A Framework for Designing Explanations for AI-Assisted Decision Making

ANONYMOUS AUTHOR(S)

In this paper, I conduct a systemic literature review of existing research on how AI explanations can impact AI-assisted decision-making across different contexts. In order to understand the varied design of explanations presented in the literature, I formulate a 2-D map to classify these explanations based on the aspect of the decision that is made *salient in the explanation* and the *scope of the explanation*. Along the first dimension, explanations may be more *process-centric* or more *outcome-centric*, depending on which aspect of the decision is more salient in the explanation. Along the second dimension, explanations may be valid *locally* or *globally*, depending on the scope of information that is provided. Additionally, I also define a set of design levers to control when the explanations are presented, how selective the explanation is, and how many decision outcomes are presented in the explanation. Finally, using this framework, I present a set of claims for how explanation design can impact some of the goals relevant to AI-assisted decision-making, while also identifying gaps for future research.

ACM Reference Format:
Anonymous Author(s). 2018. A Framework for Designing Explanations for AI-Assisted Decision Making. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advances in artificial intelligence (AI) have led to its widespread adoption in decision-making contexts [10]. In particular, AI models are being increasingly used to assist experts and authorities in their decision-making, ranging from aiding judges with pre-sentencing trials [11] to helping doctors with medical diagnoses [7]. More recently, explanation mechanisms have begun to play a major role in improving various aspects of this human-AI decision-making.

As a result, research has been done to design and test many different explanations types in various contexts [1, 2, 7, 15, 23, 27, 28]. For example, studies have experimented with explanations that provide more information about the AI model's process, such as revealing the set of features and their importance as determined by the model [2], providing a relevant example from the model's training data [6], or simply revealing the model's error rates and accuracy [15]. In this paper, I aim to present a systemic review of existing literature on how explanations can impact AI-assisted decision-making across different contexts.

Based on my review, I formulate a 2-D map to classify different types of explanations identified from the prior literature. Based on the idea that the act of 'explanation' is to provide some information about a phenomenon [17, 18], my 2-D map classifies explanations based on the kind of informational support they provide. In particular, I characterize explanations based on two factors - 1) the aspect of the decision that is made *salient in the explanation* where explanations may be more *process-centric* or more *outcome-centric*, and 2) the *scope of the explanation* where explanations may provide information that is valid *locally* or *globally*.

Through my analysis, I also reveal how the same type of explanation may be presented in different ways, for instance by varying the number of decision outcomes that are presented in an explanation. To account for these variations, I

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
Manuscript submitted to ACM

define a set of *design levers* to control the design of a particular type of explanation based on – 1) when the explanation (along with the outcome) is presented, 2) how many outcomes are presented in the explanation, and 3) how selective the explanation is.

Finally, I use this framework to analyze how different explanation types and design levers can impact the various goals that may be relevant to AI-assisted decision-making. In particular, I pick a set of generic goals that may be valid across different contexts such as the overall human-AI team performance and the extent of human over-reliance on AI. When presenting my analysis, I refer to the human as the decision maker to make their role explicit. To assist domain experts and practitioners in scoping out the role of explanations for their particular context, I also present a set of design claims that underline the impact of an explanation design on a particular outcome (e.g., human over-reliance on AI). The framework remains a work in progress as I continue to refine my analysis and include some of the more recent work (since late 2022) on designing AI explanations. However, within the context of this workshop, I believe that my analysis brings forward important questions relevant to the current state of research and provides direction for future research on AI explanations and their role in AI-assisted decision-making.

The rest of the paper is organized as follows – in Section 3 and 4, I define my 2-dimensional map and explain the 3 design levers. In Section 5, I describe a set of goals relevant to decision-makers and analyze how different design decisions can impact those goals. In Section 6, I underline some limitations of my framework as well as challenges for future work.

2 LITERATURE REVIEW METHODOLOGY

To begin my literature review, I collected a seed list of research articles that either argued for a new explanation design or presented an empirical evaluation of explanation(s). In particular, I picked the seed list of articles from two graduate level course on human-AI interaction [1]. From this list of articles, I conducted two rounds of citational analysis to collect additional articles that were either cited by one of the seed articles or cited the seed article. The articles were collected between January and February 2022, so articles published after that period have not been covered as part of the analysis. In total, I analyzed 19 research articles to extract different explanation designs and/or their empirical evaluation. All articles that were analyzed are part of the reference list.

3 EXPLANATION TYPES: A 2-DIMENSIONAL MAP

In this section, I present my 2-dimensional map that helps categorize the different types of explanations presented in the literature. Scholars define 'explanation' as the act of providing information about a phenomenon [17, 18], in this case, the decision made by the AI. Aligning with that definition, I categorize the different explanations according to the type of information they provide. The first dimension represents the aspect of the decision that is *salient in the explanation*, where explanations may be *process-centric* or *outcome-centric*. One way to explain how a decision was made is to provide more information about the process that led to the decision. For example, this may include revealing the internal features of the AI model and their importance learned by the model. I consider these explanations *process-centric*, as they make the process behind the decision more salient. Alternatively, an explanation may reveal additional details about the outcome produced by the AI model. For example, revealing the distribution of different outcomes based on particular aspects of the input. Since the process and the outcome are inherently related, process-centric explanations will still reveal some information about the outcome, and vice versa. Therefore, it helps to visualize the dimension

[1] https://haicourse.ischool.utexas.edu/Boutline/, https://www.cs.purdue.edu/academic-programs/courses/VT%20course%20syllabi/Fall%202021/CS%2059200-HAI%20F21%20Tianyi%20Zhang.pdf

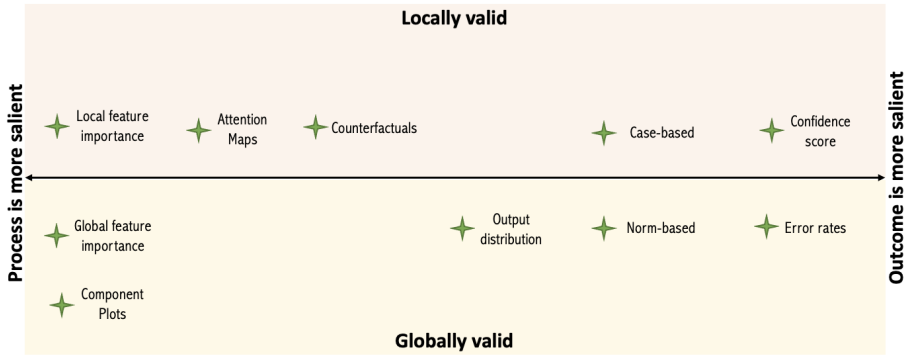

Fig. 1. A visual representation of the 2-Dimensional map, characterizing different explanations according to the type of informational support they provide.

as a continuum, where different explanations may be considered more process-centric (less outcome-centric) or less process-centric (more outcome-centric). Figure **??** maps the different explanation types along this dimension, where explanations become more outcome-centric (less process-centric) as you go from left to right.

The vertical axis in the Figure represents the second dimension along which explanations may be categorized, based on the *validity of the explanation*, where the information provided in the explanation may be valid *locally* (top in Figure 1) or *globally* (bottom). Local explanations will provide information that is valid only for a particular instance of data (or a small region around that data), and global explanations would be valid over a large range of data, and sometimes, the entire dataset. This binary distinction – also mentioned in prior literature [] – highlights the fact that information provided in a local explanation will generally not be applicable to other data points. On the contrary, global explanations will be broadly applicable and more consistent across different data points.

4 DESIGN LEVERS FOR EXPLANATIONS

As noted in the previous section, explanation types are often designed differently across different studies. In this section, I present *3 design levers* to explain and synthesize some of the variations within an explanation type that have been tested (or proposed) in prior literature. The 3 levers include varying the timings of the explanation, the outcome(s) presented in the explanation, and selectivity within the explanation. Table 1 further shows the default configuration for these design levers as well as some other configurations that have been proposed in prior literature. I now describe each of them below:

4.1 Timings of the Explanation

Generally, in a decision-making context, the explanation (along with any outcome(s)) is made available upfront to the human decision-maker. However, it may be possible to vary that timing (for instance, delay it until the human makes their decision). Bucinca et al. [5] experiment with the timings of the explanation as part of their cognitive forcing functions (CFF) to counter quick heuristic-based decision-making on the part of the humans [26]. In particular, they experiment with three different configurations – i) forcing the human to make their own decision before the AI; ii) delaying the AI's decision by a fixed amount of time (30 secs); or, iii) letting the human control when they want to see the AI's decision. The general idea of varying the timings of an explanation is also proposed by Bansal et al. [1] who

Design Lever	Default configuration	Other Examples
Timings of the explanation [1, 5, 29, 31]	upfront with the decision	Until after the human makes the decision; fixed delay (30 secs); Letting human control
Outcome(s) in the explanation [1, 6, 8, 27, 29]	Top outcome predicted by the AI	No outcome; Explain-top-2 (top 2 outcomes predicted); Adaptive (switch between top-1 and top-2); Devil's advocate (outcome predicted by the human)
Selectivity within explanation [8, 16, 18, 19, 30]	Static selection	Partial vs complete; Varying the length; Interactive selection

Table 1. Design levers and their configurations

noted how providing the explanation and the outcome upfront in their experiment was making it "almost impossible for the human to reason independently" [1]. Consequently, certain studies [29, 31] make it implicit in their design where humans are always asked to make their own prediction before being presented with the AI's prediction. And then they are given a choice to update the final prediction.

4.2 Outcome(s) present in the Explanation

The design lever is concerned with choosing the outcome(s) that are presented in the explanation. Generally, the default choice is to explain the top outcome predicted by the AI. Deviating from the default, Carton et al. [8] experimented with a design where the explanation was presented without revealing the AI's outcome. Alternatively, Bansal et al. [1] also experimented with the following 3 configurations – i) (default) only explaining the top outcome of the AI model (explain-top-1), ii) explaining the top 2 outcomes of the model (explain-top-2), and iii) an adaptive design that switched between explain-top-1 and explain-top-2 based on the AI's confidence value.

Specific to example-based explanations, Cai et al. [6] experimented with adding a comparative component to their explanations by providing examples for the top-3 outcome classes predicted by the AI. Wang and Yin [29] also experimented with case-based explanations and provided two most similar cases as part of the explanation, instead of one (default setting). Out of the two cases that were presented, one had the same outcome as the current input while the other one had a different outcome.

Other than varying whether and how many of the top outcomes are present in the explanation, the explanation may be based on a completely different outcome, such as an outcome provided by the user (analogous to the AI playing a devil's advocate role [1]). More generally, the stakeholders of public-sector decision-making have argued for presenting multiple AI outcomes, to bring more discretion in the decision-making process [27].

4.3 Selectivity within the Explanation

For many explanations, choosing the explanation type further provides scope for making the explanation more or less selective. This is in general true for all explanations involving some representation of the input features (e.g., feature importance, counterfactuals, outcome prevalence). For instance, for feature-importance-based explanations, the choice may range from presenting all the features to presenting only one or some of them. Similarly, in the case of a counterfactual, the choice concerns presenting one or more of the many changes possible in the input features.

In most cases, the choice can be broken down into two factors – *how* selective the explanation is, and, *what* gets selected. The rationale behind the first factor is clear – reduce long explanations to a cognitively manageable chunk, so as to avoid information overload [19]. As Weld and Bansal [30] note in their paper, explanations should provide "as much information as is needed, and no more". To that end, Lage et al. [16] have experimented with 3 different controls to manipulate the extent of selectivity within an explanation – i) introducing more concepts in the explanation, ii) varying the overall length of the explanation, and iii) controlling the repetition of terms in an explanation. Carton et al. [8] also experimented with a partial explanation and a full explanation, where the partial explanation only highlighted a single chunk of text that was most predictable of the outcome, whereas the full explanation highlighted all chunks of text that contributed to the model's outcome.

However, the notion of what *should* be selected is not always clear. One obvious answer is to simply select those features that have the most influence on the output; for e.g., selecting the top N features according to their importance, where N will be determined by cognitive constraints. Alternatively, researchers have argued that given how human explanations work, factors other than highlighting the most probable cause could be important as well [18, 19]. Drawing a parallel to human explanations, Miller [18] presented a series of factors that should influence the selection process. For instance, information that helps contrast with a likely but different outcome (i.e., a foil) should be selected. Abnormality should be another criterion, i.e., aspects that are unusual should be selected. It suffices to say that in order to align AI explanations with human explanations, the selection should be dynamic and context-dependent, taking into account the humans' expectations and their mental models [19, 30]. While techniques have been proposed to implicitly provide more selectivity in explanation (e.g., approximating a complex deep neural network via simple linear features [22]), little research has been done to provide further control over what gets selected and when. Doing so may require explicit modeling of what the human expects or design mechanisms that provide scope for gathering that information in real-time as the explanation is provided. While researchers have argued for *interactive explanations* and, interactivity can afford a dynamic selection of information, more research is needed to understand how these interactions may impact the overall process.

5 DESIGN SPACE FOR GOALS RELEVANT TO DECISION MAKERS

5.1 Performance

This is one of the fundamental goals aimed at measuring the performance impact of humans collaborating with AI to make the final decision. The metric is generally applicable across different domains, ranging from medical diagnosis [7], to annotating review sentiments [1] to machine translation [13]. In contexts where there is a possibility of a correct decision (such as medical diagnosis), performance may be measured in terms of the accuracy of the decision. Alternatively, in contexts where many different correct decisions are possible (such as machine translation), performance may be measured in terms of the quality of the output.

Ideally, the goal should be for the human-AI team to perform better than either human or AI alone. However, Bansal et al. [1] show that in many studies, AI alone performance is in fact better than the performance of human-AI collaboration, arguably since AI models are designed to optimize performance on a specific metric. Yet, as Veale et al. find [27], in many decision-making contexts (e.g., content moderation [14]), humans still have an important role to play in accounting for external factors and information that is not available in the model. Therefore, a more appropriate goal, one that is indeed reported in many studies [1, 5, 11, 21], is to improve human-AI team performance relative to the performance of humans alone. To that end, studies across different contexts (e.g., food nutrition, labeling toxicity,

LSAT questions, real estate pricing) have shown that using AI models with any explanation leads to better performance compared to humans alone.

Claim: Any explanation design will lead to an improved human-AI performance relative to a human-only baseline

However, researchers [1, 24, 31] have also argued that the extent to which the explanations may impact the performance will depend on the scope of complementary performance. That is, the explanations will have the most impact if there are regions where the AI is usually correct but the human is more likely to make mistakes, and vice versa.

Claim: The impact of the explanation will be limited by the scope of complementary performance between human and AI.

Alternatively, if AI-alone consistently outperforms humans on the task [11, 21], human-AI performance can be effectively improved (relative to humans-alone) by providing explanations that make humans indiscriminately agree with the AI. However, that can result in humans agreeing with AI even when the AI is incorrect – resulting in human over-reliance on AI. Therefore, in addition to human-AI performance, it becomes important to also track human over-reliance on AI.

5.2 Over-reliance

Broadly, the philosophy behind this goal is to allow humans to establish the right amount of trust in AI, and more importantly, trust the AI in the right situations. Simply put, over-reliance can be tracked as the extent of agreement between humans and AI, in cases where AI is incorrect. More generally, most studies [1, 5, 31] track both human-AI performance and human over-reliance on AI, and as a conservative goal, establish whether certain explanation designs can reduce human over-reliance on AI without adversely affecting the team performance.

In terms of investigating *why* humans over-rely on AI, most studies [1, 5, 28] have attributed the over-reliance to cognitive biases, particularly, anchoring bias and confirmation bias. If that is indeed true, it seems plausible that simply presenting the AI's outcome (as part of the explanation) before the human gets a chance to make their own decision will potentially bias the human and result in some over-reliance. This is confirmed by Carton et al.'s [8] study who found that human-AI agreement was significantly higher in all conditions where the AI's outcome was presented upfront, i.e., before the human had a chance to make their own decision – irrespective of whether the AI was correct or not, and whether an explanation was provided or not. Similarly, Wang et al. [28] had also noted that medical experts in their study – while interacting with different explanations – would avoid looking at the AI's outcome until they had formed a hypothesis of their own, in order to avoid confirmation bias [28]. Bansal et al.'s [1] analysis of how humans used AI during their study also showed that using "AI as a prior guide" was the most common approach. Based on these findings, I present my next claim:

Claim: Presenting the AI's outcome upfront with an explanation will lead to human over-reliance on AI.

Perhaps one way to counter the anchoring effect due to providing the explanation upfront is to simply delay the explanation (with the outcome) in order to provide the human with an opportunity to make their own independent decision. Once the AI's outcome and the explanation are shared, the humans can be provided with another opportunity to *update* the final decision. In that regard, Bucinca et al. [5] introduced a series of cognitive forcing functions (CFF) that introduced a delay in the timings of the explanation to disrupt the human's quick, heuristic-based decision-making process that is susceptible to biases [26]. The results from their study show that delaying the AI's explanation can reduce human over-reliance on AI without affecting the overall performance (relative to the default case of presenting explanations upfront). Furthermore, studies [29, 31] also show that when humans are asked to make their own decision

before being shown the AI's outcome, whether and which explanation was provided (either process-centric or outcome-centric) had no significant impact on the extent of human over-reliance on AI. However, specific to the case of confidence scores, Zhang et al. [31] found that in cases where the AI was highly confident (> 80%), sharing the AI's confidence made humans more likely to agree with it. The authors do note that for such cases, the base rate of disagreement between humans and AI was already very low (< 10%). More generally, these findings lead to the following claim:

Claim: Delaying the AI's explanation (with the outcome) will reduce human over-reliance on AI, relative to providing them upfront.

It is worth noting that the design choice of delaying the AI's explanation also takes away any informational support that could be provided to the human while making the decision. Therefore, it is worth investigating whether providing explanations upfront without explicitly providing a clear outcome also has a similar effect on human over-reliance on AI. In particular, Carton et al [8] who experimented with providing explanations (in the form of attention maps) without any explicit outcome found that when explanations were provided upfront, humans were less likely to make the mistake of incorrectly assigning a positive outcome, compared to both human and AI alone [2]. That is, their results indicate that providing explanations upfront without a clear outcome can lead to improvements compared to humans-alone, without causing the same extent of human over-reliance on AI.

In a separate study, Bansal et al. [1] also experimented with providing multiple outcomes as part of the explanation. Possibly, this can encourage humans to compare and contrast different outcomes rather than biasing them toward one outcome. More generally, Veale et al. [27] had also noted that AI models deployed in public-sector were sometimes designed to provide a list of outcomes (such as top-3, instead of just the top outcome) to bring more discretion into the decision-making process and to encourage the human decision-maker to decide what makes the most sense to them. The result from Bansal et al.'s study suggests that providing multiple outcomes reduced the human's tendency to agree with the AI when it was uncertain and more likely to be incorrect, relative to the condition when only one outcome (AI's top prediction) was provided.

In other words, in the absence of a clear AI outcome, it is possible to provide additional informational support to humans, without causing the adverse effects of human over-reliance on AI.

Claim: Presenting explanations upfront without a clear AI outcome will also reduce human over-reliance on AI.

Note that in certain contexts where scalability is a major concern (e.g., content moderation), many decisions are made independently by the AI, and humans are only involved in some of the cases. Furthermore, it is possible that the human is implicitly or explicitly aware of the AI's outcome, and therefore, not presenting a clear AI outcome may either not produce the desired results, or simply be infeasible.

Therefore, specific to the case when the AI's outcome is presented upfront, it is worth investigating whether different explanations can have a differential impact on the extent of human over-reliance on AI. To that end, all studies on over-reliance where the outcome was presented upfront [1, 5, 8, 12] have used more process-centric explanations, either feature importance or attention maps. In fact, there is reason to believe that process-centric explanations may not be ideal in this case. Research [25] suggests that inducing process-oriented thinking in the presence of an outcome can make the decision process more difficult for humans and result in lower decision confidence. The effect was attributed to the fact that process-oriented thinking required humans to weigh more factors (i.e., both the process and the outcome) relative to outcome-centric thinking.

[2]Specific to their design, explanations could only be provided when the AI model predicted one of the positive outcomes. Therefore, the same effect could not be observed when the AI predicted one of the negative outcomes

Furthermore, evidence exists that even confidence scores – which are mostly used as a baseline comparison [1, 23] – can be effective at making the humans suspect the AI more. However, they do little to guide humans to the right decision, and humans still have to identify cases when the AI is incorrect with high confidence [1].

More generally, for outcome-centric explanations to be effective, it is still important for them to provide information that humans can meaningfully combine with their own prior knowledge [17, 18]. Therefore, I consider this a gap in the current literature:

Gap: When the AI's outcome is presented upfront, can outcome-centric explanations reduce human over-reliance on AI, relative to process-centric explanations?

5.3 Efficiency

In addition to the performance of the human-AI collaborative process, it may also be important to track the efficiency of the process, especially as new interventions are introduced into the process. Efficiency – measured in terms of time (or another resource) required to complete a given set of tasks – can be particularly important in domains such as, content moderation and machine translation [13], where efficiency along with scalability are the main goals driving the use of AI.

While it is always desirable to make the process more efficient, certain interventions may present a trade-off between efficiency and other goals (such as performance or over-reliance). In particular, design choices explicitly aimed at disrupting the human's quick heuristic-based processing – specifically, delaying the timings of the explanation, or not presenting a clear outcome – can have implications for the efficiency of the process.

Specific to varying the outcome(s) present in the explanation, Cai et al [6] found that humans spent more time analyzing explanations that contained normative examples from the top-3 predicted classes, compared to explanations that contained examples only from the top predicted class. However, Carton et al. [8] found that the mean amount of time spent on a task did not vary significantly based on whether the explanations contained the AI's outcome or not. It seems plausible that the loss of efficiency observed by Cai et al. is due to the extra information that is provided to explain the multiple outcomes. This leads to my next design claim:

Claim: Presenting multiple outcomes in the explanation will reduce the efficiency of the human-AI decision-making process

None of the studies [5, 29, 31] that delayed the explanation timings have reported on the efficiency of the decision-making process. Although it does seem intuitive that delaying explanations will require more time to complete the task. Particularly, humans first have to make their own decision (equivalent to human-alone), and then also process the additional information provided as part of the explanation. It is still worth finding out to what extent that impact the overall efficiency of the process:

Gap: To what extent does delaying the timings of the explanation impact the efficiency of the human-AI decision-making process?

Furthermore, the efficiency of the process can also be impacted by controlling for selectivity within the explanation. Lage et al. [16] found that humans' response time on the task increased significantly when the explanations defined more concepts, or when explanations were simply longer in text. This leads to my next design claim:

Claim 8: Making explanations more selective will increase the efficiency of the process.

5.4 Subjective Perceptions

So far, I have focused on measures that capture the actual behavior of humans interacting with AI in decision-making contexts. In addition, prior studies have also collected various subjective measures to capture human perceptions (or attitudes) about AI. In fact, Bucinca et al. [4] found that the humans' actual behavior (for instance, measured in terms of overall performance) did not correlate with their subjective perceptions about the AI. Specific to decision-making in the public sector, Veale et al. [27] further noted that despite high measures of performance, many AI models do not get adopted or deployed in real settings. Therefore, it may be worthwhile to collect and account for both, measures of actual behavior as well as human subjective perceptions about AI.

In terms of subjective measures, different studies have focused on different measures, such as the perceived utility of AI [5], expectations about the AI [15], subjective understanding of the AI [29], etc. While it is unclear how different subjective measures are correlated with each other, most studies show that providing any explanation generally leads to better subjective perceptions about the AI [1, 5, 6, 15, 28]. Based on that, I present my next design claim:

Claim: Any explanation design will lead to better subjective perceptions about AI among humans.

Amongst the different measures, the perceived utility has been used most widely. Most studies use a single Likert scale response [1, 7, 13] to measure the utility of the AI as perceived by the human decision maker. In addition, the perceived utility may also be inferred by measuring the workload experienced by the human decision-maker. The workload can be measured using standard scales, such as the NASA TLX[3] (see [7]), or through a single Likert scale measure of the mental demand experienced by the human decision-maker [5].

Since explanations are generally considered helpful, it naturally follows that delaying the explanation will decrease the perceived utility of the AI. Yet, Bucinca et al. [5] found no significant difference in the subjective perceptions of workload experienced by humans assigned to different conditions, where explanations were delayed in some of the conditions. Furthermore, in terms of the outcome(s) presented in the explanation, Bansal et al. [1] found no significant differences in the perceived utility of AI in terms of presenting top-1 or top-2 outcomes. At the same time, their adaptive design which switched between explain-top-1 and explain-top-2 based on the confidence score of the AI was generally considered more useful than either of the two designs.

It is possible that any differences in subjective perceptions may only appear when the same human interacts with different designs (i.e., a within-subject comparison). Indeed, in a different context, Binns et al. [2] found that humans' subjective perceptions about different explanations varied in the within-subject study, while no differences were found in the between-subject study. Note that in the real world, a comparison between different explanations may be inevitable and therefore, a within-subject comparison may provide more realistic results [2]. I consider this as a gap in the current literature:

Gap: In a within-subject comparison, does varying the timings of the explanation or the outcome(s) present in the explanation result in different perceptions about the utility of the AI?

Furthermore, it can be expected that humans' subjective perceptions about AI may depend on the type of information that is presented in the explanation. Comparing local and global explanations, there is some evidence that at the time of decision-making, humans may prefer locally valid explanations over global explanations. In a study with medical experts, Wang et al. [28] found that their participants did not like global explanations (particularly, global feature importance and dependency plots) as they masked the relations and inter-dependency (colinearity, intersectionality) between different features, limiting their overall utility.

[3] https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf

	Process-centric (vs outcome-centric)	Locally valid (vs globally valid)	Timings of the explanation	Outcome(s) in the explanation	Selectivity within explanation
Performance	~ [1, 29]	~ [29]	~[5, 31]	~[1, 8]	~[8, 16]
Over-reliance	?? [1, 3, 12, 25]		– [5, 29, 31]	– [1, 8, 27]	~ [8]
Efficiency			?? [5, 29, 31]	– [6, 8]	++ [8, 16]
Subjective preferences		+- [15, 28]	?? [5]	?? [1]	

Table 2. Summary of the Design Space for different goals relevant to decision-makers. ++ represents an increase; – represents a decrease; +- represents mixed effects; ?? represents a gap; other than the case of over-reliance, ++ is generally better

Claim: At the time of decision-making, humans will find local explanations to be more useful than global explanations

However, global explanations may have their own role to play. Specifically, Kocielnik et al. [15] found that global explanations (feature importance and error rates) were helpful in setting the right expectations toward the AI, and correspondingly resulted in increased *satisfaction* and *acceptance* for the AI.

Claim: Global explanations can be helpful in setting the right expectation among humans

Specific to process-centric vs outcome-centric explanations, however, there is a lack of evidence comparing the humans' subjective perceptions.

Gap: Do humans have different subjective perceptions about process-centric explanations, compared to outcome-centric explanations?

6 LIMITATIONS AND CHALLENGES

In this section I describe the limitations of my approach.

- First and foremost, by combining studies across different domains, I have abstracted out and ignored many details about the context and the nature of the task that may be relevant to the design of the explanation. For instance, the context in which AI is embedded may – whether a digital screen or a voice agent – can make certain explanations (e.g., visual representations) (un)desirable. Many of those aspects will eventually determine the explanation mechanism that works best and as prior research has found, explanation mechanisms should always be tested in the specific domain [21].
- For the most part in my framework, AI models are considered black boxes, and mechanisms for generating explanations are considered independent of the AI model. However, my choice is aligned with prior literature [9] that argues for considering explanation systems separate from AI models to specifically focus on the usability of the explanation. Furthermore, new methods continue to be developed that can generate a particular type of explanation (for e.g., counterfactuals [20]) for any type of AI model.
- Furthermore, prior research also shows that humans tend to develop and update their mental models about the AI over repeated interactions. Most of the research covered in my analysis ignores this aspect, and therefore, my framework is also limited in that regard. That is, it remains unclear how the design claims I present will evolve through repeated interactions with the same AI. Future research should look into how human information needs evolve over time, and if a sequence or a combination of explanations over time will be more effective than just one explanation design.

REFERENCES

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. *arXiv:2006.14779 [cs]* (Jan. 2021). arXiv:2006.14779 [cs]

[2] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (April 2018), 1–14. https://doi.org/10.1145/3173574.3173951 arXiv:1801.10408

[3] Tommy Bruzzese, Irena Gao, Griffin Dietz, Christina Ding, and Alyssa Romanos. 2020. Effect of Confidence Indicators on Trust in AI-Generated Profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382842

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 454–464. https://doi.org/10.1145/3377325.3377498

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287

[6] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3301275.3302289

[7] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300234

[8] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 95–106.

[9] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. 2019. Accountability of AI Under the Law: The Role of Explanation. https://doi.org/10.48550/arXiv.1711.01134 arXiv:1711.01134 [cs, stat]

[10] Yanqing Duan, John S. Edwards, and Yogesh K Dwivedi. 2019. Artificial Intelligence for Decision Making in the Era of Big Data – Evolution, Challenges and Research Agenda. *International Journal of Information Management* 48 (Oct. 2019), 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[11] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 90–99. https://doi.org/10.1145/3287560.3287563

[12] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359152

[13] Jeffrey Heer. 2019. Agency plus Automation: Designing Artificial Intelligence into Interactive Systems. *Proceedings of the National Academy of Sciences* 116, 6 (Feb. 2019), 1844–1850. https://doi.org/10.1073/pnas.1807184115

[14] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (July 2019), 31:1–31:35. https://doi.org/10.1145/3338243

[15] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300641

[16] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An Evaluation of the Human-Interpretability of Explanation. *arXiv:1902.00006 [cs, stat]* (Aug. 2019). arXiv:1902.00006 [cs, stat]

[17] Zachary C. Lipton. 2017. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]* (March 2017). arXiv:1606.03490 [cs, stat]

[18] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]* (Aug. 2018). arXiv:1706.07269 [cs]

[19] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan. 2019), 279–288. https://doi.org/10.1145/3287560.3287574 arXiv:1811.01439

[20] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617. https://doi.org/10.1145/3351095.3372850 arXiv:1905.07697 [cs, stat]

[21] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–52. https://doi.org/10.1145/3411764.3445315

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.48550/arXiv.1602.04938 arXiv:1602.04938 [cs, stat]

[23] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-Aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14.

https://doi.org/10.1145/3313831.3376506

[24] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human + Machine Complementarity for Recidivism Predictions. https://doi.org/10.48550/arXiv.1808.09123 arXiv:1808.09123 [cs, stat]

[25] Debora Viana Thompson, Rebecca W. Hamilton, and Petia K. Petrova. 2009. When Mental Simulation Hinders Behavior: The Effects of Process-Oriented Thinking on Decision Difficulty and Performance. *Journal of Consumer Research* 36, 4 (Dec. 2009), 562–574. https://doi.org/10.1086/599325

[26] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. 185 (1974), 10.

[27] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (April 2018), 1–14. https://doi.org/10.1145/3173574.3174014 arXiv:1802.01029

[28] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[29] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 318–328. https://doi.org/10.1145/3397481.3450650

[30] Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486

[31] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852