 2016.

[46] R. Krueger, J. Beyer, W.-D. Jang, N. W. Kim, A. Sokolov, P. K. Sorger, and H. Pfister. Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):227–237, 2020. doi: 10.1109/TVCG.2019.2934547

[47] A. Kuznetsov, M. Novotny, J. Klein, D. Saez-Trumper, and A. Kittur. Templates and trust-o-meters: Towards a widely deployable indicator of trust in wikipedia. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2022.

[48] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[49] D. Leffrang and O. Müller. Should i follow this model? the effect of uncertainty visualization on the acceptance of time series forecasts. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 20–26. IEEE, 2021.

[50] F. Lekschas, B. Peterson, D. Haehn, E. Ma, N. Gehlenborg, and H. Pfister. PEAX: Interactive Visual Pattern Search in Sequential Data Using Unsupervised Deep Representation Learning. *Computer Graphics Forum*, 39(3):167–179, 2020. doi: 10.1111/cgf.13971

[51] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. Trustworthy ai: From principles to practices. *arXiv preprint arXiv:2110.01167*, 2021.

[52] M. Li and A. Suh. Machinelike or humanlike? a literature review of anthropomorphism in ai-enabled technology. 01 2021. doi: 10.24251/HICSS.2021.493

[53] C. Lin and M. A. Thornton. Fooled by beautiful data: Visualization aesthetics bias trust in science, news, and social media. 2021.

[54] J. Luguri and L. J. Strahilevitz. Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1):43–109, 2021.

[55] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[56] A. F. Markus, J. A. Kors, and P. R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies.

Journal of Biomedical Informatics, 113(July 2020):103655, 2021.

[57] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359183

[58] E. Mayr, N. Hynek, S. Salisu, and F. Windhager. Trust in information visualization. In *TrustVis@ EuroVis*, pp. 25–29, 2019.

[59] D. H. McKnight, V. Choudhury, and C. Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, 13(3):334–359, 2002.

[60] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[61] S. Mohseni, N. Zarei, and E. D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[62] J. W. Moore. What is the sense of agency and why does it matter? *Frontiers in psychology*, 7:1272, 2016.

[63] T. Munzner. *Visualization analysis and design*. CRC press, 2014.

[64] M. Nourani, J. King, and E. Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, pp. 112–121, 2020.

[65] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, pp. 340–350, 2021.

[66] M. Nyre-Yu, E. S. Morris, B. C. Moss, C. Smutz, and M. Smith. Considerations for deploying xai tools in the wild: Lessons learned from xai deployment in a cybersecurity operations setting. Technical report, Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2021.

[67] K. Okamura and S. Yamada. Adaptive trust calibration for supervised autonomous vehicles. In *Adjunct proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications*, pp. 92–97, 2018.

[68] L. Padilla, R. Fygenson, S. C. Castro, and E. Bertini. Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations. 2022.

[69] V. Rawte, A. Sheth, and A. Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

[70] A. Rechkemmer and M. Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2022.

[71] U. Rehman, F. Iqbal, and M. U. Shah. Exploring differences in ethical decision-making processes between humans and chatgpt-3 model: a study of trade-offs. *AI and Ethics*, pp. 1–11, 2023.

[72] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[73] J. Ritchie, D. Wigdor, and F. Chevalier. A lie reveals the truth: Quasimodes for task-aligned data presentation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019.

[74] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Fifth international conference on coordinated and multiple views in exploratory visualization (CMV 2007)*, pp. 61–71. IEEE, 2007.

[75] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016.

[76] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2015.

[77] J. Schneider, J. Handali, M. Vlachos, and C. Meske. Deceptive ai explanations: Creation and detection. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, pp. 44–55, 2022.

[78] J. Schroeder and N. Epley. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, 145(11):1427, 2016.

[79] J. Schroeder and M. Schroeder. Trusting in machines: How mode of interaction affects willingness to share personal information with machines. 2018.

[80] B. Shneiderman. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

[81] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proc. Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, 2020. doi: 10.1145/3313831.3376624

[82] B. A. Sparks and V. Browning. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism management*, 32(6):1310–1323, 2011.

[83] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim. A survey of human-centered evaluations in human-centered machine learning. In *Computer Graphics Forum*, vol. 40, pp. 543–568. Wiley Online Library, 2021.

[84] N. Sultanum, D. Singh, M. Brudno, and F. Chevalier. Doccurate: A Curation-Based Approach for Clinical Text Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):142–151, 2019.

[85] H. Suresh, N. Lao, and I. Liccardi. Misplaced trust: measuring the interference of machine learning in human decision-making. In *12th ACM Conference on Web Science*, pp. 315–324, 2020.

[86] **Padilla, L.**, M. Kay, and J. Hullman. Uncertainty Visualization. In R. A. Piegorsch, Walter W. aand Levine, H. H. Zhang, and T. C. M. Lee, eds., *Computational Statistics in Data Science*, chap. 21, pp. 405–421. Wiley, Oxford, 2022.

[87] N. Tintarev and J. Masthoff. Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on Recommender systems*, pp. 153–156, 2007.

[88] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. *arXiv*, (Ml):1–21, 2019.

[89] E. R. Tufte. The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, 7(3):15, 1985.

[90] A. Waytz, J. Cacioppo, and N. Epley. Who sees human? the stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, 2010.

[91] A. Waytz, J. Heafner, and N. Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52:113–117, 2014.

[92] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André. "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, 15(2):87–98, 2021.

[93] R. K. Wilson and C. C. Eckel. Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, 59(2):189–202, 2006.

[94] G. Wu, X. Hu, and Y. Wu. Effects of perceived interactivity, perceived web assurance and disposition to trust on initial online trust. *Journal of Computer-Mediated Communication*, 16(1):1–26, 2010.

[95] C. Xiong, L. Padilla, K. Grayson, and S. Franconeri. Examining the components of trust in map-based visualizations. In *1st EuroVis Workshop on Trustworthy Visualization, TrustVis 2019*, pp. 19–23. The Eurographics Association, 2019.

[96] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020.

[97] R. Zehrung, A. Singhal, M. Correll, and L. Battle. Vis ex machina: An analysis of trust in human versus algorithmically generated visualization recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021.

[98] Y. Zhang, Q. Vera Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020.