A Computational Cognitive Model of Informative and Persuasive Explanations of Artificial Intelligence Systems

Shane T. Mueller
Kit Cischke
Lamia Alam
Tauseef Mamun
Michigan Technological University

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

Cite as:
Mueller, S.T., Cischke, K., Alam, L., and Mamun, T. (2021). "A Computational Cognitive Model of Informative and Persuasive Explanations of Artificial Intelligence Systems" Technical Report, DARPA Explainable AI Program.

Table of Contents

Abstract

In this Report, we describe a series of Computational Cognitive Models (CCMs) that help account for major forms and goals of explanations in AI systems. A previous report on Sub-Task 2.5 (titled "A Computational Model of Explanatory Reasoning: Foundations of Explanation in Sensemaking") described the Computational Cognitive Model and an evaluation of it via simulations. This follow-on Technical Report begins by describing the basic framework of the CCM, distinguishing between learning that can occur at two levels: (1) a slow feedback-based level associated with System-I processing, and (2) a fast knowledge reconfiguration associated with System-II processing. We demonstrate how formal CCMs that simulate and represent knowledge can provide accounts for how different kinds of explanations (examples, saliency highlighting) interact with different goals (persuasive versus informative explanations) that represent the dominant modes of algorithmic explanation used in modern XAI systems. We present results from an experiment establishing that subjective assessments of satisfaction, trust, and the like can be dissociated from a better understanding of a system. We explore a representational CCM that helps understand how explanations can persuade, with or without providing strong explanatory information. Finally, we discuss these findings with respect to how they may inform the development of XAI systems.

1. Background and Prior Work

Explanation and Sensemaking

Previously (see Mueller et al., 2019), we have described how explanation is closely linked to sensemaking, and described a series of computational models that both establish distinct learning and decision functions of sensemaking, and link these to explanatory reasoning. Figure 1 shows the basic framework we proposed based on the Data/Frame model of Klein et al. (2006), but also incorporating a critical additional element: distinguishing between learning that happens via low-level intuitive system-1 feedback mechanisms, and learning that happens through deliberate knowledge exploration and reconfiguration.

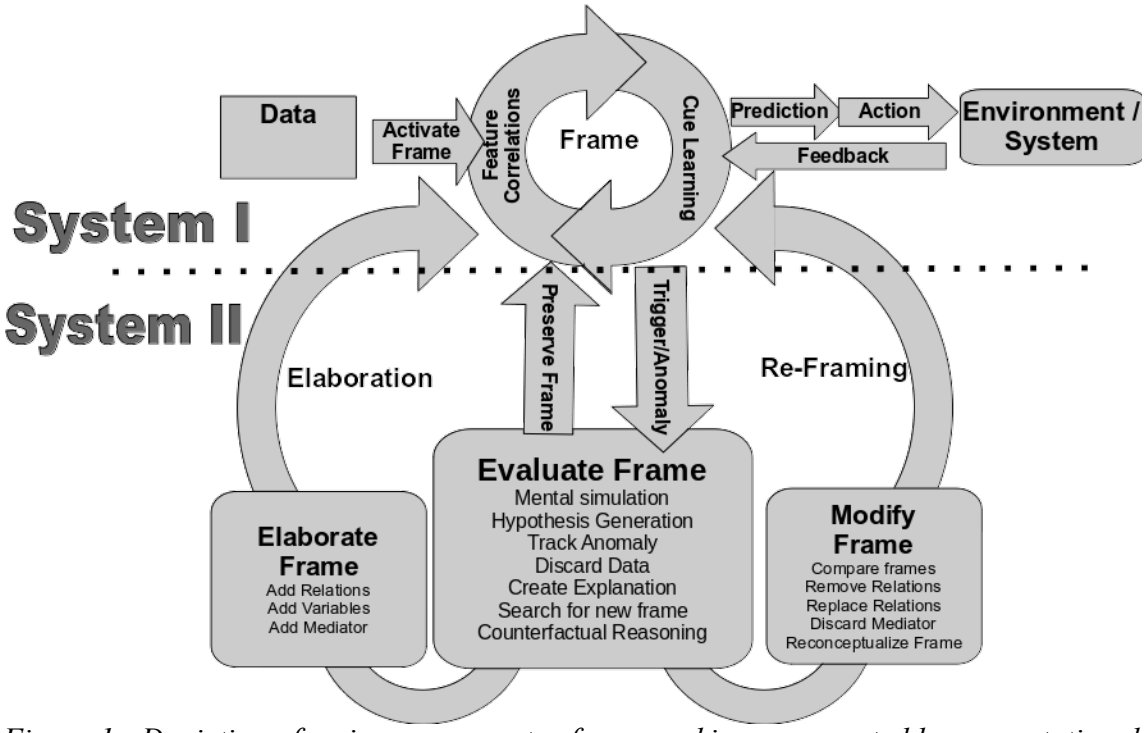

Figure 1: Depiction of major components of sensemaking represented by computational model of explanatory reasoning.

System-I and System-II Processes

Kahneman (2011) described complementary processes of fast and slow thinking in terms of the labels "System I" and "System II". According to this framework, System-I consists of intuitive processing and recognition, whereas System-II involves deliberative processing, causal reasoning, and the like. Although this is normally framed in terms of how new information is accessed, the same distinction exists during learning, with System-I involving low-level feedback and correlation-based learning (operant and classical conditioning, error-based feedback learning, hebbian learning, etc.), and with System-II involving knowledge reconfiguration. These two systems map nicely onto the top and bottom of the sensemaking processes in Figure 1. Interestingly, although System-I processes are typically associated with *fast* reasoning and System-II with *slow*, this appears to reverse during learning. Here, cue-based learning of contingencies (System-I) based on feedback and error may take hundreds of exposures to form accurate stimulus-response mappings, whereas knowledge reframing and rule learning (System-II) can produce huge benefits from minimal training or just a few examples.

These complementary learning systems are relevant to XAI explanations because many explanations that are generated automatically are fairly opaque, and difficult to verbalize or form interpretable rules from. These include examples, heatmap visualizations, and even contrast cases. Thus, learning how the system works based on these artifacts may require long-term exposure to many examples in order to learn via System-I learning mechanisms.

Example paradigms and data patterns

Although initial form of the DARPA program distinguished between autonomy and data analytics, the final systems blurred lines between these domains, and common explanatory idioms were used across many different approaches. These include the use of examples of various types, the use of feature highlighting or saliency approaches, and the use of contrast in both of these cases. When abstracted to a level that can be modeled via simulation, the overlap in the example paradigms is substantial, whereas the distinction between explanation idioms is more interesting and requires different modeling assumptions. Consequently, in this report, we will use a common set of models to illustrate understanding of explanations used for both analytics and autonomy situations.

2. Modeling Approach

The models we present here attempt to answer two general sets of questions for distinct explanatory goals: **persuasion** and **understanding**. The distinction between these can be understood by considering a credit rating algorithm that decides whether a person should qualify for a loan. Explanations might take on a number of forms: identifying a piece of data (e.g., loan repayment history) that was important in denying the loan, providing a real or fabricated example (showing a similar person whose loan application was accepted) and identifying how they differed along some set of decision features, showing the relative importance of different features in the global decision process, or showing a set of applicants who were given loans. Some of these provide a decision maker or applicant information they can use to develop a better mental model of the algorithm, and so will be able to make accurate predictions about loan acceptance, and possibly even change behaviors in order to comply with the things the algorithm values. These would enhance understanding. One example of this would involve the kinds of advice credit counselors might give after assessing a credit report. However, many potential explanations do not enhance understanding directly, but may mostly work to convince a loan officer that the algorithm is working consistently or with reasonable data. For example, an explanation that simply highlights the most important feature that led to a loan application being accepted or rejected has a much lower bar---it merely needs to persuade the user that the decision was reasonable, even if they cannot generalize from that to form a better mental model.

The cognitive mechanisms of persuasion and understanding also depend on the kind of explanation being given. For example, feature-highlighting or saliency explanation (such as LIME) works by informing a person that a particular feature should be given more weight in their mental model of how a system works, whereas, if an example is given as an explanation, they must compare features to infer which ones were most important.

Modeling Assumptions and Framework

In the models described in this report, we simplify the modeling framework used in previous models reported in Mueller et al. (2019) in several ways, focusing specifically on the reinforcement learning mechanisms involved in error-based feedback learning, and contrasting it to various shortcuts that System-II sensemaking may provide via AI explanations. In general, the sensemaking process relies on the formation of a frame and then the tuning of that frame based on

feedback. Tuning the frame is often described as using a delta-rule learning scheme, as discussed in the main report. Delta-rule learning is necessarily slow and iterative, where steps towards a "correct" understanding are limited by the learning rate, α. Higher values of α allow for faster learning, while lower values slow the process down. In either case, the mental model of the learner will eventually converge on the mental model of the system given enough trials and feedback. Example source code defining these models can be provided upon request.

We argue it should be possible to "shortcut" iterative learning with an *explanation* process. If the feedback is more specific and has some kind of semantic value that provides information about *why* a particular outcome occurred instead of simply the outcome itself, the learner can make major adjustments to their frame instead of the proportional changes that usually characterize delta-rule learning. This should allow the frame to be closely aligned with "reality" within several learning trials instead of hundreds. For example, Jeep brand vehicles have several characteristic features, including seven-slotted grilles and trapezoidal wheel arches. Highlighting these features in an explanation should allow the learner to adjust their frame more rapidly than in traditional delta-rule learning, in which over many experiences with a Jeep, an observer might eventually notice a pattern, or may be able to learn the pattern implicitly without being able to verbalize it.

3. Simulation: Example-based Explanations for Understanding

Use of examples is a important method for explanation in XAI, and for learning in general. It should be noted that example-based learning models (see Anderson et al., 1994; 1997) have been important in explaining procedural skill, and similar mechanisms are relevant for XAI. In addition, In research on pedagogy (e.g., Atkinson et al., 2000; van Gog, et al., 2011), extensive research has centered on the use of worked examples as a method of learning, which typically relates to the ability to make analogies between similar procedures to learn a general problem solving approach.

Potential benefits of Examples

Although examples are frequently used and adopted in XAI applications, they may confer many different kinds of information, many of which may have benefits, but this benefit may only occur if the user understand what information the examples are trying to convey. For example, nearest-neighbor examples may be useful in showing how other similar cases were classified or the decisions an autonomous system made in similar situations. The similarity of the examples to a test case show the user which features appear to matter, and difference may even show which features do not matter. However, in a situation where both the human and AI perform well, nearest-neighbor examples may provide no benefit. As an example, if an AI correctly labels an image as an elephant, showing another image also correctly identified as an elephant provides limited information, unless the new image differs in some interesting way. An alternative kind of example-based explanation is to show typical cases of either the chosen category or an opposite categorization. This can help a user generate a veridical mental model of the AI's concept, but it may be confusing if the user does not clearly understand how the AI is choosing examples. Yet another strategy is to select examples that illustrate feature importance. Mueller at al. (2021) followed the work of Kenny & Keane (2021) and to advocate use of a semifactual-counterfactual

sequence to explain an AI classifier or autonomous agent. Here, starting with a specific case, a small change is made (or a case selected) that does not change the outcome, followed by a slightly larger change along the same conceptual dimension that does change the outcome. This is a special case of using contrasting or counterfactual examples to help explain decisions (see Shafto et al., 2014).

Depending on how the examples are chosen, it is easy to see a direct connection to saliency approaches. If a pair of examples is chosen that differ by a single feature, and the classification or decision changes because of the difference, this is conceptually identical to a saliency or feature highlighting approach. However, this could easily backfire if the user does not understand the rationale for selecting examples. In our observations, it may frequently be unclear to users why examples were chosen, and they may be given abstract rationale like "we made this decision based on what we learned from these examples".

Description of Simulations

To investigate these issues more carefully using the computational model, we again returned to the delta-rule learning model, to investigate how different kinds of examples might lead to faster or slower learning. In our simulations, a particular input configuration was categorized as being of a particular set. The AI then provides an additional example of that set in order to justify or explain the categorization. Colloquially, the AI says, "I called this picture a bird because it looks a lot like this other bird." Strictly speaking, this will always accelerate the learning rate because instead of a single delta-rule update in every trial, the user experiences two such updates.

In this simulation, we will examine several kinds of examples. First, we can consider the value of a *random* example. A random example is simply a randomly-chosen case, and serves as a useful control condition because it provides two non-identical cases to generalize from. We define the *prototype* example as an instance where every characteristic is set to the mean value of the input set for a given class. This may or may not reflect a real item in the set. Attempting to define a prototypical "dog" with an average ear length, tail length, and color would almost certainly not map to any extant dog. However, this is mathematically consistent with existing theories on cognitive prototypes (see Estes, 1986; Nosofsky 1992). Similarly, let us define a *caricature* or *epitome* as an instance where the features are set to exaggerated or maximal values. The underlying notion here is that if a feature has some degree of importance, we shall maximize its impact. A feature ignored by the AI will be set to 0 so as to eliminate its influence on the categorization. Again, such an item may not exist in the set or even be realistic. Mueller (2020) discussed how human classification is often better for caricatures, and our use here provides a roadmap for how caricatures might be used to improve explanations.

We ran the same types of simulations as in the previous section while varying the type of example provided to the user during the learning process. Shown in Figure 2 are the results of these three simulations.

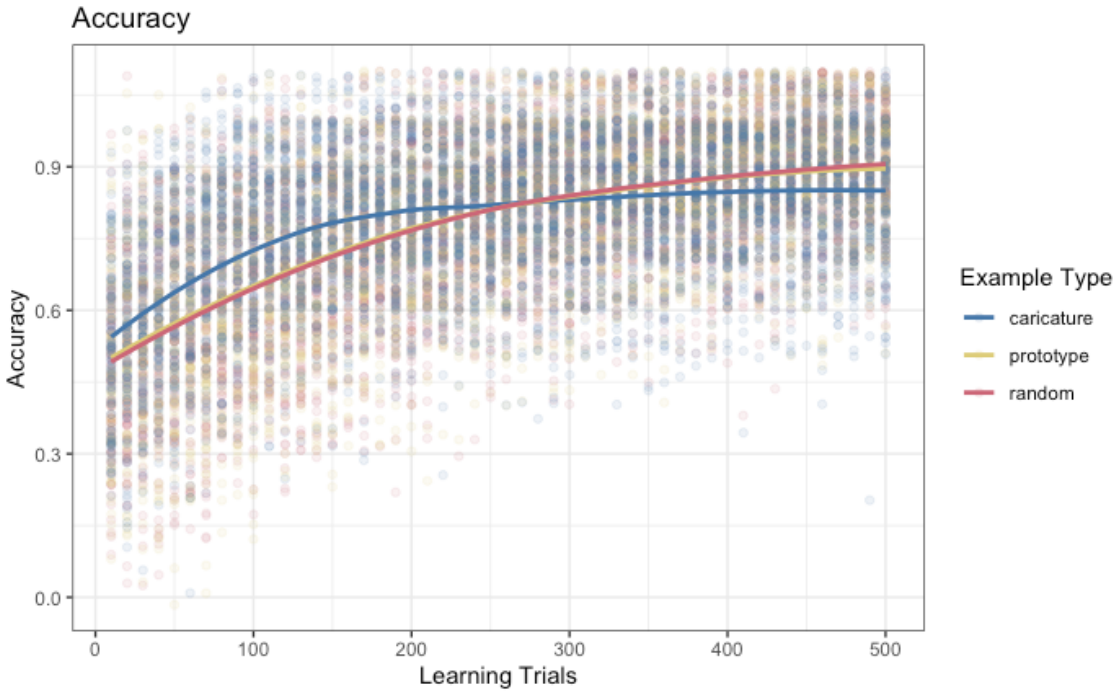

Figure 2: Caricature examples are effective early in the process, if only for mathematical reasons. Random examples are indistinguishable from prototypes.

Results

Results revealed that the effective learning rate for providing prototypical examples appears no different from random examples. This result may even be optimistic, because our simulated categories are very compact and regular. Researchers have found that exemplar-based models often provide a better representation of the shape of categories (Nosofsky, 1988), and for cases where categories are complex, random examples may provide better learning than the prototype. It is perhaps surprising that the value of prototypes is no better and no worse than selecting an entirely random example of the same classification. Continually reinforcing the mean values apparently does not increase learning over simply showing another example. Indeed, the impact of the random example actually becomes greater than the impact of the prototype in later trials. If we are, for example, classifying pictures of "dogs" and "not dogs", continually showing the prototypical dog has no advantage over showing some other random dog photo in terms of helping the user understand what the AI believes a dog looks like.

Next, we found that the caricature examples are especially helpful early in the learning process. Mathematically, this is essentially functionally the same as increasing the learning rate, because the delta-rule learning with caricatures will produce larger errors, and thus will calculate larger average differences between its current value and optimal value. Consider that the value of the caricatured coefficients are 1.5x greater than the mean values. The associative rule of multiplication allows us to reallocate that factor of 1.5 to the value of α. Alternatively, we can simply consider that multiplying larger coefficients by the same learning rate will result in larger adaptations. However, the learning rate quickly levels off, as the large changes actually result in

worse predictions. The real-world analog here is also clear. We often delineate the most extreme examples of a solution technique, classification, or other material to rapidly refine the learner's understanding. At some point, the value of the extremes becomes limited and we must use more refined examples.

Discussion of Example-based explanations

Learning processes that include effective explanations, whether through feature highlighting, counter-factual examples, or other techniques should accelerate the learning process compared to strict delta-rule learning. This should reduce the number of trials required to effectively predict an outcome from hundreds down to less than a dozen trials. None of the mathematical approaches we have described here can do that, regardless of the way that we attempt to tune underlying parameters. The frequency of an explanation may have no impact or may have a negative impact. Attempting to emphasize factors that strongly influenced an outcome does not have any real impact. This is true for small boosts, moderate boosts and even large boosts. The only way we can positively impact the prediction accuracy is by more aggressively zeroing out parameters that did not meaningfully contribute to the outcome. This still does not reach our goal of rapid sensemaking through major frame shifts.

In terms of selecting examples to accelerate learning, we find that prototypes most rapidly affect the early learning process, but lose effectiveness over time. There is no value in reusing a prototypical example over and above the use of a random example of the same class when trying to accelerate learning.

Linking contrasting example and saliency based explanations

When two contrasting or counterfactual examples are given, we suggest the cognitive operations related to processing the explanation are essentially identical to saliency information typically provided by algorithms such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017). That is, by comparing a pair of contrasting examples that have different classifications and different features, we arrive at a kind of information that is processed the same as what a saliency model produces---it shows which features are most important in making the classification versus another one. The relative effectiveness of the two approaches will certainly depend on how examples or saliency is identified, but our models will treat them in identical ways. Consequently, the next set of simulations will deal with both cases at the same time.

4. Simulation: Incorporating Saliency and example-difference Explanations in Understanding for System-I processes

In the present set of simulations, we provide a simulation model of the processes involved in either (1) saliency explanations that attempt to show a user what features are being used to make a decision, and (2) contrastive example explanations that attempt to show a user how a change in one or more features can lead to a change in the choice or decision. In all of these simulations, we focus on the simplest case in which an explanation focuses on a single feature, which we think is

probably optimal for reasoning and learning, but we consider saliency and contrastive examples and providing the same kind of information that is used in identical ways by the user.

We first investigated how System-1 style learning might incorporate these kinds of explanations. After some early experiments, we settled on comparing three learning strategies that provided system-1 based learning. The goal here is to determine whether reweighting a traditional error-based feedback learning rule to overweigh the 'explained' feature will provide a substantial (or any) learning boost. If we can establish that highlighting can impact system-1 style learning, this reduces the need for a dual-process model incorporating knowledge reconfiguration in order to learn the behavior of an AI system. In the first simulation, we examined:

- Pure delta-rule learning
- A "filtered" delta-rule, where a subset of the coefficients (dictated by the highlighted or salient feature) would be affected
- An "explanation" strategy, wherein the most important coefficients were boosted and unimportant coefficients were minimized or zeroed-out altogether

The pure delta rule learning serves as a control condition. The other two strategies serve as reasonable ways in which explanation features could be incorporated into a delta-rule learning scheme while still using that scheme. However, these schemes can be destructive—zeroing out values might help initially, but on occasion a learned value might be zeroed-out which would represent interference or unlearning of the AI system. Consequently, we investigated mixture strategies, where a pure delta-rule was mixed together with one of the others on in different percentages, so that delta-rule could be "boosted" by an explanation every *n* trials.

Strategy Rationale

The delta-rule strategy is well-known and not worth discussion here aside from noting that we used $\alpha = 0.2$. The filtered delta-rule is our first attempt to model an explanation. The AI indicates the *k* most important features that led to a classification and standard delta-rule equations are applied only to those coefficients of the ILM. The notion here is that we are emphasizing certain features that contribute most strongly the classification outcome and trying to bring them closer to the model of the AI. Other coefficients in the ILM are left unchanged.

The "single-feature boost" is based on an intuitive sense of what might happen when a particular feature is emphasized in an explanation. If another human explains that the key way to identify a Jeep is by the shape of the wheel arches and the number of slots in the grille, other features may be completely ignored. Mathematically, the model takes one of the coefficients that most strongly contributed to the classification outcome and sets this to a suitably large value while simultaneously de-emphasizing or even zeroing out the other coefficients. We do not always pick the single biggest coefficient, but choose from several factors proportionally to their contribution to the classification outcome. Thus, we are able to let the model choose between a somewhat arbitrary number of important features to boost over the course of several learning trials. There are two main variables here that may affect the efficacy of the explanations: the new boosted coefficient value; and the threshold that determines unimportant coefficients. We investigate the importance of each.

There is a third factor that controls the single-feature boost: how many times we will perform the boost. It is advantageous to be able to control this so that there isn't "thrashing" where what is learned in one trial is zeroed out in a subsequent trial.

Simulation Process

In the simulations, a random set of cases is generated with 8 input features. The AI classifies each according to its 4-feature model, and produces the "ground truth" (whether the classification is real-world accurate or not). That is to say, an input image may be classified as containing a handwritten '9' that is actually the digit '4'. After each trial, the user reevaluates their frame according to the selected strategy. If the user and AI had similar results, the frame change will be minimal. If the results are considerably different, the user will have more significant changes to their frame. In initial runs, three strategy mixes were employed: a pure delta-rule learning for reference (Strategy 1 in the plots); a strategy where there one 1% explanations (Strategy 2); and a strategy with 50% explanations (Strategy 3), though the number of explanations is still constrained by the maximum explanation limit.

Simulation Results

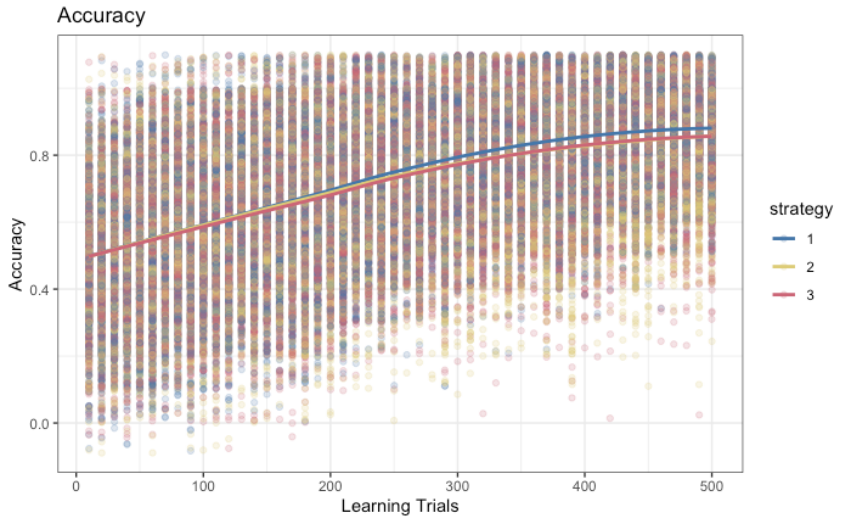

Figure 3:The prediction accuracy for all three strategies across all trials with default values for the zero threshold, boosted value, and number of explanations) is essentially the same. There is no early boost from explanation.

Simulation results (Figure *3*) shows a classification accuracy measure: each case was classified as a positive or negative value and the model was scored as correct if it correctly predicted the sign with its linear model. *Figure 3* shows the cumulative results across 500 random inputs, run 100 times for each strategy. The model starts at around 50% accuracy (chance) and shows a slow increase in accuracy over several hundred trials until it begins to asymptote. The results show a

classic System-I learning process, and neither of the strategies we examined for introducing explanation improved learning. All three strategies grow at approximately the same rate and achieve the same level of success after 500 trials. Clearly, the two explanation strategies do not produce short-term or long-term boosts in accuracy, so as a System-I explanation mechanism, it would seem that boosting certain coefficients and deemphasizing others provides no immediate benefit, as the other coefficients contributed in some way to the original classification.

If the strategy mixtures themselves have no impact on the learning rate, there are three other factors that can be adjusted to determine if any of these strategies effectively describes learning with explanations.

Altering the Number of Explanatory 'Boosts'

We suspect that allowing an explanation strategy to repeatedly boost features would result in decreased accuracy due to "thrashing". After the basic importance of features is learned, further explanations might better be ignored if they result in resetting values. This is clearly illustrated in Figure 4. Here, we compare two simulations: one in which the cognitive model stops considering explanations after 10, and another that continues to consider and use each explanation up to 500 in a row. The learner is constantly switching what the most important feature is and forgetting what they have learned about other important features or how little the unimportant features contribute. The overall prediction accuracy stays below 75% and still initially grows at a rate similar to pure delta-rule learning. The default value of 10 explanations is indistinguishable from delta-rule learning. And as more explanations are considered in the 500-explanation condition, Strategy 3 (which resets coefficients to 0) experiences a substantial performance decrement.

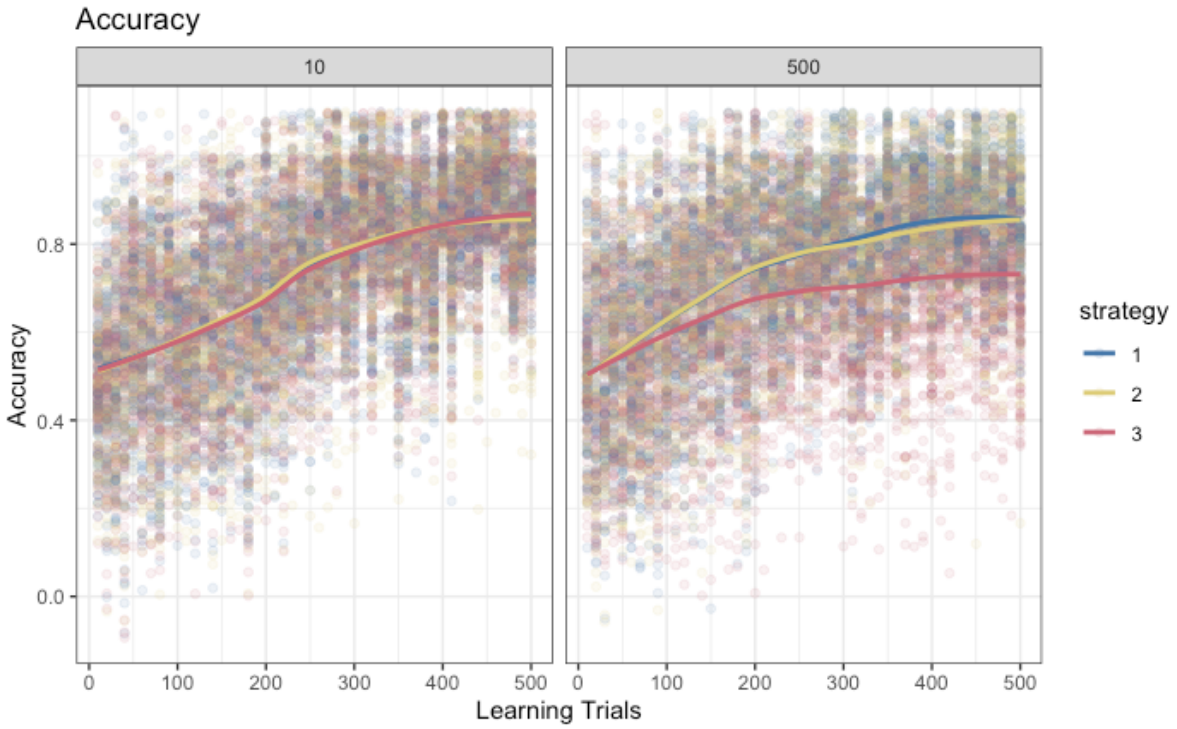

Figure 4: Allowing only 10 explanations is essentially the same as delta-rule learning, so much so that the plots overlap to the point strategy 1 is completely occluded. Many explanations actually decreases accuracy.

Altering the Value of the Boost

In the single-feature boost model (strategy 3), we examine the feature' current value before adjusting it. Initially, we ensured that the new coefficient is *at least* ±0.4 (recall that approximately half of the coefficients are negative). The rationale here is that if an explanation says that a feature is important, it should only impact the user's understanding if the user did not already know it was important. However, the specific threshold is somewhat arbitrary, so we wanted to evaluate whether the particular boost threshold mattered. To examine this, we ran simulations with this boost value set to 0.2 and 0.8. The results are shown in Figure 5, which shows that this particular value has no impact on the cognitive model's performance: a higher threshold does not lead to slower or faster learning.

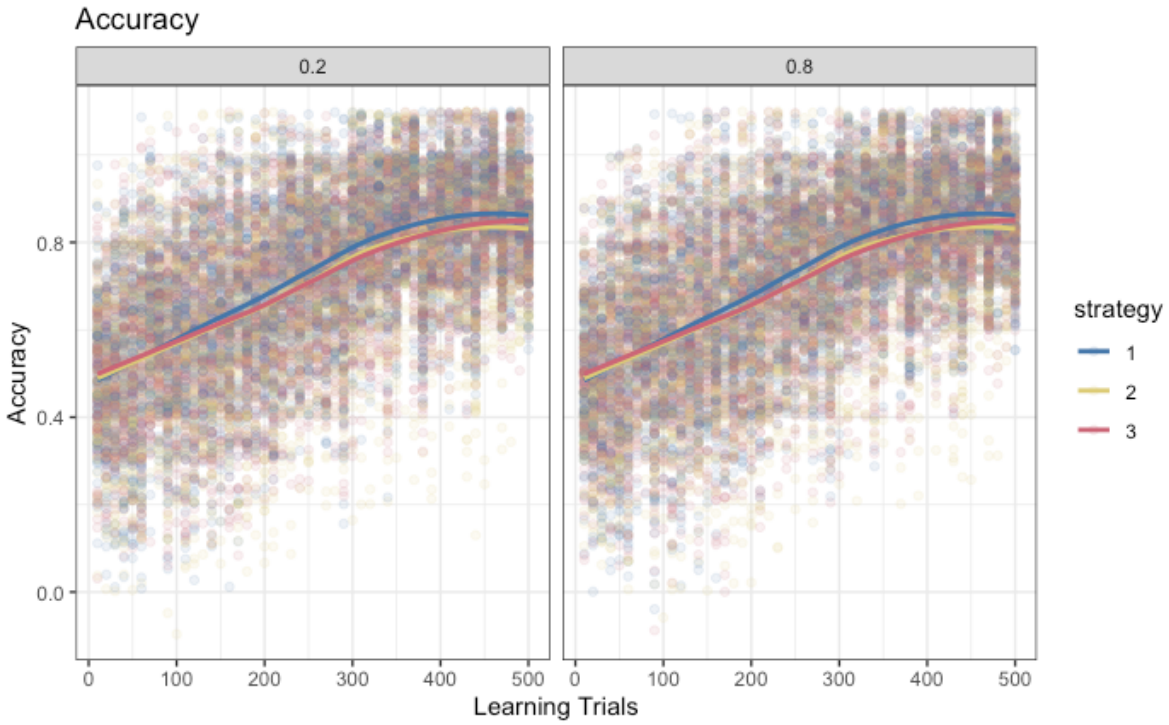

Figure 5: Boosting a coefficient in the ILM to 0.2 or 0.8 during an explanation has no real effect on the learning rate. Ultimately, pure delta-rule learning still provides the best accuracy.

Altering the Zero Threshold

Similarly, strategy 3 deemphasizes "unimportant" values in the ILM, and we utilized a threshold to ensure that we don't reduce the helpful, useful influence of other coefficients. By default, this value is set to ±0.4, so that and only values less than this threshold are zeroed out. Simulations were run that varied this parameter to a small value of ±0.2 (very few coefficients were zeroed) and to a large value of ±0.8 (most coefficients are zeroed). As shown in Figure 6, this is the first modification that has a meaningful impact. When this threshold is small, all strategies behave similarly. When it is large, this conveys a modest learning benefit. Essentially, small values takes away the power of the explanation boost and makes it functionally equivalent to pure delta-rule learning. Setting the zeroing threshold to 0.8 increases the power of the explanation, but the effect is still not seen until after 100 trials and ultimately is no better than delta-rule learning. The on-going normalization of the coefficients makes the truly important factors increasingly important and tremendously de-emphasizes the other factors.

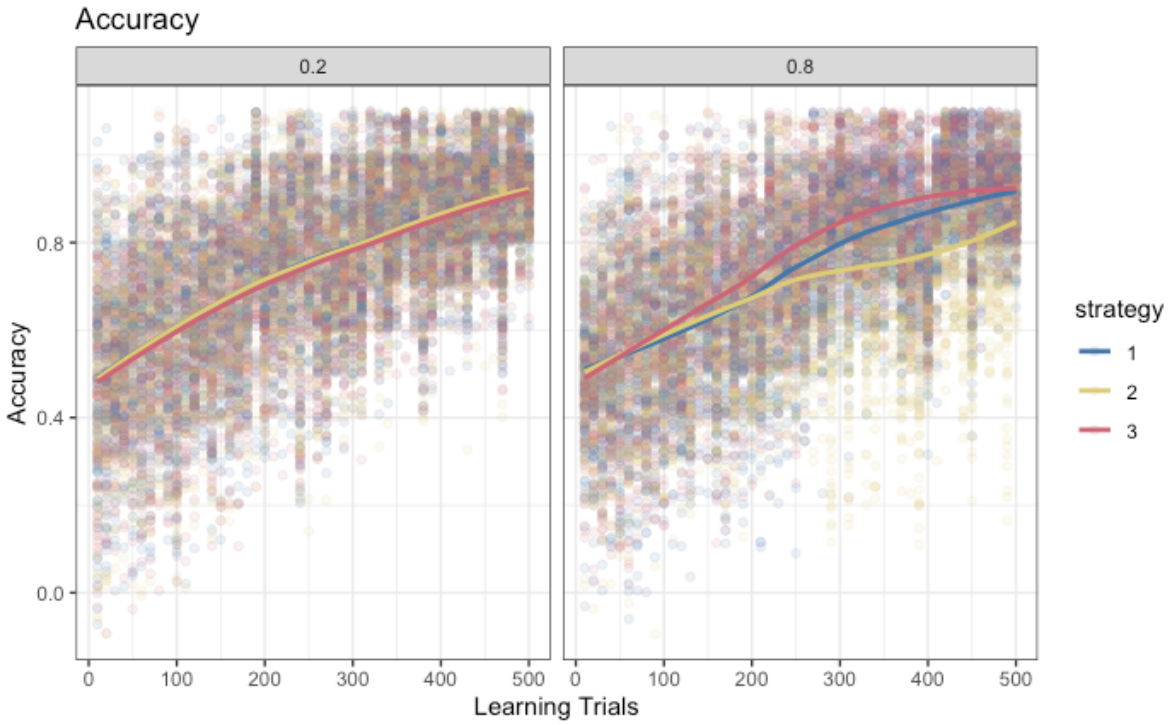

Figure 6: The effect of zeroing out coefficients less than 0.2 (left) or less than 0.8 (right). The explanations that emphasize only the strongest coefficients and zero out almost everything else have the strongest effect, but are only marginally better that delta-rule.

Prospects for explanations impacting System-1 learning

The models presented in this section illustrate the potential for adapting a system-I learning mechanism to incorporate feature-highlighting or saliency. The results are not very promising. That is to say, if we believe that highlighting or saliency will naturally lead to a better understanding via incidental or implicit learning, at best we found schemes that could improve learning accuracy be maybe 10% after hundreds of trials, in comparison to no-explanation. This also suggests that, unless this information is framed to the user in a way that supports sensemaking, it may not help them understand the system better.

Next, we consider how a System-II account might incorporate feature highlighting.

5. Simulation: Incorporating Saliency and Example-Difference Explanations in Understanding for System-2 processes

If System-1 style learning schemes seem ineffective at incorporating feature or saliency information, perhaps System-II might be better. Our previous models (see Mueller et al., 2019) envisioned System-II processes as modifications in a mental model, or movement through a space or network of mental models, and System-I as delta-rule learning within a particular mental model. In the present simulation, we use a slightly different strategy, where System-I learning is based on

delta-rule learning, but System-II learning is based on episodic memory for recent explanations, such that the predictive model is adjusted to match the frequencies of the prior explanations. As an example, if across 10 cases of a bird image classifier, a saliency map might highlight the tail and wings of bird 6 times, and the feet 2 times. This would provide a naive importance map weighing tails and wings and feet, but ignoring other aspects such as beak or the background. This might not be the complete model that the user has learned is important, so we then adjust the current model to be a 50-50 mixture of the delta-rule coefficients and this saliency-history pattern. As before, we compare three models that mix this strategy with a delta rule learning, in a proportion of either 100-0 (strategy 1), 50-50 (strategy 2), or 0-100 (strategy 3).

Basic prediction accuracy

Figure **Error! Reference source not found.** shows mean prediction accuracy (across 100 repeated runs of the model) over the first 100 trials of the simulation for three strategies. Strategy 1 is a pure delta-rule learning mode, and is thus identical to the simulations examined in the previous series of simulations. In comparison, Strategy 2 and 3 involve pure and mixed System-II explanation strategies. Unlike the System-I models, these immediately and drastically improve performance. Indeed, basic delta-rule learning is minimal over the first 100 trials, but the knowledge reconfiguration model produced an immediate benefit in the first 10 trials.

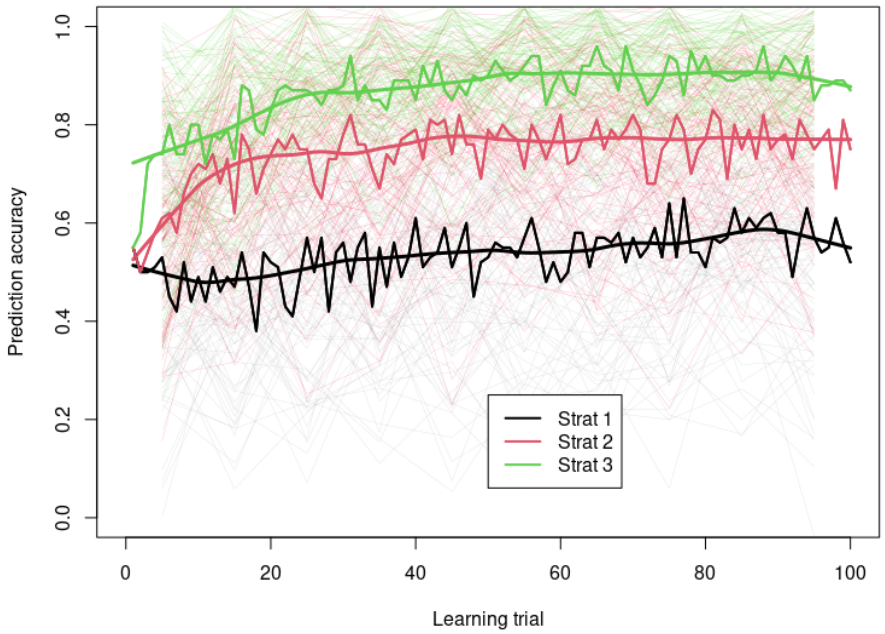

Figure 7: Prediction accuracy for three models, with black representing a pure feedback-learning model, green representing a model incorporating only historically salient features, and red representing model that mixes these two strategies. Historical saliency provides immediate and substantial learning over alternatives.

In addition, we can examine how quickly the coefficients are learned in each strategy. In Figure 7, we plot the actual coefficient values for a typical run of the cognitive model, with the dots on the

right side representing the 'true' coefficients producing the data. Here, the comparison is stark: the System-I learning involves slow drift of the coefficients toward appropriate values, whereas the System-II reconfiguration involves immediate transformation into a reasonable initial model that can then adjust toward the optimal values.

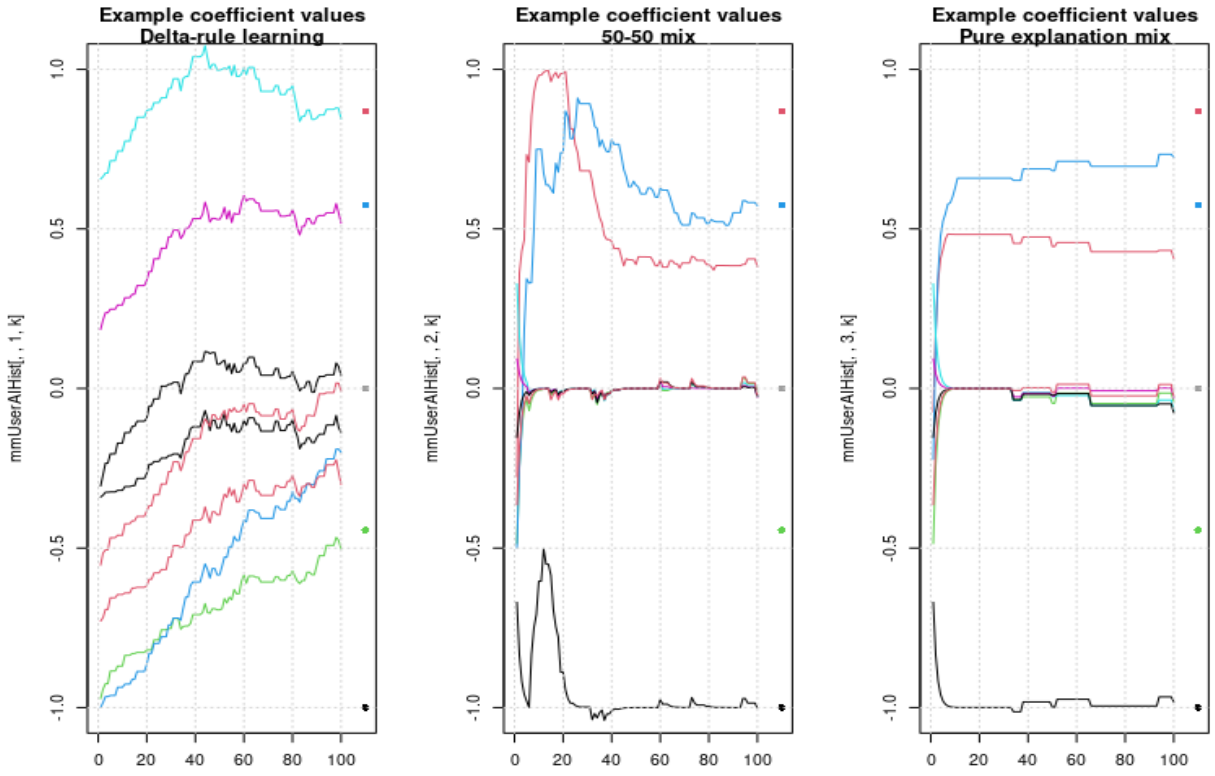

Figure 7: Learning weights of model via three different strategies. Using a history of salient features can quickly provide an accurate mental model (middle and right panels), especially in comparison to slow reinforcement learning schemes (left panel)

Discussion and implications of System-I and System-II accounts of saliency explanations

The previous two groups of simulations illustrate how one kind of explanation (saliency) can impact understanding of an AI system, in terms of its ability to predict performance of the system. These models suggest that saliency or feature-highlighting approaches like LIME (Ribeiro, et al., 2016), SHAP (Lundberg & Lee, 2017), and others have the potential to provide huge advantages in helping a user understand and predict a system (provided they are valid). However, they must be presented to the user in a way that encourages this knowledge reconfiguration, because incidental highlighting without the opportunity to explore, engage curiosity, and self-explain is likely to provide no advantage over AI-only experience.

Some suggestions for optimizing the likelihood of using System-II processes include:

- Provide history cues to allow a user to see not just the most recent salient feature, but other prior salient features
- Quantize the saliency/highlighting to simplify the causal model and memory encoding. A single important feature will be easier to remember than five weak features. In our simulations, a distribution of salient features emerged based a single feature from each example, and this was sufficient to provide extremely fast learning
- Encourage active processing and engagement with the explanation, so that systematic patterns across cases can be learned. In our simulations, the successful agent is not just learning cue-response tendencies, but a history of explanations, and this historical snapshot is what provides the greatest benefit.

Persuasive Explanations in XAI

Hoffman et al. (2018) suggested that satisfaction, trust, and other subjective assessments of an explanation follow from a better understanding of the system. Although this may be true, the psychological mechanisms by which people gain trust in a system are multi-faceted. Importantly, we often place trust in systems we do not completely understand, sometimes for good reasons (it is vouched for by a trustworthy source who does understand the system), but sometimes for not very good reasons (it has a polished interface or it helps us understand a particularly salient situation). Other concerns also come into play, including cost-benefit analysis (a system may not be trusted completely, but there is little downside if it fails) and convenience.

Some of the most popular ways of measuring the effectiveness of XAI involves subjective ratings of justifications (Hoffman et al., 2018). In these cases, the explanation involves justifying why a decision was made, and the evaluation regards whether a user finds it reasonable or persuasive. If the user finds it reasonable, they may judge the system highly according to subjective ratings, even if they are unable to make good future predictions of how it works. For simple systems this might seem unlikely, but for a complex system such as a self-driving vehicle, multi-context image classifier, machine translation system, or recommender system, the scope of behavior may be so complex that understanding a single decision might be possible even if it cannot be generalized to making better predictions. For example, a classic recommender system may suggest to a viewer that they will enjoy a particular movie. An explainable recommender may provide a justification: "You might like this movie about a bank heist because you watched three other movies about bank heists". This particular decision appears reasonable, but it doesn't mean that other recommendations will be.

The perspective that computers serve as technology for persuasion dates at least to Fogg's (1998) notion of 'captology'. One important aspect of this research is that persuasive technologies have ethical consequences, because persuasion can be used for ignoble purposes. Thus, AI explanations that serve only to convince or persuade are feasible, which may only be apparent with a more complete evaluation.

To illustrate the potential distinction between persuasion and understanding, we next present the results from a human subjects experiment.

5. Experiment on Persuasion and Understanding in autonomous analytics systems

Background

Our goal in this experiment is to provide data examining the impact of different kinds of explanations on human interactions with a simulated autonomous analytics system. Thus, the experiment attempts to address both major thrusts of the XAI program: autonomy and data analytics. The simulated AI involved an interactive scenario in which a participant played the role of a patient interacting with an AI diagnostic system. The interactions took place across a simulated 6-week event, with weekly consultations of the diagnostic system. The goal of the study was to determine how different kinds of explanations (examples, contrasts, and combined examples and contrasts) impacted both subjective ratings (i.e., persuasion) and knowledge (ability to predict future behavior) of the system. We hypothesized that both subjective ratings of the competency of the explanations and their ability to predict the AI system would improve for each of the explanations in comparison to control, but had no strong hypothesis about whether one explanation type would be better than others.

Method

Participants. 140 undergraduate students enrolled in the MTU introductory psychology course took part in the experiment for partial course credit.
Design. Participants were randomly assigned to one of four explanation conditions: Control (no explanation), examples, examples+contrast, and descriptive case information).

Procedure. Participants took part in the study via an online survey tool. They were asked to play the role of a patient experiencing stomach pain and related gastro-intestinal issues. The scenario involved approximately 40 screens, simulating repeated interactions with an AI diagnostic system ("Medibot.ai"), in which an initial diagnosis was made (IBS), and after initial success in treatment, the symptoms returned, and a rediagnosis was made (Celiac). Previous studies showed that subjective ratings (indications of persuasion) were enhanced for example and feature-based explanations in comparison to control (see Alam, 2020; Alam & Mueller, 2020), and harmed specifically during the middle rediagnosis events. However, those studies did not measure whether participants had a better understanding of the system.

Six times during the scenario, participants answered six 6-point Likert-scale questions regarding the system, related to: (1) satisfaction with the system; (2) sufficiency of explanations; (3) completeness of explanations; (4) usefulness of explanations; (5) accuracy of explanations; and (6) trust in explanations. Furthermore, we asked participants to make specific predictions about the case that would be answerable if they had an accurate understanding of the causal model relating symptoms to diagnosis. This served as a measure of knowledge and understanding, and is essentially identical to the prediction measures used in the previous simulations.

Results

Figure **Error! Reference source not found.** shows that each kind of explanation (examples, examples+contrast, and case information) was persuasive, insofar as it received higher subjective ratings than the scenario in which no explicit explanation was provided. There are three general patterns to note: (1) all subjective ratings followed the same basic pattern, suggesting that there is very little true difference to how participants answered these questions; (2) all explanation condition led to higher ratings than the control during the diagnosis weeks (time 1 and 2); (3) minimal difference exists between explanation forms; and (4) under all conditions—even the control condition—satisfaction measures rose to close to the maximum of 7 at the end of the scenario when the correct diagnosis was resolved.

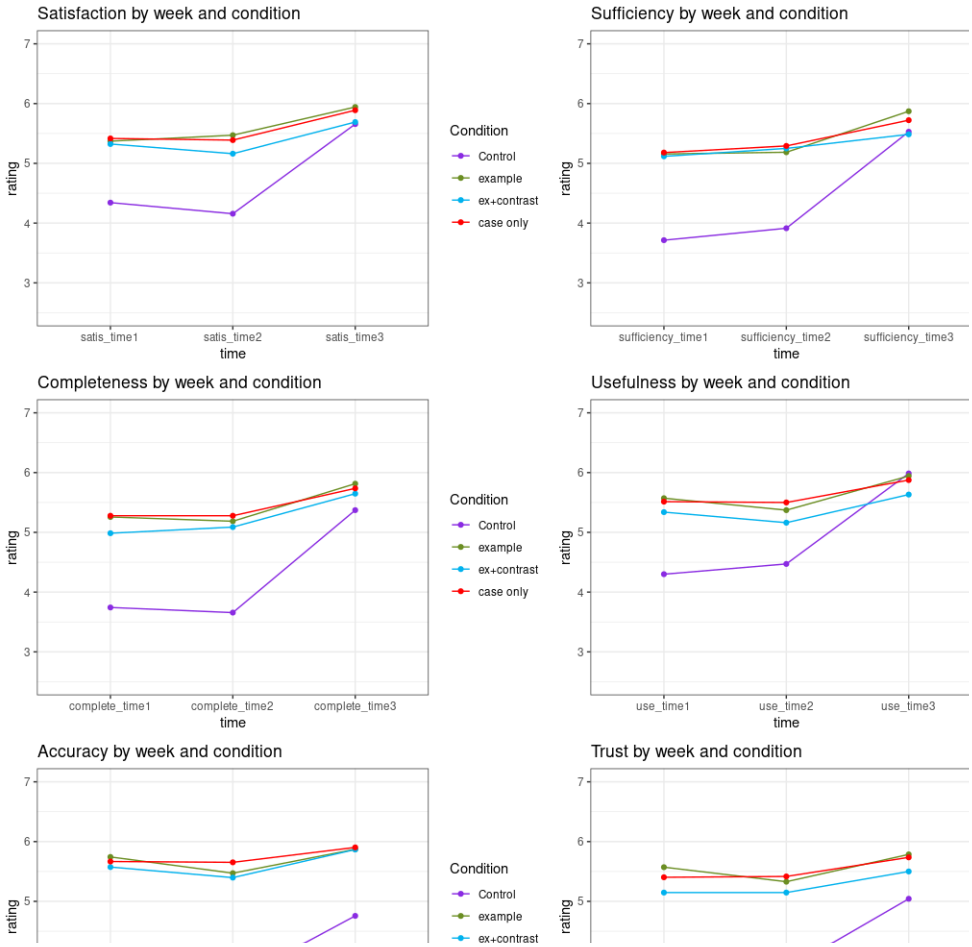

Figure 7: Prediction accuracy for three models, with black representing a pure feedback-learning model, green representing a model incorporating only historically salient features, and red representing model that mixes these two strategies. Historical saliency provides immediate and substantial learning over alternatives.

Statistical tests of the main effects and interactions showed these effects were statistically significant, with significant interactions of time and condition and main effects of condition attributable to differences from the control condition. Thus, these data provide support for the hypothesis that explanations can be persuasive for autonomous and analytic systems.

The prediction accuracy tells a different story, however. Here, different predictions were required each week, and so changes in accuracy across weeks may not stem from learning or misconceptions. A Type-III factorial repeated-measures ANOVA predicting accuracy by time and condition showed significant effects of time ($F(5,125)=11.4$, $p<.001$), condition ($F(3,125)=9.3$,$p<.001$) and the time x condition interaction ($F(15,125)=2.1$,$p<.007$). However, the effect of both condition and time: condition were not significant when the first week data were removed, so that the benefit of explanation was only apparent during the first initial week. Consequently, evidence supporting the benefit of explanation on understanding is mixed in this study, and isolated to the very initial predictions.

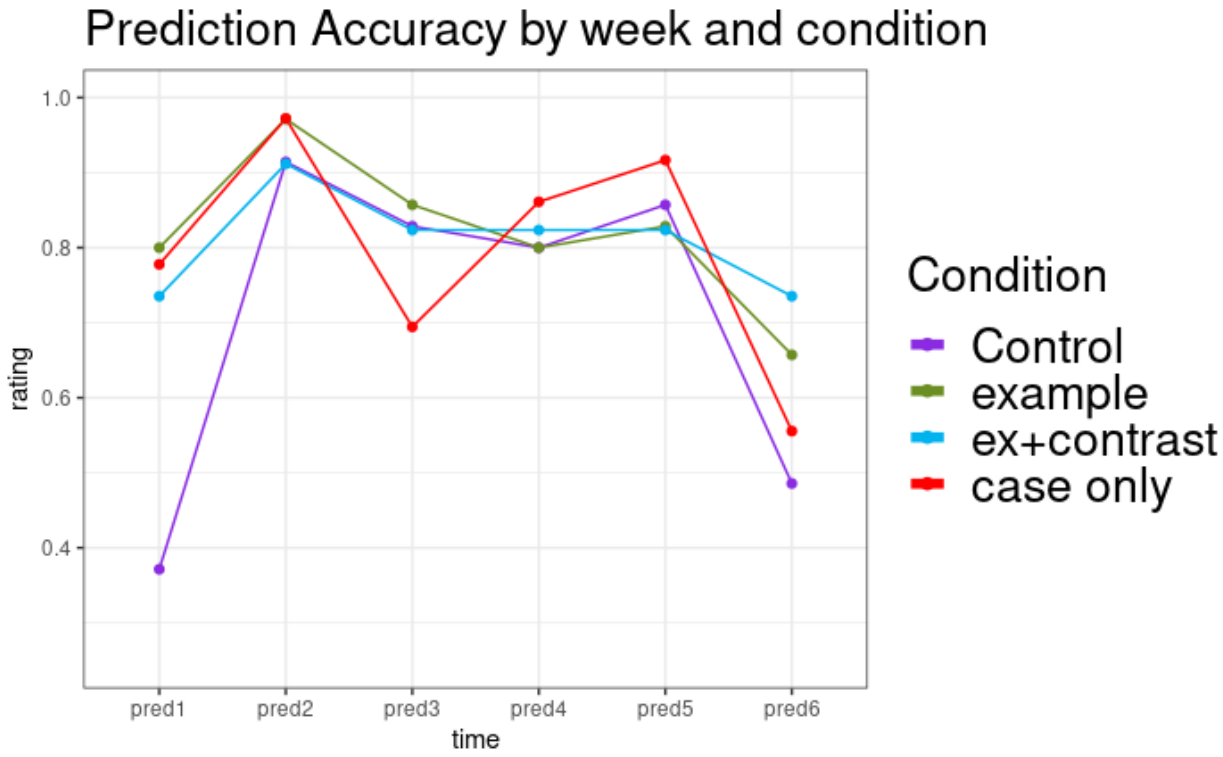

Figure 8: Prediction accuracy for knowledge questions across six simulated weeks. In comparison to subjective ratings, explanations did not confer an advantage in understanding in comparison to the control group.

As we have argued in the present section, this suggests that subjective ratings may be impacted by persuasive explanations, even when knowledge is not impacted. In the present experiment, this may partly stem from a ceiling effect—participants had a good understanding of the AI prediction

model because it was not very complicated—but it still suggests that subjective assessments might be made independently from increased knowledge.

This presents an interesting challenge: how might a computational model based primarily on developing and learning the AI model account for the findings that explanations can be more or less satisfying even when they have no impact on the underlying mental model? We will address this question in the final two sections of this report.

7. Simulating the Impact of Example-based Explanations on Persuasion

Our model of the persuasive capability of example-based explanations is heavily rooted in the JDM literature on similarity judgments. Our basic premise is that an example may serve as a good explanation for a decision if the human user can use it to rationalize the decision. What this means depends on whether or not the AI system is correct, but an explanation may be satisfactory if it helps the user understand the reasoning for making the choice or judgment, ensuring that the system is behaving consistently. If a decision can be assessed as correct (i.e., correctly labeling an image of a hammer), several reasonable examples might be persuasive:

- Images showing other very similar hammers that were also called hammers
- Images showing a variety of different hammers that were also called hammers
- Images showing similar non-hammers (hatchet or paint scraper) that were correctly identified.

In comparison, other kinds of explanations might not be persuasive in this case:

- Images of non-hammers that were labeled hammers (error to class)
- Images of hammers that were incorrectly identified (error from class)
- Images of non-examples that were incorrectly identified as other non-hammer categories (irrelevant error)

These are predicated on the goal of explanation being to justify a particular case and persuade the user it is reasonable—and not to provide broader evidence of whether the system is likely to correctly or incorrectly make a similar classification in the future. When the goal of explanation is to increase justified trust and mistrust, rather than just uniformly enhance trust, the examples that provide the appropriate persuading impact depend on whether the system is generally good or bad at making the distinction in the test case (hammer).

Thus, what serves as a desired persuasive example is highly dependent on the goals of the developer and the context and capabilities of the system. To understand this, we will examine a simplified representation in which a human must make judgments about an AI classifier that has the same information (i.e., identical feature representation) but a slightly different decision rule (see Figure 9). Here, the vertical line represents the human decision rule and the diagonal line represents the AI rule. The two triangles in the upper right and lower left represent areas of ambiguity in which the human and AI disagree about the correct response. Notice that the two agree about the majority of cases.

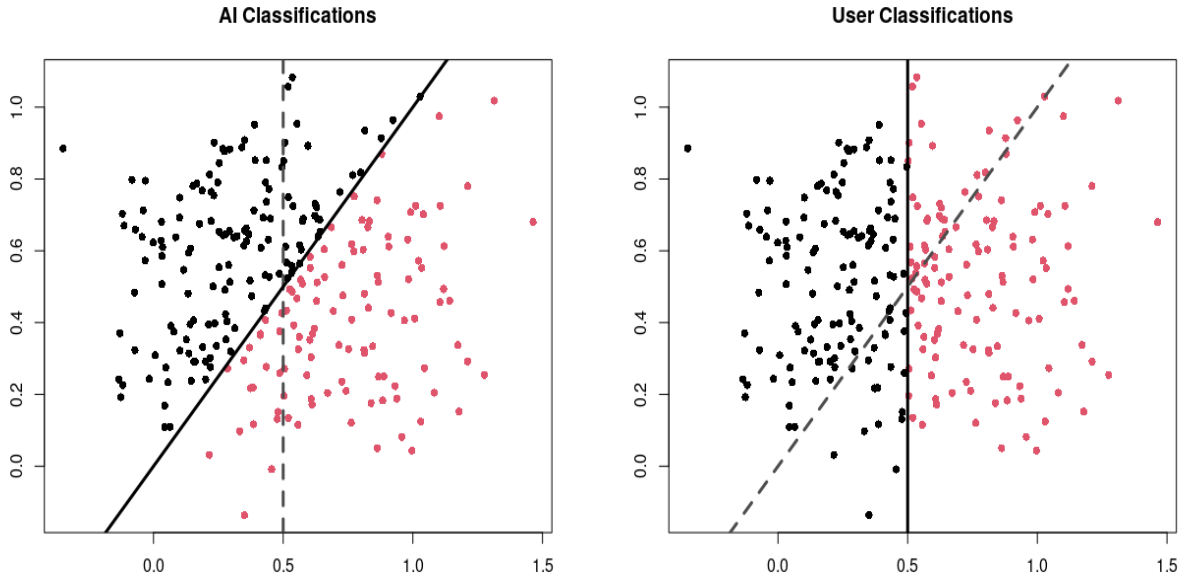

Figure 9: Depiction of a two-dimensional classification in which features of cases are identical for user and AI, but they use a slightly different decision rule.

This simple representation of knowledge can help illustrate the distinction between explanations that persuade and those that inform. For example, informative examples would help a user form an accurate mental model of the AI, and especially identify how the decision criteria differs from their own. This would allow them to correctly predict where the AI will succeed and where it will fail. This might include showing caricatures and prototypes that illustrate typical and diagnostic/extreme cases, or contrasts on either side of a decision boundary that help illustrate the distinction between two classifications.

We have already discussed the utility of examples as a method of decreasing the learning time for a particular system. Showing the user an additional example of an item with the same label or categorization during a particular trial results in double the learning in a delta-rule learning environment.

Moreover, we demonstrated that caricatures or epitomes result in faster learning compared to prototypical examples or random examples. This should make some intuitive sense, as the larger coefficients in the improper linear model (ILM) of the AI will more strongly affect the coefficients in the ILM of the user when the delta-rule learning equations are applied. This is sufficient when the user has few or no preconceived notions about the classifications produced by the AI. Consider a casual hiker in the woods using a smartphone app that classifies trees based on their leaves. If the app calls a tree a "northern red oak" and then shows an extreme example of such a tree and its leaves, the user will more quickly learn to identify the tree without the app.

In a context where the human has experience and knowledge, the task of the explanation is one of persuasion or trust-building. In these cases, we argue that the selection of the example used for

explanation must be more carefully selected in order to be "satisfying" or "acceptable" to the user. We define *acceptable* as "something that is similar to the case being explained and within the range the human considers the same category". Consider an image classifier that distinguishes between bats and birds. Classifying an image of a typical songbird as a "bird" and showing a picture of an ostrich as a justification will not be acceptable to the user and will build no trust in the system.

We can examine these relationships graphically. In the examples, we have a simple classifier that only looks at two features and places the result into one of two categories. The features are strengths of the only two characteristics used by the AI classifier. There is some small disagreement between a knowledgeable human and the AI. The human breaks the two categories at the vertical line at 0.5. The AI breaks the categories at the diagonal line. There are two primary results in this situation. Either the AI and user agree or they do not. Within those two main results are situations that should drive the example selection in order to make the explanation satisfying.

A Typical Sample

In Figure 10, we see an input that is fairly typical. The AI and human agree on the classification. It is near the prototypes for both the AI and human. Any example inside the circle is likely to be a satisfying or acceptable explanation, as they are sufficiently similar and of the same classification. Using the prototype as an explanation would be acceptable. A caricature or epitome would *not* be acceptable, as it violates the property of similarity in our definition. (Here, we are representing similarity through a simple distance measure.)

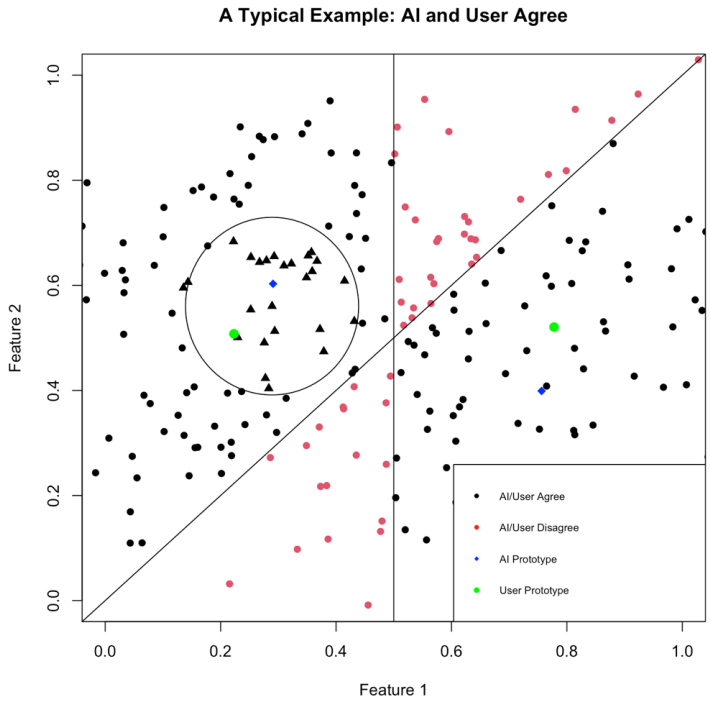

Figure 10: An input that is close to the prototypes. An example in the circle would be an acceptable explanation.

An Atypical Sample

In Figure 11, we see an example of an atypical sample and its categorization. The value of feature 1 is somewhat lower than the mean value and the value of feature 2 is somewhat higher. Again, the range of acceptable explanations is in the circle. In this scenario, neither the prototype nor the caricature are acceptable explanations as they are too dissimilar to the current sample. Consider an image of an ostrich as an atypical bird. An example of an emu or some other large bird would be acceptable; a small songbird would not. It is therefore important not to rely solely on a prototype or a caricature when selecting examples for persuasive explanation.

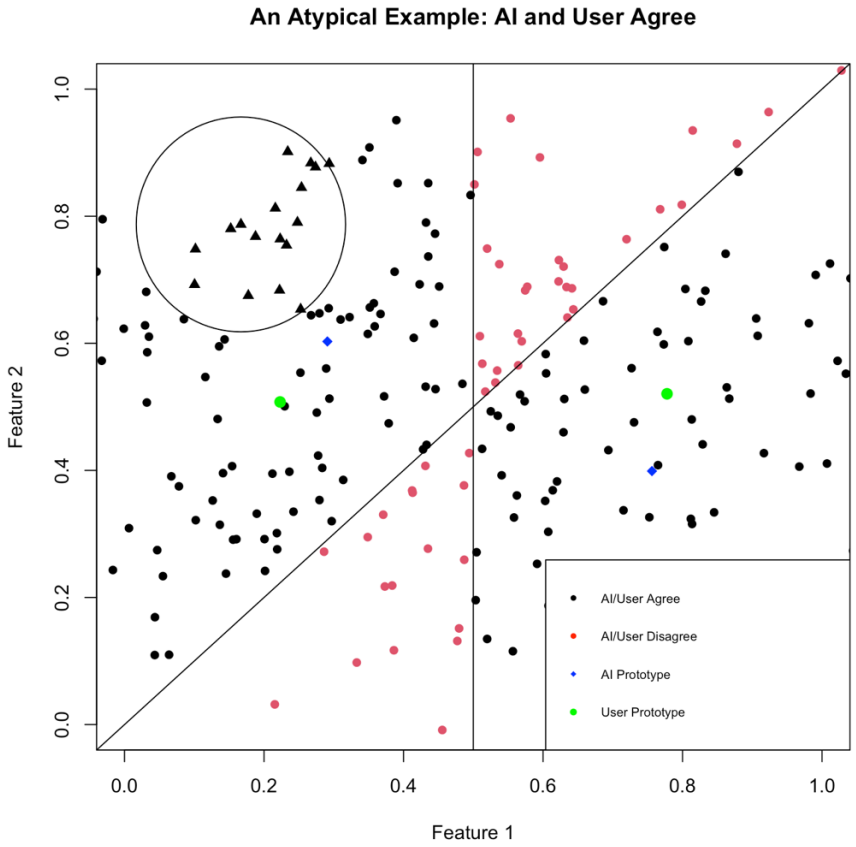

Figure 11: A categorization of an atypical sample for which a prototypical example would be a poor explanation.

Boundary Examples

In Figure 12, we see a sample that falls near the boundary of the classification. It is important to understand what has happened in this situation. The AI has been trained on a set of inputs that were all correctly labeled with the correct category. However, its mental model for future categorizations is slightly different from human's mental model. Therefore, some of those future classifications may be incorrect from the user's point of view. The AI should never be able to search its training set, find an acceptable example (based on similarity) that has the wrong category, and present that as the explanation. These should not exist. A sample from new input set

may be misclassified and presented as an explanation based on similarity. This violates the second part of our definition of acceptable: that it is of the same category.

Notice again that this boundary case is not near to the caricature or the prototype, so these would again be unacceptable examples.

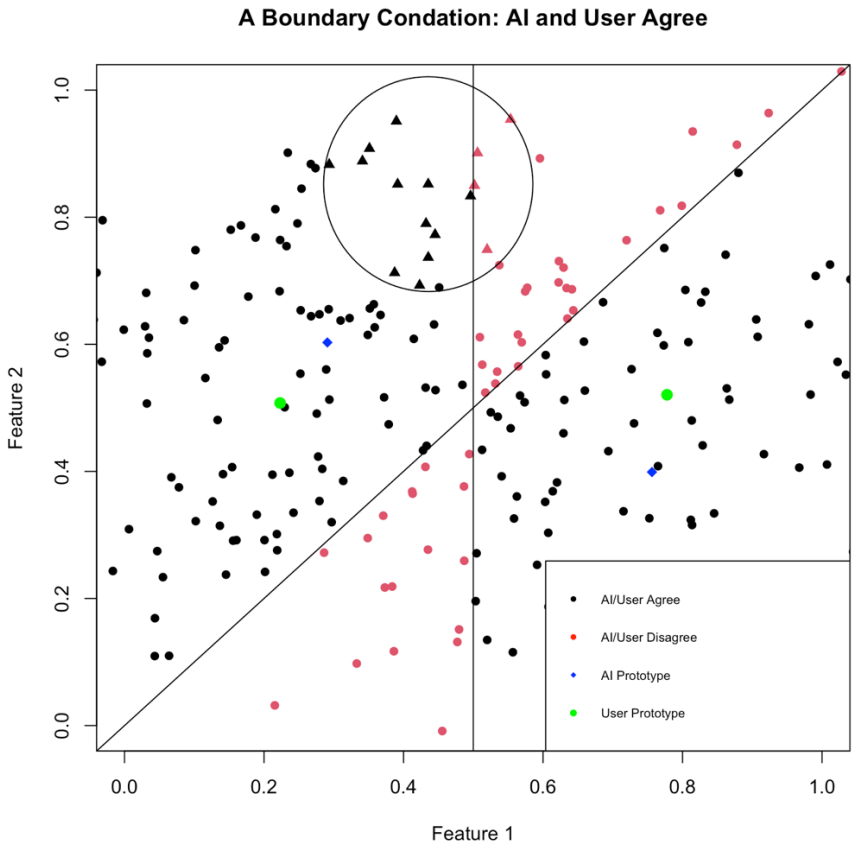

Figure 12: A boundary example where similar examples may actually be the wrong category for the user.

AI/User Disagreement with "Justified Mistrust"

As shown in all of the diagrams, there are certain cases where the trained classifier will categorize some sample differently than the user. There is actually a case here where the AI can still provide a satisfying example. Figure 13 shows a case where a sample was incorrectly classified as category A by the AI (because it is above the diagonal line) but as category B by the user (because it is to the right of the vertical line). Recall there are no examples in the training set that can be used to explain this, as every item in category B would have been correctly labeled as such during training. There are examples to the left of the vertical line that match the user's categorization and could be used to show to the user the similarities that were detected that led to the AI's classification. A satisfying example would still fit our definition, that is, classified the same as the user's classification and sufficiently similar. While this example will not convince the user that their own

classification is incorrect, but the satisfying example will be similar enough to the sample that the user will understand what led the AI to make the errant classification. Ideally, the user will be able to successfully identify other samples that will be incorrectly classified.

As before, the prototype and caricature will be unacceptable, but doubly so in this scenario. Not only will they be dissimilar, they will also be of the wrong categorization.

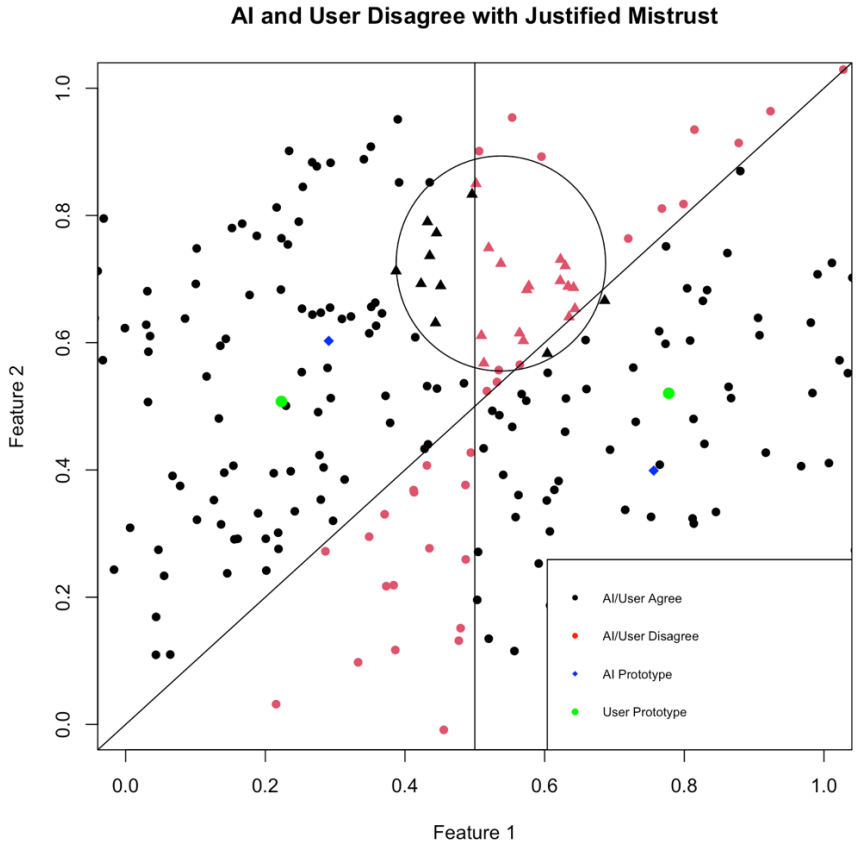

Figure 13: A satisfying explanation can still be found in the leftmost part of the circle, even when the user and AI disagree.

Disagreement with Unacceptable Examples

I don't see how this can happen. There are no items in the training set that correspond with the red dots. The AI would never provide an example of the opposite category. All other examples would be the "justified mistrust" variety – unless I'm missing something.

8. General Discussion

In this report, we describe work developing and evaluating computational models of explanatory reasoning that directly address the psychological processes and information involved in the kinds of explanations provided by XAI systems. These models inform our understanding of several important concepts within the XAI community, and connect them to formal psychological theory. We will examine several of these contributions next, discuss implications for the development of better explanatory and interpretable AI systems.

Sensemaking and the System I/System II reasoning distinction.

In our earlier tech report describing computational modeling of explanatory reasoning (Mueller et al., 2019), we focused on establishing that explanatory reasoning is a sensemaking process, and thus the goal of good explanations should be to help a user quickly reframe their misunderstandings in order to develop an accurate picture of how a system works. An important contribution of that work was establishing ways in which complementary psychological learning processes (implicit feedback-based System-1 learning and deliberate reasoning-based System-2 knowledge reconfiguration) are both important, but the reframing processes are central to what we think of as *explanation.* In the present report, we expanded this, focusing in particular on specific explanation modes (examples, contrasts, feature saliency) and goals (understanding versus persuasion). We suggest that designers of XAI systems should consider each of these elements, and ways in which their explanations can reinforce (1) both information and persuasion goals; and (2) encourage fast knowledge reconfiguration rather than slow feedback-based learning. We suspect that many potential explanations are implemented or presented to users in such a way that they encourage persuasion and not knowledge, and require substantial feedback-based learning instead of efficient knowledge reconfiguration. To address these more powerful goals and modes of reasoning, XAI developers will have to go beyond using standard algorithms and develop explanatory systems that support reasoning, including comparison, history, pattern detection, and knowledge retention.

Knowledge and information gain versus persuasion as a goal of explanation.

Hoffman et al. (2018) advocated a basic measurement model for explanations that supposes explanations promote a better understanding of the system (an accurate mental model). Yet the most common mode of explanation algorithms consist of justifications: information displays that attempt to answer the question of why a particular case produced a specific decision or choice by the AI system. This interaction mode encourages myopic local views of a system rather than systematic global views, and when coupled with evaluations focusing on satisfaction and trust, encourages the development of explanation systems that focus on *persuasion* (convincing the user the system is good) rather than information. Furthermore, natural biases of developers mean that the persuasion is likely to be biased toward selecting positive evidence that supports the accuracy of the system rather than demonstrating its weaknesses.

The models we explore only touched the surface of the complex contextual distinctions that are likely to make an explanation good for knowledge gain versus persuasion. Informational explanations typically help either establish patterns (typical cases) or boundaries (extreme cases or border cases), and to the extent they succeed, help establish an intuitive understanding of how

a system works. We suggest that persuasive explanations tend to provide an example or feature that allows a user to recognize the system is making a decision in a reasonable, systematic, or consistent fashion, even if it does not enable them to easily learn the bigger picture. We suspect that there can be substantial overlap between the content of informational and persuasive explanations, and the distinction may be about how the information is framed or explained to the user, and if interfaces permit comparison, combination, and synthesis of a number of results.

Explanations in the form of examples.

In this report, we document several models that account for how examples serve as explanations—both informative and persuasive. But in order to reason about examples, our models that use examples incorporate an implicit model of how the examples were generated. Many XAI systems using examples are vague or shield a user from the process of how an example is generated, but examples may only be useful if a user understands what they do and do not represent. In addition, we have assumed that users can naturally compare positive and negative examples in order to understand differences between features and outcomes. This assumption is likely to be reasonable for example cases that are very similar, differing by only one or two features. As the differences grow in size and number, this 'example arithmetic' is unlikely to be possible, and XAI developers relying on examples should be aware of this when selecting examples, so that they can either provide guidance about the relevant feature differences or select examples that differ minimally along dimensions that matter.

Explanations in the form of saliency/highlighting.

Similarly, we documented several models that hypothesize ways in which feature highlighting can provide explanation. One substantial lesson from these models is that although saliency alone may provide reasonable cues for persuasive explanations, it makes for poor intuitive system-1 learning cues. However, our models suggest that by allowing or encouraging memory for a number of historical cases, rapid knowledge reconfiguration might occur. This makes the prediction that one-off saliency examples (and by analogy, one-off simple example-based contrasts) make for poor informational explanations, but across even a small handful of examples a reasonable mental model might be induced---provide the user is allowed the time and freedom to explore these patterns.

Saliency and heatmaps are some of the most common XAI explanations, but they have a variety of implementations and interpretations. Our modeling suggests that applying an importance filter to select a single decisive feature to highlight may be effective, because it makes the relationship clearer and more memorable, as long as a variety of examples can be deployed to provide the user with numerous cases to form a larger mental model of the system.

Future applications for computational modeling of explanation.

To support common ground in explanation, many researchers have advocated developing user models that help establish what the user understands, and what they are prepared to learn. We suggest that simplified versions of the models presented here are ideal for representing this user model. Because the learning models provide several different benchmarks for how well a user

might perform given different assumptions about their learning approach (i.e., system-1 vs system-II), comparing a user's performance to a computational cognitive model's performance on the same cases might help establish if they need special instructions to try to reframe their understanding, or specific examples that might help them see particular patterns. Thus, just as a good user must develop an accurate mental model of the AI system in order to perform well, a good XAI system needs to develop an accurate mental model of the user in order to explain well. Although it will require substantial investment in order to make accurate computational cognitive models that represent the impact of specific explanations, we believe the present work is a foundational step toward achieving that goal.

4. References

Alam, L. 2020. Investigating the Impact of Explanation on Repairing Trust in Ai Diagnostic Systems for Re-Diagnosis (Publication No. 28088930) [Master's Thesis, Michigan Technological University].

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC medical informatics and decision making,* 21(1), 1-15.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1322–1340.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 932–945

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of educational research, 70(2)*, 181-214.

Estes, W. K. (1986). Array models for category learning. *Cognitive psychology*, *18*(4), 500-549.

Fogg, B. J. (1998, April). Captology: the study of computers as persuasive technologies. In *CHI 98 Conference Summary on Human Factors in Computing Systems* (p. 385).

Hoffman, R. R., Mueller, S. T., Klein, G., & Littman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kenny, E.M., & Keane, M.T. (2020). On generating plausible counterfactual and semi-factual explanations for deep learning. arXiv preprint arXiv:2009.06399.

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems*, *21*(5), 88-92.

Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768-4777).

Mueller, S. T. (2020). Cognitive Anthropomorphism of AI: How Humans and Computers Classify Images. *Ergonomics in Design: The Quarterly of Human Factors Applications*, 28(3), 12-19.

Mueller, S. T., Alam L., Mamun, T., Tan, Y., Hoffman, R. R., & Klein, G. (2019). A computational model of explanatory reasoning: Foundations of explanation in sensemaking. *DARPA XAI Program Tech report.*

Mueller, S. T., Tan, S. Y. Y., Linja, A., Klein, G., & Hoffman, R. R. (2021). Authoring guide for Cognitive Tutorials for Artificial Intelligence: Purposes and the Methods Development. *DARPA XAI Technical Report.*

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*(4), 700.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. *Essays in honor of William K. Estes*, *1*, 149-167.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55-89.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, 36(3), 212-218.