Patterns

CellPress
OPEN ACCESS

Review

How transparency modulates trust in artificial intelligence

John Zerilli,[1,*] Umang Bhatt,[2,3] and Adrian Weller[2,3]

[1]Institute for Ethics in AI and Faculty of Law, University of Oxford, St Cross Building, St Cross Road, Oxford OX1 3U, UK
[2]Leverhulme Centre for the Future of Intelligence and Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK
[3]The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK
*Correspondence: john.zerilli@gmail.com

https://doi.org/10.1016/j.patter.2022.100455

THE BIGGER PICTURE Recent advances in artificial intelligence (AI) and machine learning have brought the study of human-AI (HAI) teams into sharper focus. An important set of questions for those designing HAI interfaces concerns trust—specifically, human trust in the AI systems with which they form teams. We review the literature on how perceiving an AI making mistakes violates trust and how such violations might be repaired. In doing so, we discuss the role played by various forms of algorithmic transparency in the process of trust repair, including explanations of algorithms, uncertainty estimates, and performance metrics.

SUMMARY

The study of human-machine systems is central to a variety of behavioral and engineering disciplines, including management science, human factors, robotics, and human-computer interaction. Recent advances in artificial intelligence (AI) and machine learning have brought the study of human-AI teams into sharper focus. An important set of questions for those designing human-AI interfaces concerns trust, transparency, and error tolerance. Here, we review the emerging literature on this important topic, identify open questions, and discuss some of the pitfalls of human-AI team research. We present opposition (extreme algorithm aversion or distrust) and loafing (extreme automation complacency or bias) as lying at opposite ends of a spectrum, with algorithmic vigilance representing an ideal mid-point. We suggest that, while transparency may be crucial for facilitating appropriate levels of trust in AI and thus for counteracting aversive behaviors and promoting vigilance, transparency should not be conceived solely in terms of the explainability of an algorithm. Dynamic task allocation, as well as the communication of confidence and performance metrics—among other strategies—may ultimately prove more useful to users than explanations from algorithms and significantly more effective in promoting vigilance. We further suggest that, while both aversive and appreciative attitudes are detrimental to optimal human-AI team performance, strategies to curb aversion are likely to be more important in the longer term than those attempting to mitigate appreciation. Our wider aim is to channel disparate efforts in human-AI team research into a common framework and to draw attention to the ecological validity of results in this field.

INTRODUCTION

The study of human-machine systems is central to a variety of behavioral and engineering disciplines, including management science,[1–3] human factors,[4–7] robotics,[8–13] and human-computer interaction.[14–24] Recent advances in artificial intelligence (AI) and machine learning have brought the study of human-AI (HAI) teams into sharper focus. An important set of questions for those designing HAI interfaces concerns trust: specifically, human trust in the algorithmic systems with which they form teams. Trust in machines has been defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."[17,25] More precisely, trust is "a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another."[26] Trust is therefore a subjective attitude and attribute of the vulnerable party, to be distinguished from *trustworthiness*, which is an *objective* attribute of the trustee. Just as human collaboration would be impossible without some degree of trust between team members, some form of trust in algorithmic systems is necessary for HAI teams to perform smoothly and effectively. It follows too that if trust is ever violated, its repair will be crucial in any attempt to rehabilitate team performance.

Here, we briefly review the literature on how perceiving an AI make mistakes violates trust and how such violations might be repaired. In doing so, we discuss the role played by various forms of algorithmic transparency in the process of trust repair. We then identify and discuss two important questions left open in this literature: first, concerning what effects the size, frequency, type, and distribution of errors have in the violation and repair of trust, and second, concerning how various forms of transparency—in particular explanations of algorithms, confidence and performance metrics, and dynamic allocation strategies—fare comparatively in the process of trust repair. We suggest that while transparency may be crucial for facilitating trust in AI and thus for counteracting aversive behaviors, transparency should not be conceived solely in terms of explainability. Our final section discusses some of the pitfalls of HAI team research. In particular, we worry that the ecological validity of results in this field is not sufficiently appreciated—at least in practice.

We should lodge three important caveats at the outset. The first concerns the nature of the trust in question, given that trust is, in the first instance, an interpersonal attitude between humans, not between humans and machines. Interpersonal trust has been the subject of investigation in organizational and social psychology for several decades,[27–33] and in these fields, trust is understood to be influenced by at least two factors: (1) the competence of the trustee and (2) the degree to which the trustee exhibits good faith/benevolent intentions—e.g., in a contractual setting, the desire to support the other party's efforts in performing the contract—but, more generally, the absence of ill will or ulterior motives in the trustee.[29–33] Recast into language more appropriate for artificial agents, we can take *competence* to denote a system's accuracy and *good faith* to denote a system's transparency, as judged by a range of criteria including, but not limited to, its explainability. It is true that good faith is not, strictly speaking, the same thing as transparency, and that transparency is often a means of verifying good faith (as well as accuracy). However, it is also true that transparency can itself be an *expression* of good faith on the trustee's part, as when someone who is "open" or "forthright" is understood to harbor no ill will or hidden agenda. In other words, while good faith encompasses more than transparency, it often encompasses *at least* that much. Note also that, throughout this paper, we take transparency to mean any information provided about an AI system beyond its model outputs. By explainability, we mean information that specifically helps to understand how or why a system produced its outputs.[34]

Secondly, accuracy and transparency are by no means the only antecedents of trust in embedded AIs.[35,36] Other important, if less marked, determinants of trust in automation include ergonomic and demographic factors, team size and composition (e.g., in terms of active versus passive users), and task type and complexity.

Finally, we note that the AIs considered in this paper are all examples of what some have termed "embedded AIs," as opposed to AI-enabled virtual agents (e.g., Siri or Alexa) and robots (e.g., Pepper or Roomba).[35] Embedded AIs are forms of AI that are "invisible to the user, embedded inside of a computer or other tool" and which thus lack "a visual representation or a distinguished identity."[35] Common examples would be smartphone apps, e-mail spam filters, ranking algorithms, and recommender systems. Less obvious examples include business systems and automated decision software (e.g., customer credit rating algorithms, offender recidivism risk tools, etc.).

THE EFFECTS OF ERROR AND TRANSPARENCY ON TRUST

In an ideal world, only systems that are trustworthy would be trusted. Distrust may be justified whenever a system performs considerably worse than a human (or human team) acting alone, or whenever a system is opaque or ethically suspect. But distrust is problematic when the distrusting behavior to which it leads—what has been termed algorithm "aversion" —is really an *overreaction* to having witnessed the system's mistakes.[5,14,15,37] In the most extreme case, algorithm aversion results in a refusal to engage with a system at all or a blatant disregard of its recommendations—an attitude we term "opposition."

Conversely, there is such a thing as *too much* trust—algorithm "appreciation"[3]—or overtrust, where a human is so impressed by a system that they cease actively monitoring its outputs[4,5] and in the limiting case follow its every recommendation without question—an attitude we term "loafing." As one might have guessed, appreciation is not a problem for systems that pass a very high threshold of accuracy[38,39] (see Box 1). Accordingly, the AIs of interest to HAI team research are generally trustworthy in the sense that they are adept at performing a particular task, but not so adept that overtrust ceases to be a problem (cf., Bansal et al.[37]) and yet not so error prone that algorithm aversion becomes rational. Both aversion and its opposite, appreciation, are inappropriate attitudes toward systems that are generally trustworthy in this sense.[4,14]

To our knowledge, these various attitudes have never been cast within a single frame of reference. Papers overwhelmingly tend to problematize overtrust or distrust, failing to demonstrate that both phenomena should be understood as part of a broader inquiry into HAI teams, and that any one system can engender any of the above attitudes. Hence, we envisage opposition and loafing as lying at opposite ends of a spectrum, with algorithmic "vigilance" representing an ideal mid-point between them and aversion and appreciation lying mid-way between this ideal and each of the two extremes (Figure 1). Algorithmic vigilance, as we will use the term, is an attitude of active user engagement and healthy skepticism. It marks the level of trust that a human (or human team) should display toward an AI from the point of view of optimal HAI team performance. Confusingly, this attitude is sometimes given the name "complementarity," presumably to indicate that some ideal division of labor has been struck between human and machine, such that humans will focus on tasks too difficult for machines and vice versa.[37]

But complementarity in this sense may be compatible with human loafing (see Box 1), so we prefer the term vigilance.

What counts as vigilance may differ from case to case depending on the AI under consideration. If vigilance is observed over time t, each non-ideal attitude of trust v_{oppose}, v_{avert}, $v_{appreciate}$, and v_{loaf} might be modeled as a related function (Figure 2A).

In human factors engineering and human-computer interaction, overtrust has been extensively researched for close to four decades.[4] In human factors, the phenomenon goes by the

Box 1. Which AIs are the target of human-AI team research?

Human-AI (HAI) team researchers hail from a variety of behavioral and engineering disciplines, including management science, human factors, robotics, and human-computer interaction. HAI team research is concerned with the alleviation of user distrust and overtrust in AI, where such attitudes are likely to impede optimal HAI team performance. An AI that fares worse than a human (or human team) acting alone will rightly arouse distrust. An AI that is vastly superior to a human (or human team) acting alone will unproblematically elicit overtrust. But systems of the first kind are unlikely to be deployed, unless the HAI team deploying them can still outperform humans acting alone, while systems of the second kind are rare in team settings, since humans may be superfluous once a machine can perform so much better than a human (or human team) acting alone.[37] That leaves a wide range of AI systems as the focus of HAI team research. Humans acting alone will be better than some of these, but not better than the HAI team comprising them; the rest of these systems will be better than the humans acting alone, but not better than the HAI team comprising them. For many systems in this range, the attitude conducive to optimal HAI team performance will be vigilance, since both aversion and opposition, as well as appreciation and loafing, will impede optimal HAI team performance (see Figure 1 for the meaning of these terms). Is there a way of schematically demarcating the range of such systems? Perhaps surprisingly, no one has ever attempted to specify the class of systems that is the proper target of HAI team research. But without a clear, shared understanding of which systems require vigilance, which do not, and which should not be used at all, investigations into a large array of systems, each having different levels of reliability, make for a cluttered and confusing terrain.

Assume (plausibly) that every human (or human team) interacting with an AI in the specified range will introduce human errors (e.g., Dietvorst et al.[15] and Bansal at al.[37]) Assume further (for simplicity) that all errors are equally significant, be they human or AI. Let the rate at which humans introduce errors be denoted H, and the rate at which humans spot AI errors be denoted S. As we said, either the AI acting alone will fare better than the humans acting alone, but not better than the HAI team; or the humans acting alone will fare better than the AI acting alone, but not better than the HAI team (we ignore the case where humans acting alone can outperform both the AI and the HAI team, as the AI here would simply not be deployed). Then whenever $H < S$, humans will be spotting more AI errors than the errors they themselves introduce. By contrast, whenever $H > S$ the AI will fare better than both the humans acting alone and the HAI team, because in this case the humans will be introducing more errors than those they are able to spot in the AI's outputs. Systems performing at or higher than the level at which $H > S$ will be best served by removing humans from the loop altogether.[15,37] If for whatever reason humans are kept in the loop, however, the emergence of appreciation and loafing will not be detrimental to HAI team performance.

To illustrate, we plot user trust as a function of system reliability in Figure 3. The plot depicts a well-calibrated user trust function over a range of system performance levels (trust is said to be "well calibrated" when user expectations match system capabilities). Assume that performance at the $H < S$ level marks the point at which a system performs better than a user alone, but not better than an HAI team (e.g., assume that a human alone makes 200 errors, an AI alone makes 100 errors, but that the HAI team will make only 40 errors, because the human spots all 100 AI errors and introduces only 40 of their own for a net total of 40 HAI team errors). As a system's performance gradually improves on this benchmark, S falls because there are progressively fewer errors for the user to spot (e.g., AI_2 will make 99 errors, AI_3 will make 98 errors, etc., while—we assume for simplicity—a user will continue to introduce 40 errors). Eventually a system will reach the point at which $H = S$ (e.g., AI_{61} will make 40 errors). Any system whose performance exceeds this level (i.e., when $H > S$) will perform better than the HAI team (e.g., AI_{62} will make 39 errors, but while the human will spot all 39 they will introduce 40 of their own, for a net total of 40 HAI team errors). Thus, when a system performs at the $H < S$ level, vigilance will be the ideal user response. When a system performs at the $H > S$ level, loafing will be the ideal user response.

We said that systems performing at or higher than the $H > S$ level will be best served by removing humans from the loop altogether. However, this may not be technically, ethically, or politically feasible. In any event, as we noted, the emergence of appreciation and loafing in such cases will not be detrimental to HAI team performance. But to the extent that these systems are not invulnerable to errors that a human might witness, the risk of aversion and opposition will persist. Strategies to mitigate this risk, such as allowing humans to manipulate the algorithm even if doing so may degrade the system's performanc,[15] are still preferable to giving aversion and opposition free rein (again, so long as the HAI team performs better than the humans acting alone).[15]

names of "automation complacency" and "automation bias."[40] Although similar, these effects are not the same. Automation complacency describes the state of passivity, diffidence, or deference into which the user of a system falls when uncritically relying on a technology they deem more proficient than themselves.[41] In effect, it is the failure to attend to the possibility that a system may be wrong through failure to *seek out* either confirmatory or disconfirmatory evidence.[7] Automation bias is a more extreme variant of this attitude and manifests when a human user actively prefers a system's signals over actual—i.e., overtly—contradictory information, including information from more reliable sources such as the user's own senses.[7,41] Crucially, it is the perception of a system's superior performance that induces these states: they are rarely observed when a system is considered liable to even occasional error.[5,7,42–44]

By contrast, algorithm aversion has not been nearly as well researched or theorized. But some results are notable. Users of AI in many lab-based settings have been shown to display unrealistically high levels of trust initially, only for that trust to drop precipitously in response to seeing a system err.[5,14,15] Users then typically retreat to human judgment, even when doing so leads demonstrably to even more errors.[5,14,15] For example, during an incentivized task, when given the choice between relying on their own judgment exclusively or relying on an algorithm's

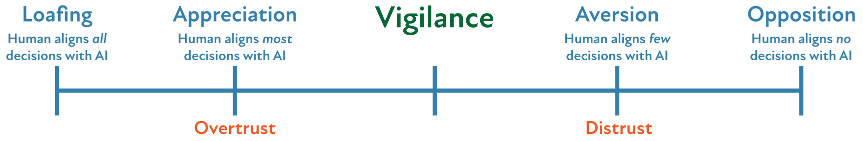
Loafing
Human aligns *all* decisions with AI
Appreciation
Human aligns *most* decisions with AI
Vigilance
Aversion
Human aligns *few* decisions with AI
Opposition
Human aligns *no* decisions with AI
Overtrust
Distrust

Figure 1. Scale of user attitudes toward AI in human-AI teams

forecasts exclusively, most participants who had not seen the algorithm perform chose to rely on the algorithm exclusively, while most of those who had seen the algorithm perform (and hence err) chose to rely on human judgment, despite observing the algorithm's better performance.[14] It has been suggested that this effect is greater for obvious errors than for subtle ones, because obvious errors can quite drastically upset a user's initially high expectations of a system's competence.[5] Moreover, a user's expertise can affect their perception of machine errors.[3] Users who are expert or self-confident in tasks that have been delegated to automation tend to ignore machine advice[45] and, as a result, make less-accurate predictions relative to lay people willing to follow machine advice.[5,14,15]

The pattern of trust → error → distrust, in which trust becomes difficult to restore despite impressive system performance, could be explained by users' "diminishing sensitivity to error." Over the course of five studies, Dietvorst and Bharti[46] found that participants displayed error intolerance when confronted with decision makers that were highly reliable on average but incapable of perfect forecasts, and error tolerance when confronted with decision makers that were less reliable on average but that had at least a chance of making near-perfect forecasts. If users have diminishing sensitivity to error, it would plausibly explain why AIs that make even a single error are penalized so harshly: users' hopes for near-perfect automated forecasting having thus been dashed, the more volatile and error-prone decision-making option (human judgment) suddenly looks like the most appealing one (human forecasters can at least *stumble* on near-perfect forecasts after all). In any event, errors seem to have a stronger impact on trust than correct outputs.[5,7] This phenomenon is indeed so pronounced that cumulative feedback about a system's superior performance presented at the end of a task session may not be enough to counteract users' misgivings after having had their expectations disappointed over the course of a task session.[5]

Curiously, while higher levels of trust generally lead to greater reliance, trust and reliance are not monotonic. An untrustworthy system may rightly arouse distrust (measured subjectively by self-evaluation and report) and yet continue to be relied upon (judging by actual usage data).[5,7,47] The converse of this situation has also been observed, so that even when the subjective feeling of trust eventually *recovered* after witnessing a system failure, immediate post-failure behavior (e.g., scrupulous cross-checking) did not revert to the pre-failure norm.[7]

In the remainder of this section, we single out four categories of transparency—explanations, performance metrics, dynamic allocation strategies, and confidence information—that we think have special significance in HAI team coordination.

Explanation as a form of transparency

No doubt the most pertinent form of transparency is explanation, which can enhance a user's understanding of how an algorithm works and hence why it might commit the sorts of errors it does.[5,37,48] While important, and a well-attested means of establishing appropriate levels of trust,[49] explanations can easily backfire. Some explanations of AI systems, for example, appear to induce automation complacency.[5,37,50] Feature importance explanations—which discover which input features exert the most influence on a model's outputs—are particularly prone to misleading users in this regard,[51–53] although similar example-based explanation methods have, admittedly, been shown to be conducive to HAI team performance.[54,55] In the same vein, when explanations are provided before users are in a position to assess a situation for themselves, users may be led to anchor on the first data they receive, conditioning subsequent deliberation.[37] More perversely, "too much transparency can cause people to incorrectly follow a model when it makes a mistake, due to information overload."[24] On other occasions, poor or confusing explanations can lead to algorithm aversion.[24]

Performance metrics

Many of these results could easily lead one to the cynical conclusion that the best way for AI systems to promote the right amount of trust is simply by shielding users from information about the system's decisions—in effect, by being less transparent.[5,14] (Dzindolet et al.[5] report that "eliminating operators' awareness of an automated decision aid's obvious errors [through blinding the participants to the decisions of the aid] was useful in promoting appropriate automation reliance if participants were continually reminded of their and their aid's performance. Unfortunately, applying these techniques outside the laboratory is problematic. It would not be reasonable to provide someone with an automated decision aid but not allow them to see the decisions the aid has made.") Yet there is reason to believe that a better calibration of trust to a system's actual level of accuracy can be achieved by providing more of the *right kind* of transparency: not just cumulative performance feedback (delivered at the end of a task session), but *continuous* performance feedback that allows the user to maintain a better picture of the system's relative superiority in real time[5,56] (see Figure 2B). Some researchers have even noticed a pattern in the way accuracy information interacts with user attitudes. Metainformation about low-reliability automation runs the risk of promoting overtrust (as measured by higher trust ratings), but metainformation about high-reliability automation seems to have the opposite effect. Presumably this is because, in the first case, users are placed on notice, ready to step in and override the system when it fails, which could, perversely, contribute to a sense that the system is actually more reliable than it is; while, in the second case, metainformation may consolidate users' unrealistic expectations, which are inevitably contradicted on witnessing errors, with the attendant fallout.[57]

User control and dynamic allocation

Because explanations ultimately satisfy a need to be in control, an effective alternative strategy may be to allow users a degree

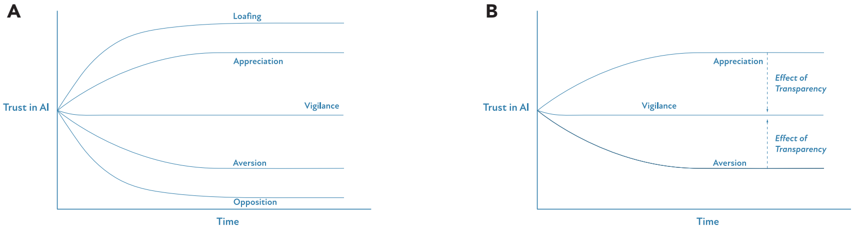

Figure 2. User trust in automation after witnessing system failures

(A) Five possible trust trajectories over time. Notice that the default attitude toward automation is generally one of high trust that falls by some measure in response to seeing a system err. The vigilant user of AI recalibrates their initially unrealistic estimate of a system's capabilities gradually, but not to the point where their attitude becomes aversive.

(B) The hypothesized role of transparency in trust calibration

of latitude over whether to accept an algorithm's outputs at face value. For instance, provided that they can modify its forecasts, users are apparently willing to take an algorithm seriously even after seeing it make occasional mistakes. What is more, the precise degree of control seems to be irrelevant: the ability to modify a forecast even slightly may be sufficient to induce appropriate reliance.[15] Control can be exercised in various ways, including through cognitive ''forcing'' functions that prompt users to request additional information in the form of explanations should they desire them.[50]

The static versus dynamic nature of task allocation is also important, because tasks in which control flexibly shifts between human and machine in accordance with user needs are better at sustaining operator vigilance.[47] HAI teams in which allocation is dynamic can be further divided between those in which the allocation is *adaptable*, where users dictate the allocation, and those in which the allocation is *adaptive*, where the allocation is automated.[47,58] Allocation can then proceed along several lines, but perhaps the most intuitive is along lines of difficulty. A human is likely to find some tasks easy that a machine will find hard and others hard that a machine will find easy. (From the machine's perspective, difficulty can be understood in terms of the degree of uncertainty exhibited in regard to a specific prediction.)[59] Generally, human trust in AI is higher when tasks involve objective calculation—to the point of trusting the AI even after seeing it make mistakes[60]—and lower when tasks involve social and emotional intelligence.[2] Both adaptive and adaptable forms of allocation can go some way toward achieving an optimal division of labor from the point of view of difficulty. For example, under adaptable allocation, humans can reserve all the tasks they consider easy for themselves and delegate the remaining ones to a machine. Under adaptive allocation, a machine could vary the difficulty of the tasks it reserved for the human, so that it referred to both moderately difficult as well as easy tasks to them, in an attempt to keep users vigilant (e.g., via so-called ''catch trials''). In one study, adaptable allocation was found to have a marginal advantage over adaptive allocation, and (unsurprisingly) happens to be easier to design.[58] However, adaptive systems may be able to leverage uncertainty information in ways that are more effective than adaptable systems (catch trials for one)[61] (see Box 2).

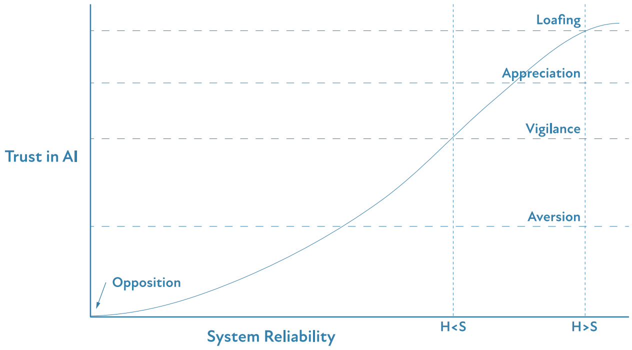

Figure 3. Trust versus reliability

Confidence information

A different form of transparency involves presenting users with system confidence information. There is growing evidence that suitably formatted confidence data (e.g., in the form of uncertainty estimates, confidence intervals, confidence levels, etc.) may improve trust calibration.[6,55,64] To the extent that humans have the capacity to incorporate an AI system's uncertainty appropriately, this will result in better performance. However, we highlight two significant challenges: (1) humans are often poor at handling numeric information, so presentation and design may be important.[69,70] Indeed, there is evidence that humans may be prone to ''information overload'' so that providing confidence measures might lead to worse performance.[37] (2) In fact, it is typically challenging to provide reliable, well-calibrated uncertainty estimates. An unfortunate property of current AI systems is that they are prone to being overconfident on examples where they might perform poorly.[71]

OPEN QUESTIONS

There are at least two important sets of issues whose resolution is outstanding. First, it is unclear what effects the size, frequency, type, and distribution of errors have in the loss and recovery of trust after users witness automation errors. Second, we know little about how different forms of transparency compare in the course of rebuilding that trust. In particular, almost nothing is known about how explanations, confidence data, performance metrics, and dynamic allocation strategies measure up against *each other* from the standpoint of optimal HAI team performance.

Error size, frequency, type, and distribution

Beyond common intuitions, little is known about the precise effects of an error's size on trust violation and repair. It is reasonable to suppose that an error's size need not refer to simply to its deviation from an ideal quantity or range, as in the case of risk scores that are off by some measure. By referring to an error's size one could equally well intend to convey, more generally, how *surprising* the error given widely held assumptions among

Box 2. Example of dynamic task allocation

Allocation strategies can help a human maintain algorithmic vigilance.[47] In an adaptable HAI team, the user dictates the allocation *a priori*.[58] This allows humans to select which tasks they want to outsource to machines. Humans may elect to keep easy tasks for themselves, leaving harder tasks for machines, or may instead keep the difficult tasks (e.g., tasks requiring the exercise of discretion), allowing machines to focus on rote tasks. In an adaptive HAI team, by contrast, the machine dynamically determines the allocation strategy.[58]

A large body of research in aviation demonstrates the potential advantages of adaptive allocation.[62] Air traffic controllers manage aircraft flow and intervene if aircraft separation is too low.[63] The controller is provided with an automated decision aid to handle multiple tasks. In these scenarios, an adaptive allocation strategy is usually preferred.[62,64] One advantage of adaptive strategies is that they can accommodate the use of ''catch trials.'' The point of a catch trial is to ensure that the controller is alert and situationally aware.[65,66] They may take the form of randomly generated system errors to ''catch out'' the user or (more commonly) abstentions in which the system declines to recommend a course of action in a specific instance, leaving the user to fall back on their own skills.

When both the human and machine find a task easy, it likely does not matter which agent provides a response (although decision fatigue is an ever-present risk).[67,68] More interesting are cases in which both machine and human struggle with a task. One approach here would be to select an agent at random. If the human is selected, then the human must make a decision without the machine's recommendation; if the machine is selected, then the human would be shown the machine's recommendation before making a decision (i.e., the human would have a choice whether to accept the machine's recommendation). Future work might explore the efficacy of similar tie-breaking strategies when machines and humans both struggle with the same tasks.

users about how the world ought to be. As we already suggested, mistakes on easy tasks (i.e., obvious mistakes) may be judged more harshly and be more corrosive of trust, than those on tasks perceived to be more difficult. We also noted evidence that continuous performance feedback may be an effective means of encouraging appropriate reliance after users witness automation errors. But it is not clear whether this kind of feedback is powerful enough to withstand the blow dealt to trust by the commission of large or obvious automation errors (e.g., Dzindolet et al.[5] found that such feedback is only effective when users are shielded from seeing obvious errors altogether). Again, beyond common intuitions, little can be said about the precise effects of the *frequency* of errors either. But, as one might expect, users do seem able to recover more readily from isolated or acute system failures than they do from chronic ones.[48,72]

Less still is known about the effects of distinct types of error on trust. Some studies purport to show that false alarms and misses affect trust differently, with false alarms having a greater negative impact than misses; while some report no significant difference along this dimension,[36] interpreting these conflicting results by suggesting that the *consequences* of false alarms versus misses determine the effects observed. In a contest between a false alarm that poses only a ''minor inconvenience'' (e.g., a trigger-happy smoke alarm) and a miss that could be lethal (a smoke alarm that operates intermittently), it is the former that will have less deleterious effects on trust than the latter. But as they note: ''the relative influence of *other* types of automation failures, such as breakdowns and error messages, has yet to be determined'' (our emphasis).

Perhaps least understood of all is the effect of the *distribution* of system errors over time. For example, are two large errors in quick succession as detrimental to trust as two large errors spaced apart (e.g., one at the beginning and one in the middle of a task session)? If so, are such ''clustered'' errors also more difficult to repair than temporally dispersed ones? We do not know. There are some indications that the earlier during a session that an error occurs, the sharper and more significant the decline in trust and the more difficult it will be to recover, despite reliable performance otherwise.[7] This makes sense—if an acquaintance betrays your trust very early on in your dealings with them, you may find it harder to ''forgive and forget'' a single infraction than if you had been friends for 20 years. Nonetheless, such adverse events can be beneficial too, inducing appropriate reliance (as against algorithm aversion). The studies by Manzey et al.,[7] for instance, revealed that participants exposed to automation failures earlier on in a task session were less susceptible to both automation complacency and automation bias. But beyond this we know little.

Comparative performance of transparency regimes

We already noted some of the drawbacks of AI-generated explanations in fostering well-calibrated user trust. Most notable among these is the risk of overtrust. What requires further investigation is whether the merits of various alternatives to explanations, on balance, make them more suitable than explanations. In particular, which forms of transparency are most effective in mitigating the risk of aversion and opposition after seeing an AI make a mistake? This latter question is more important than the question over which forms of transparency will best mitigate the risk of appreciation and loafing, because AI systems can be expected to improve over time, and perhaps radically. In that event, a trust surfeit arising from the use of explainable algorithms will not prove nearly as hazardous as a trust deficit arising from the use of alternative algorithms—at least in safety-critical domains. Hence the bar that any of the alternatives to explainable algorithms will have to meet may need to be set progressively higher, roughly in line with gains to system accuracy.

Be that as it may, model confidence data (e.g., uncertainty estimates) have been shown to be more helpful to users than explanations in at least one study.[55] In another, confidence data ''helped pilots make better decisions about task allocation and compliance with [system] recommendations and thus resulted in improved performance and safety.''[6] Even so, the precise experimental setup was limited to a restricted range of confidence levels (high, low, and variable) and a binary solution space (the presence of ice on the jet wing or jet tail). As the study's authors noted, more realistic experimental conditions are necessary before one is warranted in drawing firmer conclusions. Indeed, greater *comparative* investigation of the efficacy of confidence data and explanations—under as close to real-life scenarios as possible—is what is really needed.[55] The same goes for dynamic performance metrics displaying an AI's superior "running average" against its human counterpart/s. As we noted earlier in this section, whether continuous performance feedback of this sort mitigates aversive tendencies emerging after users witness large or obvious errors is not known. Allowing users to manipulate algorithmic outputs may be all that it takes to set the reverse of these tendencies in motion.[15] It is possible, too, that adaptive allocation paradigms, which exploit the full possibilities of model uncertainty, will prove more effective overall in promoting vigilance than adaptable allocation. But again, whether any of these paradigms are preferable to explanations and to what extent remains unclear.

Incorrect, deceptive, or misleading transparency

Recall our definition of transparency as any information provided about an AI system beyond its model outputs. While transparency is often beneficial, we briefly note several potential dangers.[73] Just as model outputs can be wrong, so too can additional transparency information. Since this information might be relied upon in making decisions, incorrect transparency can cause harm. Incorrect transparency might be unintentional[74] or could be deliberately deceptive.[75–77] Furthermore, even correct information might be misleading. In human communication, we often leave certain points unsaid, assuming our counterpart has background knowledge of the context. This creates the potential for information to be misleading if it is not carefully presented.[78] Hence, ideally, algorithmic transparency should satisfy what linguists would call *pragmatic* desiderata. However, these are not easy to measure or satisfy in practice and remain an important focus of machine learning research.

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

We have considered how various forms of algorithmic transparency may promote user vigilance. More broadly, however, we have sought to provide a practical framework for the study of HAI teams that (1) brings the same phenomena investigated by a variety of fields under a unified descriptive apparatus, (2) clarifies the scope of the technical systems that are the proper target of these investigations, and (3) identifies the overriding concern of these investigations with the maintenance of algorithmic vigilance. Our hope is that, by presenting the above research within this framework, we might inspire those who study HAI teams to seek to forge stronger connections despite the persistence of disciplinary boundaries (in practice if not in principle). At the moment, HAI team research is siloed. To take just one case, the authors of a recent (and high-quality) peer-reviewed study took themselves to be challenging "the widespread assertion that people are averse to algorithms" on the basis that the participants in their study "were quite willing to rely on algorithmic advice before seeing the algorithm err."[3] Human factors engineers would be unmoved by the finding that humans are prepared to trust—indeed overtrust—algorithms, having invested great efforts over the years in dealing with the problematic consequences of this very tendency. In our view, HAI team research should comprise a unified branch of study with a basic modus operandi and lingua franca, albeit drawing from expertise across several autonomous subfields. Our framework offers a pragmatic way forward.

Perhaps the greatest challenge in the study of HAI teams, however, is simply resisting the urge to overgeneralize experimental results.[47] Indeed, we think that ecological validity is an underappreciated problem in this area. Findings in aviation and shipping contexts are of questionable value in court and law enforcement contexts, which in turn may have little bearing on how the automation of medical diagnoses should be approached.[38] In legal and medical contexts, initial trust in automation is actually quite low, presumably due to the expertise of the users involved.[79,80,] This is at odds with the general findings we reviewed above.

Insofar as ecological validity *is* acknowledged, too often it features as an afterthought: a mere warning to readers of the limitations of the study concerned along with a reminder to keep those limitations in mind when applying results in real-world settings.[7] This is a good start, but it has not prevented occasionally sweeping claims being made about how "people" using "algorithms" react in this or that situation[3,5,14,15] (cf. Carton et al.[51]).To illustrate, we can take an otherwise excellent and justly influential study whose authors fell into this trap. At one point, the authors state their take-home message as follows: "observing an automated decision aid make errors leads to distrust of the automated decision aid, unless an explanation is provided explaining why the aid might err."[5] A little further down the same page (p. 715), however, one finds the customary discussion of limitations. First, they noted that "the task was very simple and artificial." Second, the study necessarily ignored "[t]he effect of one person's view of the automated aid's trustworthiness on other group members' reliance decisions," because the study limited itself to examining the dyad of a single user with an automated aid; and so on. When the findings of a branch of study are taken up with the vim and vigor typical in HAI team research, ecological concerns become too important to squeeze into general disclaimers. How can we be certain that the limitations do not vitiate the generalizations entirely? Ideally, authors should premise all substantive claims so that even such rudiments as titles and abstracts are expressed tentatively. In the illustration just given, the take-home message cannot quite be: people distrust automated aids whose errors they witness unless an explanation is provided. Something more tentative is called for: *in very simple automated tasks involving a single person*, people *tend* to distrust automated aids whose errors they witness, unless an explanation is provided. Every algorithm, every interface, every task, is unique after all.

Perhaps the most effective way to meet the ecological challenge is for HAI team research to proceed in a *task-specific* fashion that takes account of the precise nature of the task and its setting. Note that task specificity is distinct from domain specificity. Domain-specific investigation would confine research and its results to a more or less widely defined *domain* of activity (such as maritime shipping or criminal justice). Task-specific investigation, by contrast, would confine research by the nature of the task under consideration (such as adjudication between disputing parties, regardless of whether it is carried out by a court of law, a mediator, or a human resources officer). Since the basis of investigation and extrapolation in the latter case is the similarity of the tasks undertaken, regardless of domain, task-specific investigation may harness results from research conducted across what are in fact very distinct domains of activity (as the examples just given show). Conversely, a task-specific orientation may mean that results from one experiment are not presumed to generalize to another setting, despite the fact that both tasks occur within the same domain (e.g., results from an experiment testing the behavior of judges using recidivism risk algorithms in sentencing or bail applications may fail to generalize to a setting in which judges use algorithms to determine the likelihood of a repeat psychotic episode in a parent suing for child custody).

Our impression is that sweeping claims are more typical in the literature of organizational behavior and machine learning than they are in those of, say, ergonomics and human factors. The latter fields have always had several parallel streams of inquiry running alongside one another (e.g., one for ocean navigation, one for aviation and air traffic control, one for autonomous vehicles, another for nuclear power, etc.), and this has meant that conclusions in these fields have always been implicitly circumscribed. It is in the nature of task-specific research to constrain the applicability of results.

An emphasis on task-specific inquiry may seem in tension with our call for HAI team research to espouse greater cross-disciplinary cohesion and coordination. But what we are calling for in the latter case is simply an end to the kind of siloed research in which differences in terminology serve no purpose, and where people from one field are unaware of discoveries in another relating to the exact same subject matter. Cross-disciplinary activity, as such, is compatible with task-specific investigation: the field of human factors itself offers an excellent model of domain- *and* task-specific research worth emulating at a larger scale. This transition may not be easy to achieve. HAI team researchers, whose main experience is in machine learning, may find it especially difficult. The machine learning community on the whole values task-*independent*, model-agnostic and scalable, general models to solve as many variations of a problem as possible. This work is not misconceived. Indeed, there is a delicate balance to be struck between the necessity of controlled and (to a sometimes considerable extent) contrived experimental conditions on the one hand, and real-world applicability on the other. We appreciate that experimental conditions must strive to isolate the psychological processes underlying team behaviors, and that without a certain amount of artifice in experimental design there can be no generalizable results at all. However, our aim here is to direct attention to the importance of real-world applicability and, more specifically, to intra-ecological generalizability. We propose that task specificity is an effective means of securing this form of generalizability. Then, within a task-specific orientation, the familiar give-and-take between laboratory and real life can proceed in accordance with the principles of sound applied science. But task specificity is an imperative if the machine learning community is to contribute meaningfully to HAI team research.

ACKNOWLEDGMENTS

J.Z. is part-funded by the Leverhulme Trust (ECF-2020-428). U.B. acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Center for the Future of Intelligence (CFI) and from the Mozilla Foundation. A.W. acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. We are grateful to Simone Schnall and Reuben Binns for very helpful discussion and comments.

AUTHOR CONTRIBUTIONS

Conceptualization, J.Z., U.B., and A.W.; methodology, J.Z., U.B., and A.W.; investigation, J.Z. and U.B.; writing – original draft, J.Z., with U.B. and A.W. contributing parts to middle sections; writing – review & editing, J.Z.; Box 1, conceptualization and writing, J.Z.; Box 2, conceptualization and writing, U.B.; project administration, J.Z., U.B., and A.W.

REFERENCES

1. Lewandowsky, S., Mundy, M., and Tan, G. (2000). The dynamics of trust: comparing humans to automation. J. Exp. Psychol. Appl. *6*, 104.

2. Lee, M.K. (2018). Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc. *5*, 2053951718756684.

3. Logg, J.M., Minson, J.A., and Moore, D.A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. *151*, 90–103.

4. Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. Hum. Factors *39*, 230–253.

5. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., and Beck, H.P. (2003). The role of trust in automation reliance. Int. J. Human Comput. Stud. *58*, 697–718.

6. McGuirl, J.M., and Sarter, N.B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Hum. Factors *48*, 656–665.

7. Manzey, D., Reichenbach, J., and Onnasch, L. (2012). Human performance consequences of automated decision aids: the impact of degree of automation and system experience. J. Cogn. Eng. Decis. Making *6*, 57–87.

8. Bainbridge, W.A., Hart, J.W., Kim, E.S., and Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. Int. J. Soc. Robot. *3*, 41–52.

9. Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), pp. 251–258.

10. Gombolay, M.C., Gutierrez, R.A., Clarke, S.G., Sturla, G.F., and Shah, J.A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams. Aut. Robots *39*, 293–312.

11. Robinette, P., Howard, A.M., and Wagner, A.R. (2015). Timing is key for robot trust repair. In International Conference on Social Robotics (Springer), pp. 574–583.

12. Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), pp. 1–8.

13. Andrist, S., Bohus, D., Yu, Z., and Horvitz, E. (2016). Are you messing with me? Querying about the sincerity of interactions in the open world. In 2016 11 th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), pp. 409–410.

14. Dietvorst, B.J., Simmons, J.P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. *144*, 114.

15. Dietvorst, B.J., Simmons, J.P., and Massey, C. (2018). Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manag. Sci. *64*, 1155–1170.

16. Montague, E., and Xu, J. (2012). Understanding active and passive users: the effects of an active user using normal, hard and unreliable technologies on user assessment of trust in technology and co-user. Appl. Ergon. *43*, 702–712.

17. Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). 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.

18. Schmidt, P., Biessmann, F., and Teubner, T. (2020). Transparency and trust in artificial intelligence systems. J. Decis. Syst. *29*, 260–278.

19. De-Arteaga, M., Fogliato, R., and Chouldechova, A. (2020). A case for humans-in- the-loop: decisions in the presence of erroneous algorithmic scores. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12.

20. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., et al. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–13.

21. Yang, F., Huang, Z., Scholtz, J., and Arendt, D.L. (2020). 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.

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

23. Weerts, H.J., van Ipenburg, W., and Pechenizkiy, M. (2019). A human-grounded evaluation of shap for alert processing. In Proceedings of KDD Workshop on Explainable AI.

24. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14.

25. Lee, J.D., and See, K.A. (2004). Trust in automation: designing for appropriate reliance. Hum. Factors *46*, 50–80.

26. Rousseau, D.M., Sitkin, S.B., Burt, R.S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. Acad. Manag. Rev. *23*, 393–404.

27. Siegrist, M., Earle, T.C., and Gutscher, H. (2003). Test of a trust and confidence model in the applied context of electromagnetic field (EMF) risks. Risk Anal. Int. J. *23*, 705–716.

28. Siegrist, M., Gutscher, H., and Earle, T.C. (2005). Perception of risk: the influence of general trust, and general confidence. J. Risk Res. *8*, 145–156.

29. Epley, N., Waytz, A., and Cacioppo, J.T. (2007). On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. *114*, 864.

30. Evans, A.M., and Krueger, J.I. (2009). The psychology (and economics) of trust. Social Personal. Psychol. Compass *3*, 1003–1017.

31. Thielmann, I., and Hilbig, B.E. (2015). Trust: an integrative review from a person- situation perspective. Rev. Gen. Psychol. *19*, 249–277.

32. Lewicki, R.J., and Brinsfield, C. (2017). Trust repair. Annu. Rev. Organ. Psychol. Organ. Behav. *4*, 287–313.

33. Fiske, S.T. (2018). Stereotype content: warmth and competence endure. Curr. Dir. Psychol. Sci. *27*, 67–73.

34. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M., and Eckersley, P. (2020). Explainable machine learning in deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648–657.

35. Glikson, E., and Woolley, A.W. (2020). Human trust in artificial intelligence: review of empirical research. Acad. Manag. Ann. *14*, 627–660.

36. Hoff, K.A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors *57*, 407–434.

37. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16.

38. Goddard, K., Roudsari, A., and Wyatt, J.C. (2014). Automation bias: empirical results assessing influencing factors. Int. J. Med. Inform. *83*, 368–375.

39. Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2019). Algorithmic decision-making and the control problem. Minds Mach. *29*, 555–578.

40. Parasuraman, R., and Manzey, D.H. (2010). Complacency and bias in human use of automation: an attentional integration. Hum. Factors *52*, 381–410.

41. Pazouki, K., Forbes, N., Norman, R.A., and Woodward, M.D. (2018). Investigation on the impact of human-automation interaction in maritime operations. Ocean Eng. *153*, 297–304.

42. Bagheri, N., and Jamieson, G.A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency". Hum. Perform. Situat. Aware. Autom. Curr. Res. Trends *1*, 54–59.

43. Banks, V.A., Eriksson, A., O'Donoghue, J., and Stanton, N.A. (2018). Is partially automated driving a bad idea? Observations from an on-road study. Appl. Ergon. *68*, 138–145.

44. Banks, V.A., Plant, K.L., and Stanton, N.A. (2018). Driver error or designer error: using the perceptual cycle model to explore the circumstances surrounding the fatal Tesla crash on 7th may 2016. Saf. Sci. *108*, 278–285.

45. Lee, J.D., and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. Int. J. Human Comput. Stud. *40*, 153–184.

46. Dietvorst, B.J., and Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol. Sci. *31*, 1302–1314.

47. Chavaillaz, A., Wastell, D., and Sauer, J. (2016). System reliability, performance and trust in adaptable automation. Appl. Ergon. *52*, 333–342.

48. Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics *35*, 1243–1270.

49. Lai, V., and Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 29–38.

50. Bu̧cinca, Z., Malaya, M.B., and Gajos, K.Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Human Comput. Interact. *5*, 1–21.

51. Carton, S., Mei, Q., and Resnick, P. (2020). Feature-based explanations don't help people detect misclassifications of online toxicity. In Proceedings of the International AAAI Conference on Web and Social Media, *14*, pp. 95–106.

52. Shen, H., and Huang, T.-H. (2020). How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, *8*, pp. 168–172.

53. Kenny, E.M., Ford, C., Quinn, M., and Keane, M.T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif. Intell. *294*, 103459.

54. Jeyakumar, J.V., Noor, J., Cheng, Y.-H., Garcia, L., and Srivastava, M. (2020). How can I explain this to you? An empirical study of deep neural network explanation methods. Adv. Neural Inf. Process. Syst. *33*, 4211–4222.

55. van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating XAI: a comparison of rule-based and example-based explanations. Artif. Intell. *291*, 103404.

56. Wang, L., Jamieson, G.A., and Hollands, J.G. (2009). Trust and reliance on an automated combat identification system. Hum. Factors *51*, 281–291.

57. Seong, Y., and Bisantz, A.M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. Int. J. Ind. Ergon. *38*, 608–625.

58. Sauer, J., Kao, C.-S., and Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. Ergonomics *55*, 840–853.

59. Bhatt, U., Antoran, J., Zhang, Y., Liao, Q.V., Sattigeri, P., Fogliato, R., Melancon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. (2021). Uncertainty as a form of transparency: measuring, communicating, and using uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21, Association for Computing Machinery), pp. 401–413.

60. Dijkstra, J.J. (1999). User agreement with incorrect expert system advice. Behav. Inf. Technol. *18*, 399–411.

61. De, A., Okati, N., Zarezade, A., and Rodriguez, M.G. (2021). Classification under human assistance. In Proceedings of the AAAI Conference on Artificial Intelligence, *35*, pp. 5905–5913.

62. Parasuraman, R., Mouloua, M., and Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. Hum. Factors *38*, 665–679.

63. Metzger, U., and Parasuraman, R. (2005). Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. Hum. Factors *47*, 35–49.

64. Papenmeier, A., Englebienne, G., and Seifert, C. (2019). How Model Accuracy and Explanation Fidelity Influence User Trust (IJCAI Workshop on Explainable Artificial Intelligence).

65. Davies, D.R., and Parasuraman, R. (1982). The Psychology of Vigilance (Academic Press).

66. Gugerty, L.J., and Tirre, W.C. (2000). Individual differences in situation awareness. Situat. Aware. Anal. Meas. 249–276.

67. Chaparro, A., Groff, L., Tabor, K., Sifrit, K., and Gugerty, L.J. (1999). Maintainingsituational awareness: the role of visual attention. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, *volume 43* (SAGE Publications Sage CA: Los Angeles, CA), pp. 1343–1347.

68. Warm, J.S., Dember, W.N., and Hancock, P.A. (1996). Vigilance and workload in automated systems. In Automation and Human Performance: Theory and Applications, R. Parasuraman and M. Mouloua, eds. (Lawrence Erlbaum Associates, Inc), pp. 183–200.

69. Reyna, V.F., and Brainerd, C.J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. Learn. Indiv Differ *18*, 89–107. https://linkinghub.elsevier.com/retrieve/pii/S1041608007000428.

70. Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. Science *333*, 1393–1400. https://www.sciencemag.org/lookup/doi/10.1126/science.1191181.

71. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q. (2017). On calibration of modern neural networks. In International Conference on Machine Learning, pp. 1321–1330.

72. Biros, D.P., Daly, M., and Gunsch, G. (2004). The influence of task load and automation trust on deception detection. Group Decis. Negot. *13*, 173–189.

73. Weller, A. (2019). Transparency: motivations and challenges. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (Springer), pp. 23–40.

74. Ehsan, U., and Riedl, M.O. (2021). Explainability pitfalls: beyond dark patterns in explainable AI. https://arxiv.org/abs/2109.12480.

75. Heo, J., Joo, S., and Moon, T. (2019). Fooling neural network interpretations via adversarial model manipulation. Adv. Neural Inf. Process. Syst. *32*, 2925–2936.

76. Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. (2020). You shouldn't trust me: learning models which conceal unfairness from multiple explanation methods. In Proceedings of the 2020 European Conference on AI.

77. Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics and Society, pp. 180–186.

78. Gigerenzer, G., Wegwarth, O., and Feufel, M. (2010). Misleading Communication of Risk.

79. Linkov, F., Sanei-Moghaddam, A., Edwards, R.P., Lounder, P.J., Ismail, N., Goughnour, S.L., Kang, C., Mansuria, S.M., and Comerci, J.T. (2017). Implementation of hysterectomy pathway: impact on complications. Women's Health Issues *27*, 493–498.

80. Christin, A. (2017). Algorithms in practice: comparing web journalism and criminal justice. Big Data Soc. *4*, 2053951717718855.