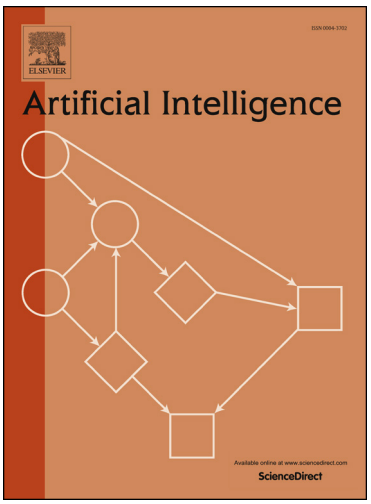
Journal Pre-proof



Explaining Black-Box classifiers using *Post-Hoc* explanations-by-example: The effect of explanations and error-rates in XAI user studies

Eoin M. Kenny, Courtney Ford, Molly Quinn and Mark T. Keane

PII: S0004-3702(21)00010-2
DOI: https://doi.org/10.1016/j.artint.2021.103459
Reference: ARTINT 103459

To appear in: *Artificial Intelligence*

Received date: 27 February 2020
Revised date: 22 December 2020
Accepted date: 21 January 2021

Please cite this article as: E.M. Kenny, C. Ford, M. Quinn et al., Explaining Black-Box classifiers using *Post-Hoc* explanations-by-example: The effect of explanations and error-rates in XAI user studies, *Artificial Intelligence*, 103459, doi: https://doi.org/10.1016/j.artint.2021.103459.

Explaining Black-Box Classifiers Using *Post-Hoc* Explanations-by-Example: The Effect of Explanations and Error-Rates in XAI User Studies

Eoin M. Kenny[a,b,c,*], Courtney Ford[a], Molly Quinn[a,b], Mark T. Keane[a,b,c]

[a] *School of Computer Science, University College Dublin, Dublin, Ireland.*
[b] *Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland.*
[c] *VistaMilk SFI Research Centre, University College Dublin, Dublin, Ireland.*

Abstract

In this paper, we describe a *post-hoc* explanation-by-example approach to eXplainable AI (XAI), where a black-box, deep learning system is explained by reference to a more transparent, proxy model (in this situation a case-based reasoner), based on a feature-weighing analysis of the former that is used to find explanatory cases from the latter (as one instance of the so-called *Twin Systems* approach). A novel method (COLE-HP) for extracting the feature-weights from black-box models is demonstrated for a convolutional neural network (CNN) applied to the MNIST dataset; in which extracted feature-weights are used to find explanatory, nearest-neighbors for test instances. Three user studies are reported examining people's judgements of right and wrong classifications made by this XAI twin-system, in the presence/absence of explanations-by-example and different error-rates (from 3-60%). The judgements gathered include item-level evaluations of both correctness and reasonableness, and system-level evaluations of trust, satisfaction, correctness, and reasonableness. Several proposals are made about the user's mental model in these tasks and how it is impacted by explanations at an item- and system-level. The wider lessons from this work for XAI and its user studies are reviewed.

Keywords: Explainable AI, Trust, User Testing, Convolutional Neural Network, Case-Based Reasoning, Deep Learning, k-Nearest Neighbors

1. Introduction

Recent impressive advances in Artificial Intelligence (AI) have encountered significant challenges in concerns about the interpretability of AI systems (especially deep learning ones), the imposition of new data regulations (such as the GDPR [1, 2]), and public disquiet about the fairness of these systems in making life-impacting decisions [3]. The development of methods that provide explanations – so-called eXplainable AI (XAI) – is seen as a panacea to these problems, on the working assumption that people will be assured when the inner workings of these systems are revealed [4], that more transparent AI systems will be auditable (to meet data regulations [2]), and that these methods will help to reveal algorithmic bias and unfairness [5, 6]. There is now a plethora of XAI techniques that propose to "explain" black-box models; for instance, black-box classifiers, such as convolutional neural networks (CNNs) for image-analysis, are now "explained" by saliency map visualizations of the network [7], or by re-casting the network as a rule-based decision tree [8], or by finding class-level prototypes for the network's predictions [9], and so on. However, the evaluation of these XAI techniques in carefully-designed user studies has lagged behind technique development [10]. Most papers fail to report user studies, while others report pilot studies that are inconclusive, or indeed, larger studies with methodological flaws (e.g., poorly controlled materials, incorrect statistical analyses). This state of affairs raises the appalling vista that many popular and highly-cited XAI methods in the literature may

*Corresponding author
Email addresses: eoin.kenny@insight-centre.org (Eoin M. Kenny), courtney.ford@ucdconnect.ie (Courtney Ford), molly.quinn@insight-centre.org (Molly Quinn), mark.keane@ucd.ie (Mark T. Keane)

not actually explain anything at all, as they may provide explanations that users cannot understand, find too complex, or indeed actively mislead them (e.g., if they elicit inappropriate trust). In this paper, we try to address this imbalance by evaluating a *post-hoc* explanation method for black-box, CNN classifiers, in three carefully-designed user-studies that attempt to assess the validity of this method. As such, we hope that the paper provides some guidance on how XAI user-studies might be properly conducted, while indicating some of the issues, complexities, and pitfalls that can arise in such tests.

Broadly speaking, XAI methods divide into those that attempt to improve *interpretability* by (i) directly conveying the workings of the model (so-called *transparency*), or (ii) justifying how/why the model arrived at its predictions (so-called *post-hoc explanation*) [11]. One type of *post-hoc* explanation involves explanation-by-example, where instances or examples are provided as evidence for why the model produced a given prediction. For example, a user interacting with a property-loan system could be told *you were refused the loan because your profile was the same as person-X and they were refused a loan for the same amount.* Or, a user debugging an image-classification system could be told *the system classified this test image as a "0", because it looks very like this other image that was labelled as a "0"* (even though the original test instance has a ground-truth of "6", see Fig. 1). Given the long-standing use of case-based explanations in law, business and medicine, it has been repeatedly-argued in AI – and specifically, in case-based reasoning (CBR) – that explanations-by-example are intuitively and easily understood by people [12, 13, 14, 15, 16, 17]. Additionally, although this research area also suffers from a dearth of user-testing (e.g., one review [18] found <1% of papers in this area has user evaluations), it has been shown that such explanatory cases can improve people's performance/understanding of an AI system (see e.g., [19, 20]), though the method may not always be the best explanation strategy (see e.g., [21]).

We have advanced a framework for *post-hoc* explanation, called the *twin systems* approach [22, 18], that is tested in the user studies reported here. This approach has been applied to a wide range of artificial neural networks (ANNs), from multi-layered perceptrons (MLPs) to convolutional neural networks (CNNs), for many different domains [22]. Stated simply, this approach extracts the feature-weights from an ANN and applies them to a twinned k-NN model to find the nearest-neighboring explanatory cases for a prediction. Kenny and Keane [22] explored a selection of different feature-weighting methods and found a contributions-based method to be the most accurate in reflecting the ANN's function. It is a variant of this method – called COLE-HP – that is tested here in an ANN-CBR twin-system involving a CNN classifier operating on the popular MNIST dataset [23]. Specifically, in the context of a model-debugging task [24, 25], we record users' evaluations of the classifier's performance with/without *post-hoc* explanations while varying the error-rates of the system; specifically, users are asked to rate the "correctness" and "reasonableness" of the model's predictions, along with assessing their overall trust and satisfaction with the classifier [26].

A critical aspect of these studies is that they test the actual outputs of the model, the "real" errors it produces, rather than artificial proxy outputs (which can raise validity issues; see [20, 27]). So, in the three experiments reported here, we evaluate how people's understanding of the predictions of a black-box classifier and their overall assessment of that system (i.e., trust and satisfaction) are impacted by:

- The provision of *post-hoc* explanations-by-example; namely, the effects of *explanation*.
- Their experience of the model making errors; namely, the effects of *error-rates*.
- Potential interactions between explanation and error-rate effects.

1.1. Outline of Paper

In Section 2 we discuss background to the current work, our notion of explanation, and previous work on *post-hoc* explanations in AI systems. In addition, we also review the relevant literature related to perceptions of errors, algorithmic aversion, and model-debugging tasks involving the MNIST dataset. In Section 3, we review the the novel feature-weighting method we have proposed, within the twin-systems framework for CNNs, and show how it generates *post-hoc* explanations-by-example for the MNIST domain. In Section 4 we present our experimental paradigm used here to test the impact of these explanations and the variables explored. In Section 5, we advance a new psychological analysis of people's mental models for XAI user-studies – the Tricorn User Model – and show how this model applies to the model-debugging task used in

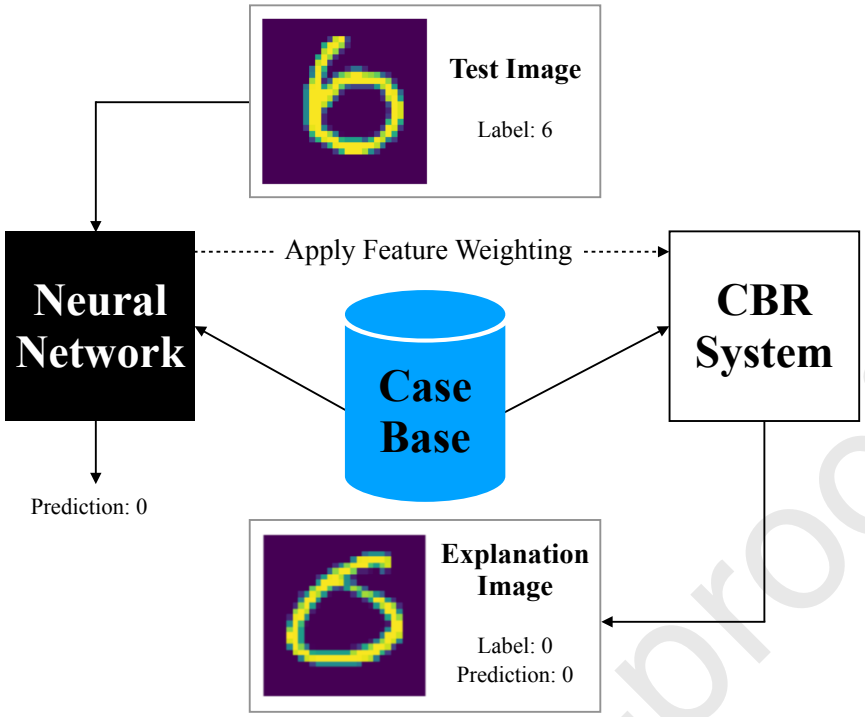


Figure 1: The Twin-Systems Explanation Framework: A deep learning model produces a misclassification for a test image in MNIST, wrongly labelling a "6" as a "0". This prediction is explained by analyzing the feature-weights of the network for that prediction and applying these to a twinned k-NN to retrieve a nearest neighbor to the test-image in the training set. This explanatory image shows that the model relied on an image of a "0" that looks very like a "6" to make its prediction of a "0". As such, though the model is "incorrect" in its classification it performs somewhat "correctly" and, possibly, "reasonably" in the use of the labelled-data it was given.

the current studies. Sections 6-8 report three user studies examining the effects of *post-hoc* explanation and error-rates in this debugging task. Finally, Section 9 discusses the results of the current studies and tries to tease out any possible lessons going forward.

2. Related Work: *Post-Hoc* Explanation, Trust in Systems, and Debugging Tasks

The current work brings together diverse strands of research in Artificial Intelligence, Cognitive Science, and Cognitive Psychology (see [28] for the broader canvas). It advances a model-agnostic method for *post-hoc* explanation of black-box systems, makes novel proposals on mental models in XAI, and then tests these proposals in three user-studies involving a classifier-debugging task. This work draws on XAI methods using *post-hoc* explanation, user-studies of explanation, trust in automated systems, and algorithmic aversion. In the current section, we sketch the relevant findings that converge on the current work. So, we consider related work from five interlocking areas:

- *Post-Hoc Explanation in XAI*: Methods for *post-hoc* explanation in XAI.
- *Factual, Post-Hoc Explanations*: Twin-system methods for computing factual, *post-hoc* explanations.
- *User Studies on Post-Hoc Explanation*: Key findings from user-studies on *post-hoc* explanation.
- *Trust in Automated Systems*: Findings on people's trust in automated systems and their perceptions of system-error (e.g., *algorithmic aversion*).

- *User Studies on Debugging AI*: Findings from tasks where people debug AI classifiers involving the MNIST dataset.

2.1. Post-Hoc Explanation in XAI

As an area XAI has many issues, but foremost amongst these, perhaps, is some clarity on what the term "explanation" actually means. Several recent XAI reviews have pointedly noted the lack of clear definitions for the notions of explanation, interpretability, and transparency [29, 27, 12, 30, 31, 13], echoing long-standing discussions in CBR [14, 15], recommender systems [32], philosophy [33, 34, 35], and psychology [36]. While the exact meaning of these terms remains a matter of debate, these reviews make useful taxonomic distinctions. For example, Sørmo *et al.* [14] emphasise the distinction between explaining how the system reached some answer (what they call *transparency*), and explaining why the answer is good (*justification*). Recently, this distinction has been echoed by dividing *interpretability* into (i) *transparency* (or *simulatability*) which tries to reflect how the AI system produced its outputs, and (ii) *post-hoc interpretability* which is more about why the AI did what it did, providing some after-the-fact rationale/evidence for system outputs [12, 13]. With respect to the interpretability of deep neural networks (DNNs), Gilpin *et al.* [37] propose using a proxy model "that behaves similarly to the original model, but in a way that is easier to explain, or by creating a saliency map to highlight a small portion of the computation that is relevant" (p. 3); they identify linear proxy-models (e.g., LIME by Ribeiro *et al.* [5]) and decision trees (e.g., Frosst and Hinton [8]) as common options for such proxy models.

Post-Hoc Explanation has been further sub-divided into (i) textual explanations of system outputs, (ii) visualizations of learned representations or models (e.g., saliency maps [7]), and (iii) explanations-by-example (see [12, 38, 39]). So, *post-hoc explainability* can be cast as a type of "explanation-by-justification", an after-the-prediction explanation step where some evidence/information/visualization is given to elucidate the predictions made by the AI system. Furthermore, recently, *post-hoc* explanation-by-example has been split into several different example-types, (i) *factual examples*, the classic case-based explanation method, where the user is presented with a similar case(s) (e.g., "you were refused the loan because you profile is similar to person-x who was also refused the loan"), (ii) *counterfactual examples*, where the user is presented with contrasting cases that change the class of the prediction (e.g., "if you had a higher salary, you would have the profile of a person who got the loan"), or, indeed, (iii) *semi-factual examples*, where the user is presented with positive cases that do *not* change the class of the prediction [40] (e.g., "even if you had a higher salary, you would still not have the profile of a person who got the loan"). Most research in the literature has focused on factual *post-hoc* explanations (see next sub-section), though there is a rapidly expanding interest in techniques to compute counterfactual explanations [1, 41, 30, 42, 43, 44, 45, 40]; however, there is very little research on semi-factual explanations (but see [40] for the only current example in the AI literature and [46] for a similar approach called *a fortori* reasoning). In the next sub-section, we briefly review the methods proposed to compute factual, *post-hoc* explanations.

2.2. Factual, Post-Hoc Explanation-by-Example: A History of Twin Systems

The current work focuses on factual, *post-hoc* explanation where some explanatory example or case is provided to justify why a given prediction was made by an AI classifier. This type of explanation is often called *justification*, and some distinguish it from *explanation proper*, though traditionally is is one of the longest-standing XAI techniques in the literature [5, 7, 16, 14]. For years, proponents of CBR have argued that k-NN models provide intuitive and plausible examples to explain system predictions [12, 47, 48, 13, 14, 15, 16, 17]. Two decades ago, it was proposed that k-NN models could be used as transparent, white-box proxies for opaque, black-box neural networks (i.e., multi-layered perceptrons) by analyzing the feature-weights of the latter and applying them to the former to find explanatory nearest-neighbors within the training-set for test-instances (see Fig. 1; [49, 50, 51, 52, 53]). Kenny and Keane [22] generalized this idea in their twin-system approach, and applied it to deep learning models (i.e., CNNs), arguing that there are many hybrid-system options for pairing black-box and white-box techniques for XAI; for example, decision trees, linear models, and CBR systems have all been used as interpretable counterparts for ANNs in XAI [8, 54, 50, 5, 55, 9]. Kenny and Keane [18, 38, 22, 56] focused on the pairing between ANNs and

CBR – so-called ANN-CBR twins – and examined a selection of feature-weighting techniques to determine those which performed best. Here, we briefly summarize the prior work on feature-weighting techniques for explanation (for a detailed review see [56]).

Historically, several versions of the feature-weighting idea have previously been explored in the AI literature, going back as far as the 1990s. A South Korean group explored several feature-weighting schemes for describing MLPs (e.g., *sensitivity*, *activity*, *relevance*, and *saliency*) to find the best one to apply in k-NN to retrieve explanatory cases. Across multiple papers and tests in several domains they found that *sensitivity* and *activity* tended to do best [51, 52]. This group also advanced an important distinction between *global* and *local* feature-weighting schemes, where global methods take the input space as isotropic, deriving a single ubiquitous feature-weight vector for the entire domain, and local methods calculate a specific set of weights for each test query [49]. In North America, Caruana *et al.* [53] described a local weighting method for MLPs which used a query-case's hidden-layer activation-vector (i.e., its latent-feature representation) to find explanatory cases by computing the Euclidean distance between the query-vector's latent features and all training-cases, to find explanatory cases with the most similar latent features. This approach was, arguably, more precise than that of the Korean group, as it takes full advantage of each training and test case's individual representation. Crucially however, this method does not consider each latent-feature's weighted *contribution* to the classification when searching for explanatory cases (which we subsequently found to be important [22]). Caruana *et al.* [53] may also lose some interpretability by considering the similarity between latent representations, as opposed to the input features, which are more transparent. Finally, in Europe, a CBR group in Ireland (see [57, 58, 59, 19]), explored other local feature-weighting methods. Nugent and Cunningham [59] built an artificial local dataset around a query by systematically perturbing the features of it before querying labels for these artificial cases in the MLP. They then proceeded to build a local linear model (similar to LIME [5]) using this new local dataset; the coefficients of the linear model were then used to weight k-NN searches for explanatory cases. Significantly, this group also performed user tests of this method (see next section).

More recently, there has been a growing but distributed interest in this approach to interpretability for black-box systems. For example, Biswas *et al.* [60] revisited the Korean work to elaborate its application to unbalanced datasets. Kenny and Keane [22] have revisited the *sensitivity* method [50] and compared it to more recent methods (e.g., DeepLIFT [61] and LIME [5]). Similarly, in the the deep-learning literature, Frosst and Hinton [8] distilled a CNN into a decision tree for the purposes of improved interpretability; however, they do not propose to use the tree as a proxy model, rather they use it as an independent model, for both accuracy and interpretability. Also, Papernot and MacDaniel [62] explain CNNs using nearest neighbors from the training data in a technique called Deep k-Nearest Neighbors (DkNN). However, as in [53], they use the penultimate latent-layer activations in the ANN to fit a k-NN for explanations, without doing feature-weight mapping. In the next sub-section, we consider the extent to which this computational research on *post-hoc* explanation has been tested in user studies, and whether there is any evidence that factual, example-based explanations actually work.

2.3. Post-Hoc Explanation-by-Example: User Studies

The current work concerns itself with the user evaluation of a computational method for factual, *post-hoc* explanations. This work arises out of case-based reasoning research which, historically, assumes that this method is psychologically-intuitive and "naturally" explains predictions [16, 63, 14]. However, even in the CBR literature, this assumption has not been extensively tested in user studies (e.g., in a review of 1000+ CBR papers on explanation [18] found <1% reported user tests); however, there is a growing sense that this user-testing deficit is beginning to be addressed (see e.g., [30, 44, 21, 43, 42, 64, 65, 25]).

So, an increasing number of studies have attempted to directly test *post-hoc* explanations. Nugent and Cunningham [59], were amongst the first to do this, reporting a small user-study on the use of case-based explanations in a blood-alcohol-prediction domain [57, 58, 59, 19, 46]. They found that factual and semi-factual cases improved a user's perception of how correct a prediction was for a simple binary classification model. Similarly, in an unsupervised learning domain, Kim *et al.* [55] found that case-based prototypes helped users understand clusters better. Cai *et al.* [66] found that case-based explanations made users feel they had a better understanding of a system, and its capabilities to be of a higher quality.

More recently, in the last year, there have been a clutch of important studies on *post-hoc* explanations. Dodge *et al.* [21] tested for the effects of four distinct *post-hoc* explanation strategies on a user's global and local fairness evaluations of a machine learning model. They showed that counterfactual-explanation strategies did best and that the case-based strategy lagged. However, they summarized case information statistically (e.g., "the training set contained 10 individuals identical to X, 60% of these re-offended"), rather than presenting specific individual cases (as is done in most studies). They also found that case-based explanations seemed to generally make users feel a decision was less fair, but that local explanations of specific query instances were more effective than global ones to expose fairness issues. Yang *et al.* [67] tested user trust in example-based explanations of a classifier's predictions for images of tree leaves (N=33), finding that specific visual representations improved trust in the system (specifically, "appropriate trust"); their classifier had an accuracy of 71%, but notably, their participants were perhaps less expert (i.e., not botanists), and trust was assessed item-by-item.

Finally, Buçinca *et al.* [20] reported two experiments involving the influence of example-based explanations on an AI-model making predictions about fatty-ingredients from pictures of food; they provided explanations in two different modes, based on multiple cases (four photos of similar food dishes) and a single case with highlighted features (photo of one food-dish with identified ingredients). They found that the provision of these explanations improved performance on the fat-estimating task and that the different modes had different effects on system-level measures (of trust and satisfaction). Specifically, they found that case-based explanations significantly impacted a participant's trust in the system. Importantly, this latter study is very similar to the current ones, though the domain and task differ, in that it explains system predictions using image-based, nearest-neighbors to the test query.

In summary, these older and more recent studies show that examples and nearest-neighbors can indeed function as "explanations" to support people's use of AI systems. Specifically, they provide evidence that *post-hoc*, factual examples can help to explain the predictions of an AI system and support improvements in user performance. However, it should also be said many of these studies suffer from methodological flaws in the control of materials used, providing consistent information to users, and properly manipulating the provision of explanations. So, some of these findings deserve closer scrutiny and examination. In the present studies, we attempt to better control such factors in a classifier debugging-task. Before that however, the next section considers a related but different literature on people's trust in automated systems to see what it suggests for the current tests on error-rates.

2.4. Trust in Automated Systems: The Effects of Error

As well as considering the effects of explanations, the current work also tests for the effects of error-rates; that is, how people's experience of errors being made by the AI system impacts their perception of it. Traditionally, the relevant literature on this topic tends to focus on how error-rates impact people's trust in automated systems [68, 26, 69]. Several studies have assessed the extent to which people's experience of errors impacts their evaluations of an AI system. deVries *et al.* [70] showed that people's experience of low (20%) versus high (60%) error-rates in a route planner significantly impacted trust; for related work see [71, 72, 73]. Dzindolet *et al.* [74] presented people with items from a pseudo-computer program that supposedly detected camouflaged soldiers in 200 pictures; they found that, when people thought the program was making errors, they immediately started to distrust it, unless they received explanations as to why those errors arose. Indeed, the work of Dietvorst *et al.* [75] on algorithm aversion has shown that any forecasting error made by an automated system leads people to prefer their own or other people's forecasts, even when the automated system is demonstrably better than the human forecasters. Ribeiro *et al.* [5] reported a user study with graduate students in machine learning (N=27) that showed a lowering of trust for a "bad classifier" that made 2 incorrect classifications out of 10. Lastly, as an aside, it is also interesting to note that these findings in AI are mirrored in a great many studies in the field of human factors, on the effects of error-rate on performance [76].

So, overall, this research seems to show that people have a low tolerance for error in AI systems. Repeated studies show that error-rates of 20% significantly impacts people's trust in a system, and some studies suggest that "any" occurrence of errors undermines people's view of the model's competence.

2.5. Debugging a Classifier: User Studies Using MNIST

The current work explores a task in which people are, essentially, instructed to debug a black-box classifier. Several user studies have evaluated AI classifiers using such debugging tasks, to either assess the adequacy of datasets and/or the AI model itself. Often, these tests are carried out in domains where, arguably, people have high levels of expertize (e.g., in the deciphering of hand-written numbers [24]). While on the face of it, such debugging tasks may appear to be somewhat specialized, they do represent a key concern in the use of opaque AI models; namely, whether the dataset being used by the model is adequately annotated and/or whether the model is using the data in an acceptable way (see e.g., Ribeiro *et al.*'s [5] study involving LIME-explanations). So, broadly speaking, the literature takes this debugging task to be representative of what end-users will be doing when they are considering key dimensions of a model's classification competence.

As in the current studies, several previous studies have specifically assessed explanations for classifiers operating over the MNIST dataset, using this debugging task [77, 24, 25, 78]. Bäuerle *et al.* [24] built an interface to present misclassifications of MNIST items to users, grouping them together to speed-up the judgement task to aid model debugging, but their user study was really just a pilot (N=10). In their study, a small user sample was tasked to debug (resolve) several types of errors in the MNIST labelled dataset (e.g., interpretation errors, similarity errors). Their results showed that the user-relabeled datasets considerably reduced the number of incorrect classifications produced by the system in subsequent training, showing the feasibility of a human-in-the-loop approach. The XAI DARPA program reports several groups (notably Rutgers University) that also carried out initial user evaluations of AI classifiers using MNIST. Glickenhaus *et al.* [25] provided a preliminary report on these evaluations, though the accounts are not very detailed; notably however, they report that explanations only impacted errors, not correct items (a finding we return to here). Also, the Rutgers group assessed the utility of explanation-by-example at three levels ('most helpful, unhelpful, and random') on the CAFÉ and MNIST datasets, and found that feature highlighting explanations assist users in detecting errors, as well as increasing their mental model understanding [25]. Finally, Ross and Doshi-Velez [78] had users (N=11) make plausibility and reasonableness judgements about the robustness of different deep learning methods to adversarial attacks involving the MNIST dataset in a pilot study; though this was not about explanation or trust *per se*. However, they did ask users to make first- and second-choice predictions of the MNIST image that they expected a "reasonable classifier" to make, followed by a qualitative assessment of differences between classifications made by the several system-defenses (or explanations).

However, it is hard to gain significant insights from this literature; many of the studies are unpublished, are reported in insufficient detail (e.g., no description of method, materials, or no N given) and, indeed, may be methodologically questionable (e.g., studies with very low Ns between 2 and 10 participants). As such, we believe that the present studies provide a significant novel contribution, methodologically and substantively, to this literature. In the next section, prior to moving to the present user tests, we describe a novel feature-weighting method used to instantiate the CNN-CBR twin-system that was used in the current user tests.

3. Twin Systems: Feature Weighting to Find *Post-Hoc* Explanations

The present user studies assess a twin-system – pairing a CNN with a k-NN – to find factual, *post-hoc* explanations-by-example for predictions made by this black-box classifier using the MNIST dataset (i.e., an instance of an ANN-CBR twinning). Kenny and Keane [22] explored a selection of historical feature-weighting methods for this problem as well as developing a novel, contributions-based method – called COLE (Contributions Oriented Local Explanations) – and showed that this method generated the best twinnings across many different domains. One novelty of this work was its extension to deep learning models, specifically CNNs. They showed how to extract feature weights from CNNs with several hidden fully-connected layers in their output classifier. However, most recent CNN architectures have moved away from these types of outputs because of their tendency to overfit, caused by the massive number of parameters required to train the model [79]. Additionally, the techniques used required saliency maps that rely on heuristics in the feature-weighting calculations. Here, we rectify these issues by proposing a precise computational method

Algorithm 1 COLE-HP Feature-Weighting Extraction in Classification CNNs

Input: $D \in \mathbb{R}^m$, training data.
Input: $g(.)$, feature extractor half of CNN to explain.
Input: $f(.)$, classifier half of CNN to explain (a linear classifier).
Input: $W \in \mathbb{R}^{c,n}$, the weight matrix in f, where c is the number of output classes and n the number of extracted features in layer X (here the penultimate CNN layer after applying GAP).
Output: C, the contributions of features to the training data predictions for the CNN, these are used to fit a k-NN classifier for the CBR twin.

$C \leftarrow EmptyArray \in \mathbb{R}^{m,n}$
for i `in range` D **do**
    $I_i \leftarrow D[i]$
    $\hat{y} \leftarrow f(g(I_i))$
    $\vec{x_i} \leftarrow g(I_i)$
    $C[i] \leftarrow \vec{x_i} \odot W[\hat{y}]$
end for
return C

to implement our previous weighting algorithm COLE on these CNNs without relying on saliency maps or heuristics. In the current section, we briefly describe this technique and how it was applied to the MNIST dataset, to generate explanations for the right and wrong classifications used in the present user studies.

3.1. Contributions Oriented Local Explanations (COLE)

Our previous work has shown how COLE can successfully weight a CNN system with multiple fully-connected layers in its output classifier [22, 80]. Here, the technique is briefly reviewed and expanded upon to be applicable to recent advances in CNN architectures such as ResNets [81], which typically have a linear output classifier. The COLE feature-weighting method is based on the idea that the *contributions* of features to a model's predictions are the best source of information for finding explanatory cases. In essence, the method abstracts the CNN function into a more comprehensible k-NN one, which favors similar cases for explanations based on how the learned features *contributed* to the prediction [22].

To understand COLE formally, consider a linear regression model f with n input features and a weight vector $\vec{w} \in \mathbb{R}^n$. The contribution vector $\vec{c}$ of an instance $\vec{x}$ is:

$$\vec{c} = \langle x_1.w_1, x_2.w_2 ... x_n.w_n \rangle \tag{1}$$

where $x_i.w_i$ calculates c_i (i.e., the contribution of feature i to the final prediction). We want to find explanatory cases with similarity metrics using $\vec{c}$, rather than $\vec{x}$ or $\vec{w}$, because $\vec{c}$ closer represents the actual predictive logic of f, as the output logit is ultimately given by $Linear(\sum_{i=1}^{n} c_i + Bias)$.

When using saliency maps to derive $\vec{c}$, this method is applicable to many modern CNN architectures such as VGG-16 [80, 22]. However, the use of multiple fully-connected-layers in the classifier output of networks such as VGG requires more parameters in the network to be trained, which in turn can lead to over-fitting problems. In the light of this issue, there has been a move in the computer vision community towards simpler classification outputs for CNNs. A good example of this is ResNet [81], a popular architecture that, typically, uses global average pooling (GAP) to condense the convolutional output matrix into a single dimensional feature vector. This layer is then followed by a single dense output-layer into a SoftMax function for its classification output. This step turns the CNN classifier half into a linear model and COLE can be calculated similarly using Eq. 1. However, it should be noted that a slight modification is required when dealing with multi-class classification problems (see Section 3.2 and Algorithm 1).

3.2. COLE Hadamard Product (COLE-HP)

Formally, a CNN is typically divided into two main parts, (i) a feature extractor network $g(I)$ which extracts a set of latent features $\vec{x}$ from the input image tensor I, and (ii) a classification network $f(\vec{x})$ which

gives the probabilities of all classes in the output Y, given an extracted input feature vector $\vec{x}$. Hence, as Goyal *et al.* [64] note, all networks can be defined as:

$$P(Y|I) = f(g(I)) \tag{2}$$

where $P(Y|I)$ is the probability of all individual classes, given the input tensor image I. Here f consists of two layers, and its first layer will be referred to as X (the final latent representation of I before classification). This layer X is the result of applying GAP on the final convolutional output matrix C, which works by taking the average of all the activations in each individual feature map in C as each extracted feature in $\vec{x}$.

Due to the simple linear architecture in f, which only consists of the original latent representation in layer X, and an output SoftMax layer [i.e., there is no hidden layer(s)], it is possible to implement COLE and calculate feature contributions (i.e., $\vec{c}$) of an instance $\vec{x}$ by taking the Hadamard product of:

$$\vec{c} = \vec{x} \odot \vec{w_c} \tag{3}$$

where $\odot$ is a Hadamard product, and $\vec{w_c}$ the weight vector connecting the feature layer X to the predicted class c in the output SoftMax layer. This product (Eq. 3) is taken for all training data and used to fit a k-NN classifier to implement COLE-Hadamard-product (henceforth COLE-HP), the same product is then taken for any new query, allowing nearest neighbor explanations to be found as shown in Fig. 2. In all the present work, we used Euclidean Distance as the distance metric in our k-NN classifiers.

3.3. Running COLE-HP on MNIST

Algorithm 1 was run on a CNN trained on the MNIST dataset to generate *post-hoc* explanations for the materials deployed in Expt. 1, whilst our previous approach using saliency maps was used for Expts. 2-3[1]. Many CNNs were trained to obtain a sufficient number of materials, with the classifier generally obtaining an accuracy of $\sim 99.5\%$ on the test dataset. Fig. 2 presents some of the sample explanations to illustrate the system outputs for both right and wrong classifications. Examples of classifications that the CNN got right are shown for queries using the numbers "7" and "6", along with the three explanatory cases retrieved from the weighted k-NN system acting over the training set (see Fig. 2a). For these right classifications, the explanatory examples show the training cases with the most similar classification logic to the query (i.e., the most similar feature contributions). Classifications that the CNN got wrong are shown for queries using the numbers "8" and "9" along with the three explanatory nearest-neighbors found in the training set (see Fig. 2b). Here, the CNN classifies the "8" as a "3" using examples that are quite similar the "8" but which were labelled in the dataset as "3s"; so while the prediction is wrong it appears "reasonable", and perhaps "more correct", given the labelled-data given to the model. A similar argument can be made for the wrong classification of the "9" as a "4", given the explanatory nearest-neighbors found. It is these sorts of predictions that we examine in the three user studies reported here, were people are asked to make correctness and reasonableness judgements about the CNN's right and wrong classifications. However, before reporting on these studies we consider the user models people may rely on when doing these tasks.

4. User Tests of *Post-Hoc* Explanations: An Experimental Paradigm

In the previous section, we saw the *post-hoc*, example-based explanations that the COLE-HP algorithm finds for CNN's classifications in MNIST (see Fig. 2). In this section, we present the experimental paradigm

[1]On computational costs, using saliency map techniques as in [22] took ~ 40 seconds derive weighting for the Connect-4 dataset ($67,557$ training and testing instances with 126 features), whilst it took ~ 6613 seconds for CIFAR-10 ($60,000$ training and testing instances with 1000 extracted features). The experiments used a MacBook Pro; processor: 2.9 GHz Intel Core i5; memory: 16 GB 2133 MHz LPDDR3. Subsequent queries took < 30 seconds to derive a single query's weighting and < 0.05 seconds to find explanatory cases for both tabular and image data. Regarding the use of COLE-HP for MNIST in Expt. 1, a query took < 1 second to find its weighting and an explanation due to the relative simplicity of the calculation compared to the use of saliency maps.

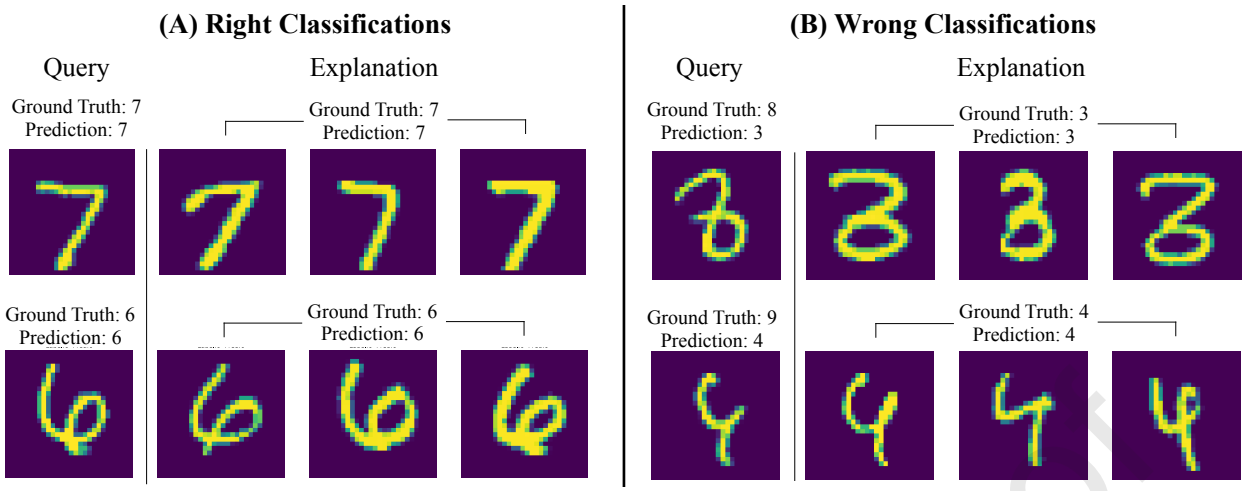


Figure 2: Example Explanations of Twin-Systems Using COLE-HP Feature-Weighting for Right and Wrong Classifications on MNIST: (A) Two right classifications by the CNN are shown with three, nearest explanatory cases from the training set (i.e., their feature contributions were the most similar). (B) Two wrong classifications by the same CNN, again with three, nearest explanatory cases, where an alternate-labelling of a very similar image was used.

used here to test for the impact of these *post-hoc* explanations and error-rates on people's mental models. So, we begin by laying out the materials used in our studies and the main independent variables explored. Then, we consider the different measures used to assess changes in people's mental models of the task. In the next main section (Section 5), we advance a theoretical analysis of this experimental paradigm with reference to three conceptual frameworks for XAI: a taxonomy of XAI user-tests (see [27], and Section 5.1), the DARPA model of explanation (see [26]; see Section 5.2), and a new proposal called the Tricorn User Model of XAI (see Section 5.3).

4.1. The Current Paradigm: Task Context and Variables Tested

The current experimental paradigm is a debugging-a-classifier task, in which participants are asked to evaluate the predictions made by a "computer program" (the CNN-CBR MNIST twin-system; see Section 3) to assess the correctness and reasonableness of the classifications made. Specifically, the people were told that *"the program labelled the number this way because of what it learned from the human-labelled numbers it was shown"*. In all the experiments reported here, participants worked their way through 25-30 different classifications made by the system. In the present experiments, three key manipulations to these items were explored: the presence/absence of the explanations (i.e., Explanation), the presentation of right/wrong classifications (i.e., Classification-Type), and the relative proportions of right/wrong classifications (i.e., Error-Rate).

Explanation. In all the current experiments, we manipulated the presence or absence of explanations for the classifications presented. Fig. 3 shows samples of the matched materials seen by participants for this manipulation. Fig. 3A shows an explanation-present item, where a correct classification of a "7" by the CNN is shown with an explanation saying that it made this prediction "because of what it learned from these labelled numbers it was shown", followed by the three nearest-neighbors of the test-instance found by the COLE-HP algorithm. Fig. 3B shows an explanation-absent item where, again, a correct classification of a "7" by the CNN is shown with the statement "the program labelled the numbers this way because of what it learned from the human-labelled numbers it was shown". This could be called a "non-explanation explanation"; it just says "X occurred because of some unspecified Y". Note, also, that in these explanation-absent items, participants complete a separate sub-task, where they are asked to note the labels of three-unrelated images that have no explanatory relationship to the classification. This sub-task ensures that the materials are matched with explanation-present-items both visually and in the time

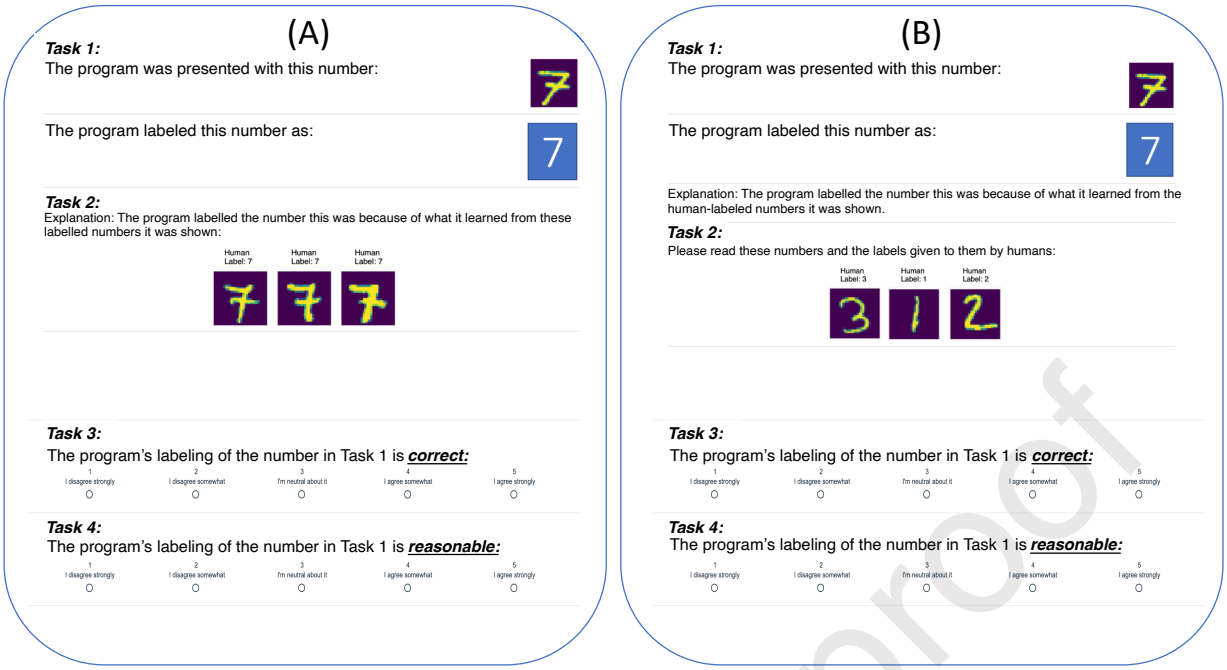


Figure 3: Examples of Correct Classifications presented in the Experiments: Showing a correct prediction made by the CNN in the (A) Explanation-Present condition with a 3-example explanation and (B) Explanation-Absent condition with a "non-explanation explanation". Note, tasks 3 and 4 present the rating scales for *Correctness* and *Reasonableness* judgements, respectively.

participants spend considering the presented information. Then, in both manipulations, participants are asked to perform two rating tasks to evaluate the classification shown at the top of the page; they are tasked to rate the "correctness" and "reasonableness" of the classification shown on a 5-point, Likert-type scale (from "I disagree strongly" to "I agree strongly").

Classification-Type. In all the current experiments, the classifications presented were manipulated on the basis of being right or wrong across the 25-30 items shown (in Expt. 1 two error-types were considered; see Fig. 4). *Right Classifications* are predictions made by the CNN that agree with the ground truth (e.g., see Fig. 3). Fig. 2A and 3 show examples of right classifications; for example, the CNN labelled the image as a "7" and the ground truth tells us that "7" was the correct label for that item (see Fig.3A). *Wrong Classifications* are predictions made by the CNN that disagree with the ground truth (e.g., see Fig. 4). Fig. 2B and 4 show instances of wrong classifications; where, a "6" was labelled as a "0" and a "3" labelled as an "8" (where an explanation is presented). Within this manipulation, two distinct classes of error were deployed, depending on the explanatory nearest-neighbors presented: alternate-labelling errors and majority-voting errors. *Alternate-labeling errors* are classification errors where the explanatory cases found have a visually-similar image, that is labelled as a different class to the ground truth (e.g., see Fig. 3b); arguably, these are "reasonable" errors as the CNN is using the experience it was given, appropriately. *Majority-voting errors* are classification errors where the explanatory cases found show two alternate-labeled images and one image labelled with the same class as the ground truth (e.g., see Fig. 4b); this explanation suggests the prediction error arose from a majority vote for the wrong class, even though there was some support for the ground-truth class. Again, people may find these to be "reasonable" errors as, arguably, the CNN appears to be using the experience it was given in a "rational" way. We expect these errors to interact with the explanation-variable; specifically, when the explanation is absent people just see a wrong classification by the model, whereas when the explanation is present people are given evidence for considering these errors as being different based on the explanatory-examples provided.

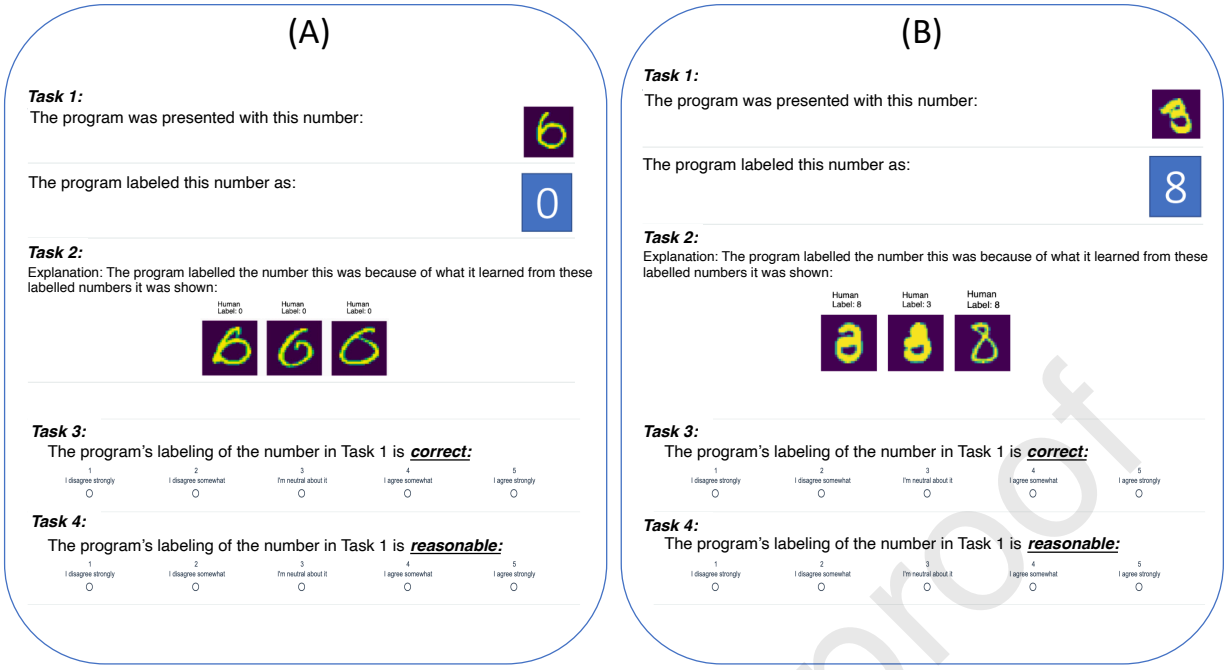


Figure 4: Examples of Wrong Classifications presented in Expt. 1 (Explanation-Present condition): Showing a prediction error made by the CNN that is (A) an *alternate-labeling-error* with its 3-case explanation, and (B) a *majority-voting-error* with its 3-case explanation.

Error-Rates. In all the current experiments, people were presented with different proportions of right/wrong items (i.e., percent error-rates). Over the three experiments, we varied errors-levels between 3-60% to determine how this variable impacted their perception of the system. As we saw in the related work, people appear to have a low tolerance for prediction error in automated systems, so on the basis of previous research one might expect rising error-rates to impact people's overall trust and satisfaction with the system.

As we shall see later, across three experiments these variables were systematically varied. Expt. 1 tested for effects of providing explanations using three classification-types (right classifications, alternate-labelling, and majority-voting errors) in the context of a 20% error-rate. Expt. 2 tested for effects of providing explanations using two classification-types (right and wrong predictions) across three error-rates (3%, 30%, and 60%). Expt. 3 tested for effects of providing explanations across two classification-types (right and wrong predictions) for four different error-rates (4%, 12%, 20%, and 28%). In the next section, we outline the different measures used in these user studies.

4.2. Current Paradigm: Item-Level and System-Level Measures

Several different measures were used in the above experimental paradigm involving this classifier-debugging task, measures that are directed at different levels of analysis: item-level measures (i.e., people's judgements of correctness and reasonableness for each presented classification) and system-level measures (people's evaluations of trust and satisfaction in the overall system[2]). Note, in the use of these measures we did not attempt to define the meanings of "correctness" and "reasonableness" in the instructions given; the aim being to let participants interpret them using their "normal" meanings, rather than to provide experimenter-defined ones. As we shall see, when people have clear agreed understandings of a term, in context, there tends to be a lot less variance in responding (as is the case for "correctness"), though when there is less agreement,

[2]Note, in Expt. 3 we also asked people to judge the "overall" correctness and reasonableness of the system

Q1. I am confident in the system. I feel that it works well.
Q2. The outputs of the system are very predictable.
Q3. The system is very reliable. I can count on it to be correct all the time.
Q4. I feel safe that when I rely on the system, I will get the right answers.
Q5. The system is efficient in that it works quickly.
Q6. I am wary of the system.
Q7. The system can perform the task better than a novice human user.
Q8. I like using the system for decision making.

Table 1: Table 1. Survey Questions of Hoffman *et al.* [26] for Trust in an AI System.

or some ambiguity, a lot more variance in responding occurs (as, we shall see, seems to occur with the term "reasonableness")

Item-level measures assess individual predictions made by a system (e.g., when a loan-application is refused and the applicant is told "If you had a higher salary then you would receive the loan"). *System-level measures* assess the overall adequacy of the system; that is, how trust/satisfaction with system as a whole is evaluated by users after interacting with it. For instance, many of the DARPA evaluations have a system-level focus, where they aim to assess whether the system as a whole, is perceived by users to be more trustworthy/satisfactory based on the explanations provided [65, 25, 64]; indeed, Hoffman *et al.* [26] have defined standardized survey-questions to assess explanation satisfaction and system trust (that are used here; see e.g., Table 1).

Traditionally, in the ergonomics of automated systems, the human in-the-loop using a system would be evaluated carrying out some key task (e.g., sorting letters with an automated post-code reader) by using various performance measures (e.g., accuracy/speed completing an item or some defined task with many items); as we shall see, the DARPA framework for explanation proposes an XAI version of this approach (see Section 5 and Fig. 5). Typically, in these evaluations, item-level effects (e.g., a person's decision time on a single item) are generally assumed to aggregate to produce overall system-level effects (i.e., faster performance on the overall task involving many items). However, as we shall see later, in the use of explanations, item-level effects may not necessarily aggregate to be reflected in system-level effects; mainly because the measures used may tap different aspects of people's mental models. In the next section, we turn to a consideration of such issues with a theoretical analysis of the current studies, by considering (i) a taxonomy of XAI tasks, (ii) people's mental models in XAI tasks, and (iii) people's mental models in the present experimental tasks.

5. Mental Models in XAI: Taxonomies, Frameworks, and Theory

Earlier, we sketched prior work on user studies of XAI, on trust in automated systems, and on the classifier-debugging task used here (see Section. 2). These literatures show that, from a user testing and psychological perspective, XAI is still a very inchoate area (as noted by [65, 26, 30, 31, 10]). So, there are still significant theoretical gaps in how to characterise XAI tasks and people's mental models in these tasks. In this section, we consider several proposals on such issues along with a new framework for mental models in XAI tasks. We also apply these proposals to the current experimental paradigm, to elucidate the psychological basis for the current work. So, in this section, we review (i) a proposed taxonomy for XAI tasks [27], (ii) the DARPA conceptual model for evaluating explanation [65, 26], and (iii) the present proposals on the *Tricorn User Model* of XAI.

5.1. A Taxonomy of XAI Tasks

Doshi-Velez and Kim [27] have advanced an influential taxonomy for XAI user studies. They propose a three-leveled taxonomy for XAI evaluations: (i) *application grounded evaluation*, where human end-users (typically with some domain expertize) are tested using the complete AI model in the task for which it was built; (ii) *human-grounded evaluation*, where human end-users (who may not be domain experts) test selective aspects of the "real" application task with the AI model; (iii) *functionally-grounded evaluation*,

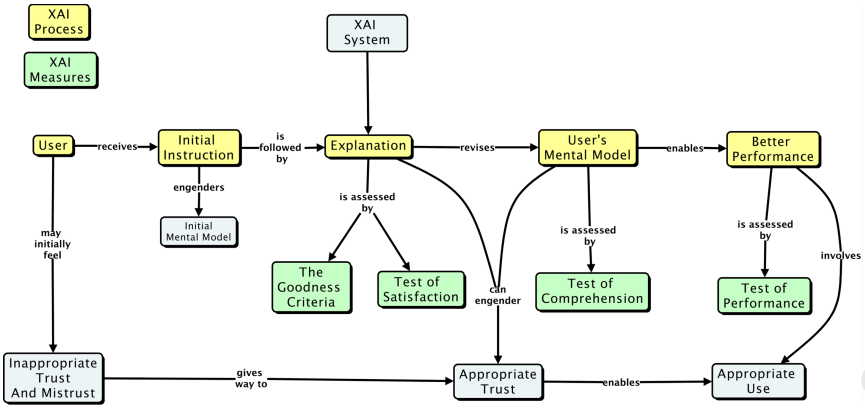


Figure 5: A conceptual model showing how the impact of explanations on user mental models and user performance can be measured by different types of tests in an XAI study (from Hoffman *et al.* [26]).

where the evaluation is based on some computational-proxy for an actual human evaluation (i.e., no human testers). As we have seen, most of the current XAI literature is carried out using functionally-grounded evaluations rather than tests involving users. In this taxonomy of tasks, end-user engagement reduces across the three levels; beginning with end-users evaluating "real" explanations generated by the complete system, moving to end-users responding to general aspects of explanation (or localized system performance) and, finally, to no user involvement at all. Implicit, perhaps, in this analysis is the view that the results of these different evaluations needs to be weighted by the level of user-engagement with the actual system.

The present experimental paradigm mainly belongs to the first level of this taxonomy (i.e., we have human experts evaluating the outputs of the AI model), though some aspects of it shade into the second level (i.e., people are not debugging a running system, as in [24]). Rather, people in our studies are making judgments about the system, evaluating actual outputs from the AI system on several dimensions (i.e., correctness, reasonableness, trust, and satisfaction), as indirect measures of an explanation's impact while performing a debugging-task on system outputs.

5.2. DARPA's Explanation Model

As part of the DARPA XAI Program, Hoffman *et al.* [26] proposed a conceptual model of how explanations might impact a user's mental model as they interact with an AI system (see Fig. 5). They propose that, faced with an AI system, users have some initial mental model, partly informed by the instructions they receive about the system. Then, as the user interacts with explanations of the system's operation, their user model is updated and develops, leading to better performance (presumably, if the explanation is effective). They point out that, at the outset, the user may have inappropriate trust (or indeed distrust) in the system, but that if the explanation strategy works, then appropriate trust will emerge from these interactions. Finally, they propose that there are a number of junctures where different measures can be applied to test the goodness of the explanation/system, the status of the mental model (in comprehension tests), and the impact of the developing mental model on performance (in performance tests). For example, with respect to trust, they proposed an 8-question trust survey consisting of questions about the explanation facilities of the system (see [26] and Table 1).

This framework is useful in laying out some of the key steps in mental model development, especially in contexts where we expect the provision of explanations to improve a users' performance with the system. However, at this level of analysis, it presents the user's mental model of the system as a single undifferentiated entity, when there are grounds for thinking that mental models needs to be further differentiated. As we shall see, this mental model can be further sub-divided into the user's knowledge of (i) the domain, (ii) the AI techniques being used by the system, and (iii) the explanation strategy. So, in the next subsection, we

present an extension to this explanation framework that involves a more multi-faceted mental model for XAI tasks.

Having said this, the current experimental paradigm can be parsed using the DARPA framework, as it sits in the central portions of the conceptual model (see Fig. 5). The instructions given to participants convey an initial mental model of the task and, as they make judgements about the system's classifications, they encounter the explanations of the XAI system. The "goodness criteria" could be aligned with the correctness and reasonableness judgements. We assume that participant's experience with the system's classifications (and their explanation) impacts the user's mental model in some way, though we do not have explicit tests for performance improvements on the debugging task (though one could easily imagine using user-feedback on error-items to fine-tune the AI model). However, we do consider measures of "appropriate trust" (in our use of the DARPA Trust and Satisfaction surveys). In the next subsection, we push this analysis further by considering how the user's mental model can be further partitioned to better understand how it might be deployed in XAI tasks.

5.3. Tricorn User Model of XAI

We have seen that the DARPA framework for explanation casts XAI tasks as involving the development of an undifferentiated mental model that emerges from a user's interaction with a system in some task/goal context. At one level of analysis this makes sense; however, a more detailed level of analysis is also possible if we divide the mental model into sub-models that capture a user's knowledge of (i) the domain, (ii) the AI model's operation, and (iii) the explanation strategy. Indeed, this partitioning of a user's mental model also implies that each of these sub-models could develop in different ways, based on a user's experience with the overall system (e.g., a given explanation strategy could impact one sub-model more than another). Hence, we advance this Tricorn User Model, that divides the user's mental model into three distinct sub-models (see Fig. 6):

- **User Model of the Domain ($\mathbf{UM^D}$).** The user has some understanding of the domain, they may know very little (low expertize) or be experts (high expertize), and this understanding may develop as they interact with AI system and explanation strategy; for instance, most people in the general population would be non-experts in a legal-sentencing domain (e.g., see [82]), whereas most of us are deep experts in written letters/numbers after years of schooling and deciphering bad handwriting (as in the current MNIST dataset).

- **User Model of the AI System ($\mathbf{UM^{AI}}$).** The user also has some understanding of the AI model that is separate from the domain model, some conception of how the AI model works; for example, people will have views about an AI classifier that it, in some sense, "learns" from experience or that it "just follows rules".

- **User Model of the Explanation ($\mathbf{UM^{XP}}$).** The user will also have an understanding of how an explanation strategy explains the system; for example, if *post-hoc* explanatory cases are being used, people understand that the explanation is asserting an evidential dependency between the nearest neighboring examples and the query. Similarly, if the explanation is a saliency map then people will be using their knowledge of such visualizations.

The final part of this framework is the task-goal context in which these sub-models are being applied. End-users may be interacting with the system to debug it, to gain some causal insight into a domain, to make a decision, and so on (see Lipton's desiderata [83]). So, these different mental models could develop differently depending on the task-goal context even when the items being considered are the same (see Fig. 6). This tripartite view of the user model has a number of significant implications for how we approach evaluations of XAI systems.

First, in every XAI user-study, though researchers may state that they are just evaluating the AI model, they are actually assessing all three of these sub-models. So, unless they specifically separate these sub-models in their independent variables (by controlling them or counterbalancing), their behavioural measures will reflect some complex interaction of changes to all three sub-models. So, for example, if some improvement

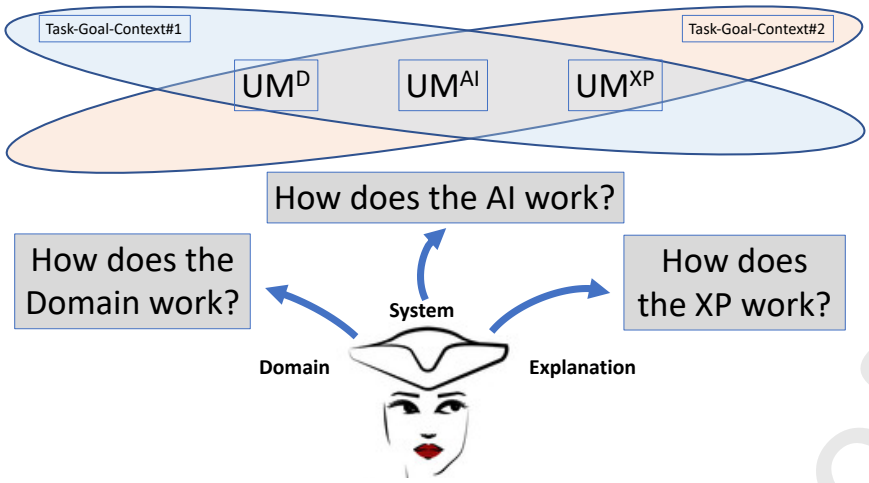


Figure 6: Tricorn User Model of XAI: The user model has three distinct sub-models for the Domain (UM^D), AI System (UM^{AI}), and the Explanation strategy (UM^{XP}), all of which may be different in different task-goal contexts.

in performance is observed on an AI system, that improvement may not reflect an improvement in the user's model of the AI (or indeed the domain); it may rather reflect their better understanding of the explanation strategy being used (e.g., a user coming to understand what the saliency map actually shows). So, without some consideration of these sub-models, we cannot be really sure what it is that we are evaluating in XAI user studies.

Second, people may have different levels of expertize in each of these sub-models in a given user study. People may know nothing about a domain (e.g., what criteria judges use in sentencing) or they may be experts (e.g., in deciphering handwritten numbers). Similarly, people may have less expertize about one explanatory method (e.g., a saliency map) than another method (e.g., an example-based explanation). Moreover, some people may have no idea about how the AI model works at all (the "it-just-follows-rules" person), whereas others may have precise ideas about how it works (e.g., the ML PhDs in Ribeiro *et al.*'s [5] studies). So, the relative level of expertize with respect to each sub-model needs to be factored into any user study.

Third, it needs to be realized that any or all of these sub-models may change and evolve in the course of a user study; users may come to a better understanding of the domain, they could gain a new insight in how the AI works or, indeed, they could learn more about an unfamiliar explanatory method (e.g., a saliency map). So, the dependent measure in the user study needs not to conflate the potential changes to all three sub-models together (as for example, many simple performance measures might do). Researchers need to select dependent measures that specifically address the sub-model they wish to assess; for instance, if the claim is that using an AI model improves decision making by informing users about the causal-dependencies in the domain, then the dependent measure in the user study needs to specifically test that causal knowledge, not some filtered version of that knowledge through the "dark glass" of some overall performance with the system.

Fourth, we need to remember, that the dynamics of change and interaction between these three sub-models will vary depending on the task-goal context. Trivially, if I am assessing the fairness of the AI system's decision-making as opposed to its causal-accuracy in an ill-defined domain, then my interaction with the system will be very different; users will learn about different aspects of the system and form very different mental models from their experience within different task-goal contexts. However, there also will be many more subtle interactions where (apparently minor) changes to the task-goal context will significantly impact what people learn or take from their interaction with the system; for instance, small instructional changes, or modifications in the way a problem is represented can lead to large, unforeseen changes in what people may take from the interaction. This is just a truism that emerges from decades of research in cognitive psychology and human-computer interaction.

Finally, this tricorn analysis points up the need for designers of XAI user-studies to think about which sub-model they are targeting with their explanation method. They need to decide whether the goal of the explanation is to improve the user's understanding of the domain or to improve their trust in the AI, or indeed, is it just about improving their ability to explain what the AI model is doing (or it is some mixture of all of these requirements).

Applying the Tricorn User Model to the Current Paradigm. As such, this framework can be applied to the current experimental paradigm. Indeed, its application suggests that most of the mental-model development in our participants should be focused on the AI model and how it is operating. If we map the current paradigm into this framework it suggests the following:

- **User Model of the Domain ($\mathbf{UM^D}$).** We assume that users are deep experts in the domain of recognising written numbers (as used in the MNIST dataset) after years of schooling and deciphering handwriting. Hence, the impact on people's mental models of the domain within the task should be minimal; that is, we would not expect them to "learn" anything more about the identification of written numbers from doing our experiments.
- **User Model of the Explanation ($\mathbf{UM^{XP}}$).** We also assume that any observed effects are unlikely to be due to changes in their understanding of the explanation strategy; people are somewhat used to example-based explanations, and while there may be some early adjustment to the way explanatory items are presented in the materials, for the most part this portion of the model should not change radically.
- **User Model of the AI System ($\mathbf{UM^{AI}}$).** This means that most of the impact on the user's mental model should fall on their understanding of the AI model (arguably, where we want it to be). So, in making judgements of correctness and reasonableness, people's mental models of the AI should be changed (in some way) that presumably may be reflected in item-level measures (of correctness and reasonableness), and in system-level measures (of trust and satisfaction).

Finally, the task context here is the debugging-task, in the sense that participants are being asked to make judgements about the classifications of the CNN; the focus is on evaluating the correctness and reasonableness of the system's operation. With this theoretical analyses in mind, we now turn to reporting the results of the three experiments performed using the current experimental paradigm (see Sections 6, 7, and 8).

6. Experiment 1: *Post-Hoc* Explanations of a CNN Classifier

Using the experimental paradigm described previously, the first experiment involved a straight test of the provision of *post-hoc*, example-based explanations for the CNN's classifications, based on the optimal feature-weighting method (COLE-HP) found in our computational experiments [22]. So, the study used 30 classifications from the CNN operating on the MNIST dataset, where 80% of predictions made where right (24 items) and 20% were wrong [of these errors, 10% were alternate-labeling errors (3 items) and 10% were majority-voting errors (3 items); see also Fig 7]. Note, that even though all of these errors were "real" misclassifications made by the CNN (over many runs) the model actually does much better; typically, the CNN was correct in over 99.5% of the its predictions. Here, we used a 20% error-rate to increase the CNN's error-rate for testing purposes, while still presenting the system as being relatively successful.

Participants were randomly assigned to two groups where the explanation was either present or absent (as described in Section 4). So, the study had a mixed design involving a 2 Explanation (present v. absent) x 3 Classification-Type (right-classification v. alternate-labeling-error v. majority-voting-error) structure, with Explanation being a between-participants variable and Classification-Type being a repeated measures, within-participants variable. We predicted that the provision of the *post-hoc*, example-based explanations would impact people's mental model of the classifier, in item-level measures, reflecting their perceived correctness and reasonableness of wrong classifications. Note, we did not expect explanations to impact the people's perception of right classifications, because we assume that explanations only play a role when things go wrong. Recall, some of the DARPA evaluations showed that explanations only impacted errors [25]. Thus,

we expected an interaction between the Explanation and Prediction-Type variables. Overall, we also expected that the provision of explanations should impact the system-level measures of trust and satisfaction; but, we accepted that such effects could be tempered by the 20% error-rate used in this study.

6.1. Experiment 1: Method

6.1.1. Participants and Design

One hundred and two people were randomly assigned to the two groups in the study, the Explanation-Present (N=51) and Explanation-Absent (N=51) conditions. Using GPOWER [84], the power analysis for two separate one-way t-tests, assuming a moderate effect size for each (d = .50), shows that an N=102 for this design ensures an alpha of .05 and power of .80. It should be noted that in all the current experiments the sample sizes used were those indicated by an appropriate power analysis designed to balance the probabilities of Type I and Type II errors. Specifically, in all the studies, GPOWER [84] was used, with the assumption of a moderate effect size. As such, there are no grounds for supposing that over- or under-sampling occurred in the studies reported. This study passed review by the university's ethics board (ref. LS-E-19-148-Kenny-Keane).

6.1.2. Materials

Each participant received 30 items to judge, from the classifications made by the CNN model for the MNIST dataset, selected from the outputs of the model (described earlier). As the model produces very few incorrect predictions, several runs were made to get the 6 error items (3 alternate-labeling errors and 3 majority-voting errors), giving an 80:20 ratio between right and wrong items. The order of these items were randomized for each participant.

6.1.3. Procedure

The study was run on the Prolific (www.prolific.com) crowd-sourcing platform with filters to select native English speakers in the United States, United Kingdom, and Ireland. Participants were paid a nominal fee for their participation. In the introductory pages to the study, instructions asked people to look at each item and complete four simple tasks, as well as to rate their satisfaction and trust in the program at the end of the study. For each item, they were asked to rate the prediction on its *correctness* "The program's labelling of the number in Task 1 is correct" and *reasonableness* "The program's labelling of the number in Task 1 is reasonable" on a 5-point Likert-scale from *I disagree strongly* (1) to *I agree strongly* (5). After rating the items, participants were given the sixteen survey questions on trust and satisfactions in blocks of eight (not reported here in detail). After they had completed the study they were given a debriefing page that said, "This study is being conducted to determine the effect of the use of explanations for the outputs of computer programs. Your responses will be used to compare the perceived correctness and reliability of computer programs such as this used with and without explanations. Thank you for taking our survey. Your response is very important to us." Finally, at the very end of the session, participants were asked to give some verbal feedback on the experience of the study; in their own words, they were asked to respond to the following question: "You may have noticed that sometimes the program did not seem to label the numbers correctly. Can you state, in your own words, in as much detail as possible, why you think the program failed on these items?".

6.2. Expts. 1-3: Qualitative Analysis of Verbal Reports

Before considering the quantitative results it is perhaps informative to consider a qualitative analysis, across all three studies, based on people's verbal reports about the program in response to a final question in which they were asked to say "why you think the program failed". Most participants gave 1-2 sentence replies to this question reflecting their opinions of the program: for example, participants' replies included (i) "it failed because the human examples were different from each other so it had no solid example to use", (ii) "I think the numbers in some of the examples were written in such an unclear way, that the computer couldn't clearly make out which number it was", (iii) "The program probably has an algorithm that takes too much assumption (sic) on what the user is trying to do with mis-marks", (iv) "It failed as

it could not recognize". Content-wise, these responses divide into those that (i) blamed-the-data (i.e., the numbers were unclear or labelled poorly by human annotators), (ii) blamed-the-program (i.e., the system failed to recognize or properly process the images), (iii) blamed-both (i.e., the data is poor and the model failed), and (iv) blamed-neither (i.e., where some other fault is mentioned or the blame is unclear). To get a qualitative sense of people's conscious perception of the model's performance we content-analyzed their verbal responses; two raters (MQ and CF) classified all the responses across our three studies (N=452) into one of these four categories (differences were resolved by discussion, and agreement on the categorizations was good; Cohen's K=0.6, $p < .001$). While the majority of people blamed-the-data for the errors occurring rather than blaming-the-program, there is a shift in the Explanation-Present conditions (N=226) towards more people blaming-the-model/blaming-both and away from blaming-the-data than in the Explanation-Absent conditions (N=226), that is reliably different, $\chi^2(3) = 11.094$, $p = .01$. In the Explanation-Absent conditions the verbal responses were blamed-the-data (68%), blamed-the-model (9%), blamed-both (18%), blamed-neither (5%). In the Explanation-Present conditions the verbal responses were blamed-the-data (53%), blamed-the-model (15%), blamed-both (25%), and blamed-neither (7%). Clearly, the explanations focus people's attention on the operation of the model, leading to an increased focus on seeing it as the source of error and a decreased focus on the data as the source of difficulties. So, this qualitative result shows that, overall, people's assessment of the system was clearly impacted by the explanations given (though, as we shall see, not necessarily in a positive way).

6.3. Experiment 1 Results: Quantitative

As predicted, *post-hoc* explanations produced by the system had a significant effect on people's perception of the *correctness* of the model's predictions (unlike *reasonableness*). Stated simply, people view errors made by the system as being more correct (or less incorrect) when explanations are present than when such explanations are absent (see Table 2 and Fig. 8a/b). Interactions were found between the Explanation and Classification-Type variables for both the correctness and reasonableness ratings. However, the item-level and system-level measures differed; people's trust/satisfaction in the overall system was not impacted by the provision of explanations. So, while explanations impacted people's evaluations of items, these effects do not aggregate into overall improvements in trust/satisfaction. In short, one could say, the explanations did not "explain away" the 20% error-rates manifested by the system (note, this error-rate is > 19% higher than the "real" error-rate of the CNN, which was < 1%).

6.3.1. Correctness Ratings

Participants' correctness ratings for the 30 predictions were collated and analyzed (n.b., on the 5-point correctness scale, where 1 is low on correctness and 5 is high). A 2 (Explanation: present v. absent) x 3 (Classification-Type: right v. alternate-labeling-error v. majority-voting-error) ANOVA with repeated measures on the second variable was applied to the mean correctness ratings of the presented items[3]. This analysis revealed main effects of Explanation ($p < .001$; see Table 2), and Classification-Type $F(2, 200) = 1235.9$, $p < .001$, partial $\eta^2 = 0.93$; as well as a reliable interaction between Explanation and Classification-Type, $F(2, 200) = 5.4$, $p = .005$, partial $\eta^2 = 0.05$.

Fig. 7 presents one view of this interaction, showing the mean correctness ratings for the 30 matched item-classifications in the experiment, by plotting Explanation-Present *by* the Explanation-Absent conditions (and dividing them into right and wrong classifications). So, the points on/close-to the diagonal are classifications where the correctness-rating is the essentially the same in both conditions (i.e., Explanation-Present = Explanation-Absent) and points below the diagonal are classifications where the correctness-rating is higher when the explanation is present (i.e., Explanation-Present > Explanation-Absent). The figure also shows us that the effect of the explanation really only occurs for the wrong-classifications (see red/yellow circles

[3]All analyses in Expt. 1 used mixed effects ANOVAs with Type III Sums of Squares, as this seemed most appropriate. In Expts. 2 and 3 we led with MANOVA analyses of each measure as they seemed more appropriate; though for comparative purposes, in Table 2, we report univariate ANOVA analyses for all experiments.

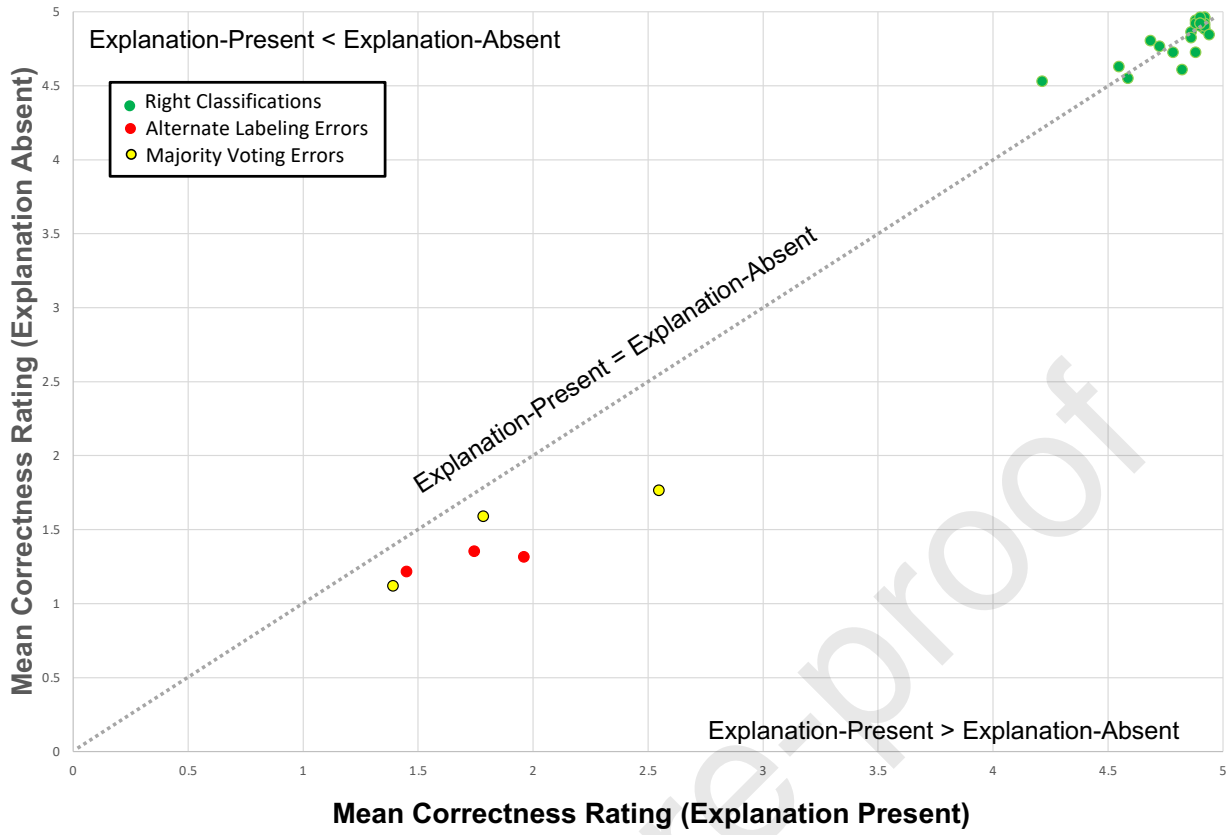


Figure 7: Experiment 1: A plot of the mean correctness ratings for the 30 matched classification-items presented in the Explanation-Present and Explanation-Absent conditions, showing right classifications (green circles) and wrong/error classifications (red and yellow circles).

in Fig. 7) and how the right classifications differ from the wrong classifications in showing no effect of explanation and receiving much higher ratings (see green circles in Fig. 7).

Fig. 8a shows another view of the Explanation by Classification-Type interaction with the means for each classification-type in the experiment; we can see that explanations have little impact on the right-classifications, but the perceived correctness of wrong-classifications increases when the explanation is provided. Specifically, for right classifications, the correctness ratings in the Explanation-Present (M=4.821, SD=0.32) and in Explanation-Absent (M=4.828, SD=0.34) conditions were almost identical and do not differ reliably (see Fig. 8a). In contrast, for error classifications the explanation conditions differ; for alternate-labeling-errors, the correctness ratings in the Explanation-Present condition (M=1.72, SD=0.70) are higher than in the Explanation-Absent condition (M=1.29, SD=0.58) and are reliably different (using Tukey HSD test, p <.001; see Fig. 8a). Similarly, for majority-voting-errors, the correctness ratings in the Explanation-Present condition (M=1.90, SD=0.75) are higher than those in the Explanation-Absent condition (M=1.49, SD=0.64) and are reliably different (Tukey HSD test, p <.003; see Fig. 8a). These results show that the provision of example-based explanations lead people to view the classification-errors of the program as being more correct (or *less incorrect*), presumably because they get some insight into why the model is making these errors. Notably, with respect to the right classifications, 99% of participants rated them higher than wrong-classifications (i.e., only one participant out of 102 gave a mean rating of the errors that was higher than that given for the right classifications). Hence, we can conclude that outliers did not substantially skew these correctness results.

To put it simply, explanations seem to mainly have an effect when things go wrong, when errors arise and

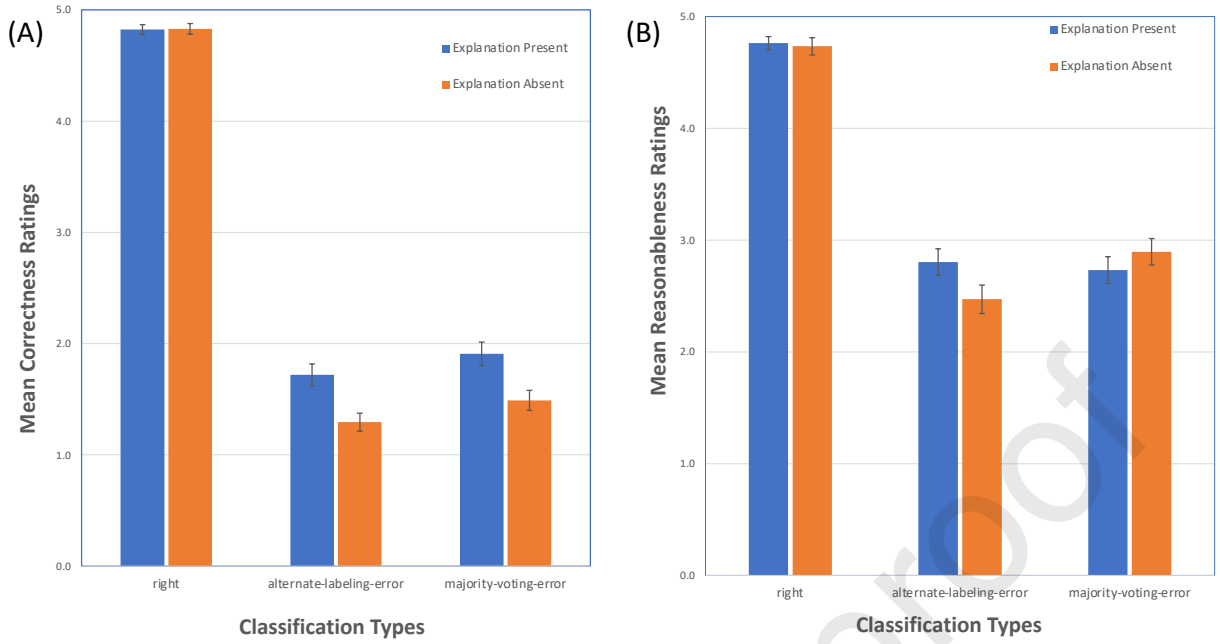


Figure 8: Experiment 1: Mean ratings for (A) Correctness and (B) Reasonableness for the three Classification-Types (right, alternate-labeling error, and majority-voting error) in the Explanation Present/Absent conditions (Standard Error bars are shown).

outputs diverge from what a user expects/desires (as was found in some of the DARPA evaluations [25]). This account seems reasonable for now, though it probably requires further exploration. Glickenhaus *et al.* [25] have queried whether this phenomenon may be due to a ceiling effect; that is, as people's ratings are already high for the right items, there is very little headroom for ratings to go higher when explanations are provided.

6.3.2. Reasonableness Ratings

Participants' reasonableness ratings for the 30 predictions were also collated and analyzed (n.b., on the 5-point reasonableness scale, where 1 is low on reasonableness and 5 is high). A 2 (Explanation: present v. absent) x 3 (Classification-Type: right v. alternate-labeling-error v. majority-voting-error) ANOVA with repeated measures on the second variable was applied to the mean reasonableness ratings of the presented items. This analysis revealed a main effect of Classification-Type, $F(2, 200) = 335.3$, $p < .001$, partial $\eta^2 = 0.77$; as well as a reliable interaction between Explanation and Classification-Type, $F(2, 200) = 3.84$ $p < .05$, partial $\eta^2 = .037$. However, the main effect for the Explanation variable was not statistically significant ($p > .05$; see Table 2). Based on the reliable interaction, we explored the planned pairwise comparisons between conditions.

Fig. 8b shows that this interaction mainly reflects the differences between the right and wrong classifications. Importantly, all the differences within the classification-types for the explanation manipulation are not reliable when using the Tukey HSD tests for right classifications (ns), alternate-labeling errors ($p = .058$), and majority-voting errors (ns). So, the pattern of results for the reasonableness measure does not parallel that found for the correctness measure.

This failure to find effects in the reasonableness rating may in part be due to people's high expertize in this domain; people are domain experts in reading cursive script and, as such, may be quite unforgiving when a program does not manifest the same level of expertize. We have not found other XAI user studies that

have rated reasonableness in this way. As we shall see in our other experiments, it does not always reveal clear effects, suggesting that it may be also interpreted differently by different participants in this task.

6.3.3. Trust and Satisfaction

Participants were also asked a number of system-level questions based on the DARPA surveys for Trust and Satisfaction (see Table 1; [26]). Statistical analyses of each of these surveys revealed no main effects of Explanation or reliable interactions between the Explanation and Trust/Satisfaction-Question variables (as one would expect if explanations increased trust/satisfaction in the system). A 2 (Explanation: present v. absent) x 8 (Trust-Question) ANOVA, with repeated measures on the second variable, showed a main effect of Trust-Question ($p < .0001$), but no effect of Explanation ($p = .38$; see Table 2). Similarly, the 2 (Explanation: present v. absent) x 8 (Satisfaction-Question) ANOVA, with repeated measures on the second variable, showed a main effect of Satisfaction-Question ($p < .0001$), but not for Explanation ($p = .24$; see Table 2). Here, the effects seen for the Trust- and Satisfaction-Question variables, merely tell us that some of these questions differ from others in their scores, which is not a focus for the present analysis. The effect-of-interest was the interaction between the Explanation and Trust/Satisfaction-Question variables, which was *not* found. So, on the face of it, these results suggest that though explanations produce item-level effects impacting people's mental models of the misclassifications they encounter, they do not "explain away" the overall performance of the system to improve people's assessment of it. That is, the provision of explanations does not change people's judgements of trust and satisfaction with a system that has a 20% error-rate. Recall, the literature on algorithmic aversion showed that people were quite intolerant, even of very low error-rates. This, of course, raises questions about what error-rates might be acceptable. Hence, in the next two studies we address this issue by systematically varying the error-rates people encounter, while also continuing to examine the effects of explanations.

7. Experiment 2: Explanations and High-End Error-Rates

Experiment 1 found that explanations only impact people's correctness judgements of the CNN's misclassifications. Specifically, it found that when misclassifications were explained using *post-hoc* examples, people rated them as significantly more correct (or *less incorrect*). However, it also found that providing item-level explanations did not impact people's overall trust and satisfaction in the system (at least, when faced with error-rates of 20%). So, while the explanations made the errors at the item-level more "correct", at the system-level they did not "explain away" the overall error performance of the system. Clearly, this result invites further exploration on what error-rate might indeed be acceptable; a question that is common to all AI systems where people encounter system-errors. So, in this experiment, we explored a very low error-baseline (3% errors, which is very close the the "real" error-rate of 1% for this CNN) contrasting it with very high levels of error (30% and 60%). This Error-Rate variable was crossed with the Explanation variable, using the same measures as in Expt. 1 (namely, judgements of correctness, reasonableness, and the trust/satisfaction surveys). So, the design was a 2 (Explanation: present v. absent) x 3 (Error-Rate : 3% v. 30%, v. 60%) x 2 (Classification-Type: Right v. Wrong) with Explanation and Error-Rate being between-participants variables and Classification-Type being a within-participant variable.

7.1. Experiments 2: Method

7.1.1. Participants

One hundred and sixty-five people were randomly assigned to the 6 groups in the study with roughly equal numbers in each condition (Ns = 27). All participants were aged over 18, native English speakers and lived in the USA, UK, or Ireland. Exclusion criteria were participation in previous studies by the lab and inattentive answering (3 people were excluded for failing to notice misclassifications noted by all other participants). Using GPOWER [84], the power analysis showed this N to be appropriate for a moderate effect-size. This study was also passed in a review by the university's ethics board (ref. LS-E-19-148-Kenny-Keane).

7.1.2. Materials

Thirty materials were generated from the CNN classifier using the MNIST dataset as before. The misclassifications were actual errors produced by the model (i.e., query-items where the classification made differed from the ground truth): 3% error-rate (had 1 wrong classification), 30% error-rate (had 9 wrong classifications), and 60% error-rate (had 18 wrong classifications). All errors were alternate-labelling errors in which the model gave a close but incorrect classification (see Fig. 4); note, majority-voting errors were too rare to test in this experiment. Explanations were presented as in Expt. 1 using the three nearest-neighbor images for the test-instance, from the MNIST training-set (see Fig. 3 and Fig. 4 for examples).

7.1.3. Procedure, Measures, and Analyses

The procedure was identical to that used in Expt. 1. As before, measures were correctness and reasonableness ratings (on 5-point scales). After rating all of the presented classifications, participants filled out the DARPA trust (8 questions) and satisfaction surveys (8 questions). MANOVAs were computed for the independent variables (Explanation and Error-Rate) involving the dependent variables of (i) right/wrong classifications for correctness ratings, (ii) right/wrong classifications for reasonableness ratings, (iii) 8 trust-question ratings, and (iv) 8 satisfaction-question ratings.

7.2. Experiment 2: Quantitative Results

Again it was found that when example-based explanations were given people perceive misclassifications as being more correct (replicating Expt. 1); people rate wrong classifications as more correct, with an explanation, presumably because it shows the model working consistently but with mislabelled data (see Fig. 9). As before, explanations were found not to impact reasonableness ratings. Increasing error-rates negatively impact people's ratings of correctness, reasonableness, and trust; notably, models with error-rates of 30-60%, are trusted significantly less than ones with a 3% error-rate.

7.2.1. Correctness Ratings

The MANOVA analyses of correctness ratings revealed significant effects for Explanation, $F(2,158) = 3.129$, $p < .05$, Wilks' $\Lambda = 0.962$, partial η^2=0.04, and Error-Rate , $F(4, 316) = 14.976$, $p < .001$, Wilks' $\Lambda = 0.707$, partial η^2=0.16. The interaction was not statistically significant. However, all of these effects occur in people's ratings of wrong classifications, not in right classifications. Univariate analyses for the wrong classifications showed main effects for the Explanation ($p < .05$; see Table 2), and Error-Rate variables ($p < .001$; see Table 2). The analyses for right classifications show no significant effects (all $p > .30$). Interestingly, Tukey HSD *post-hoc* comparisons showed people rate the wrong classifications as being more correct in the Explanation-Present ($M = 1.82$, $SD = 0.72$) than in the Explanation-Absent ($M = 1.62$, $SD = 0.58$) condition ($p < .05$). In contrast, people's ratings of the right classifications are not reliably different (Explanation-Present, M=4.73, SD=0.37; Explanation-Absent, M=4.67, SD=0.48; see Fig. 9). Error-Rate also impacts correctness ratings for wrong classifications, $F(2, 159) = 31.61$ $p < .001$, partial $\eta^2 = .284$, but not right classifications, $F(2, 159) = 1.45$, $p > .20$, partial $\eta^2 = .018$ (see Table 2 and Fig. 9). However, the effect is somewhat counter-intuitive; as people see more misclassifications (30% or 60% versus 3%) they tend to rate the misclassifications as being marginally more correct; 3% ($M = 1.23$, $SD = 0.50$), 30% ($M = 1.93$, $SD = 0.54$), 60% ($M = 1.99$, $SD = 0.64$). Tukey HSD comparisons for these wrong classifications show the differences between the 3-30% and 3-60% conditions to be reliable (all $p < .001$). None of the pairwise comparisons for the right classifications were reliably different.

7.2.2. Reasonableness Ratings

The pattern of results for reasonableness ratings is less clear. The MANOVA analyses of reasonableness ratings revealed a significant interaction between Explanation and Error-Rate , $F(4, 316) = 2.53$, $p < .05$, Wilks' $\Lambda = .94$, partial $\eta^2 = 0.03$. No main effects were statistically significant. Univariate analyses showed this Explanation x Error-Rate interaction occurs only for the right classifications, $F(2, 159) = 4.27$, $p < .05$, partial $\eta^2 = .05$; this interaction was specifically due to people rating the 60%-error condition as being more reasonable in the Explanation-Present ($M = 4.78$, $SD = .09$) than in the Explanation-Absent ($M =$

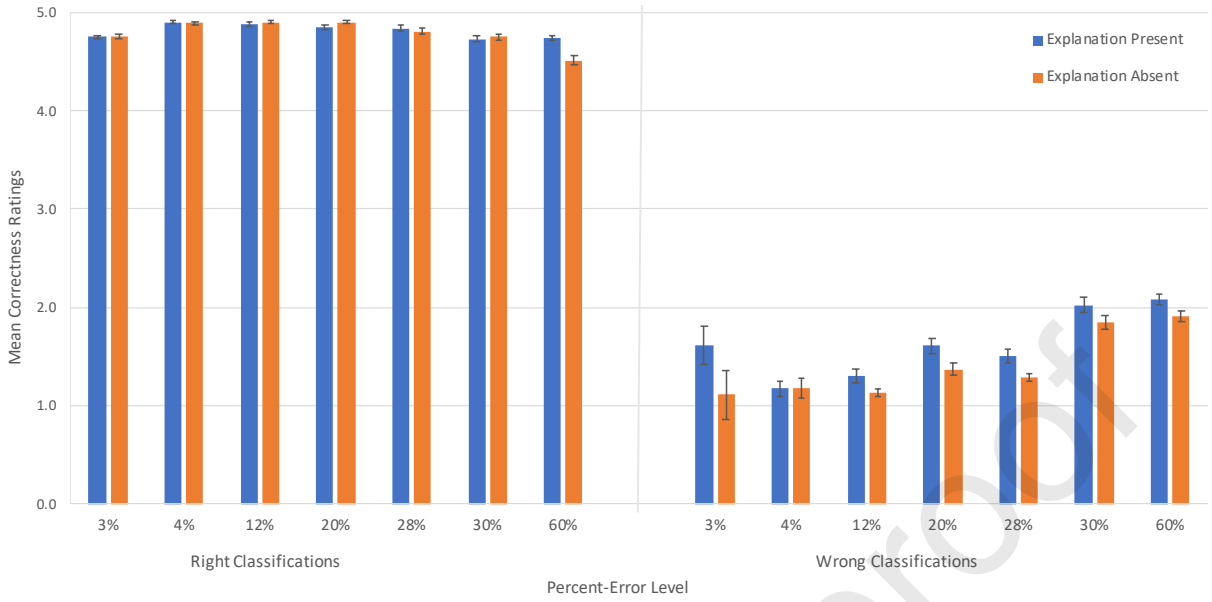


Figure 9: Experiments 1-3 Combined Results: Mean correctness ratings of right and wrong classifications when Explanations are present or absent for each Error-Rate (Standard Error bars are shown).

4.38, $SD = .08$) condition, a difference that is statistically reliable ($p < .05$). It is hard to interpret this finding. Expt. 1 found that reasonableness ratings were not impacted by the provision of explanations when the error-rate was 20%. As such, reasonableness may not be a robust measure; it may suffer from different people having quite different interpretations of what reasonableness means.

7.2.3. Trust

The MANOVA analyses of the trust survey revealed a significant effect of Error-Rate , F(16, 304) = 5.234, $p < .001$, Wilks' $\Lambda = 0.616$, partial $\eta^2 = 0.22$. There was no main effect or interaction of Explanation. Univariate main effects for Error-Rate were found in all questions (all $p < .05$), except for Question 7 ("The system can perform the task better than a novice human user"; see Table 1 for a list of all questions). Two interesting *post-hoc* Tukey HSD comparisons indicated that overall people in the 3%-error conditions reported the highest level of confidence in the system ($M = 4.28$, $SD = .67$), and rated the classifications as most predictable ($M = 4.21$, $SD = .82$). Also, users in the 3%-error condition enjoyed using the system ($M = 3.11$, $SD = .96$) significantly more than the people in the 60%-error one ($M = 2.55$, $SD = 1.15$). In summary, trust is mainly impacted by the error-rates people encounter rather than the provision of an explanation. The explanation seems to act at an item-level, affecting people's perception of the correctness of misclassifications, but those explanations do not "explain away" the perception of those failures at the system-level. Trust is impacted by rising error-rates, though not linearly; from this study, it appears that trust levels decrease sharply at 30%-errors (relative to 3%-errors) and then stay around this level for 60%-errors.

7.2.4. Satisfaction

The MANOVA analyses of satisfaction ratings of the overall system revealed a significant effect for Error-Rate , F (16, 304) = 1.755, $p < .05$, Wilks' $\Lambda = .838$, partial $\eta^2 = .085$. Pairwise comparisons indicated that Questions 6 and 7 ("This explanation of how the program works is useful to my goals," and "This explanation of the program shows me how accurate the program is") were significant at the 30% and 60%

Error-Rates, $p < .05$ for each question. The satisfaction survey questions do not tell us much in this test context (similar results were found in Expt. 1). It is unclear whether this is an issue with the measures or this particular task. Part of the problem here may be that the satisfaction questions (unlike the trust ones) seem to range over a number of distinctly different issues, and treating them as a unitary set is not sensible [85].

8. Experiment 3: The Effects of Explanation and Low-End Error-Rates

Expt. 2 tested the impacts of *post-hoc* explanations and error-rates for high-end error-rates (30% and 60%, compared to 3%). Expt. 3 (N=184) tested the Explanation and Error-Rate variables using low-end error-rates to gain a fuller profile of error-rate effects. So, the design was a 2 (Explanation: present v. absent) x 4 (Error-Rate : 4% v. 12% v. 20% v. 28%) x 2 (Classification-Type: Right v. Wrong) with Explanation and Error-Rate being between-participants variables and Classification-Type being a within-participant one. It also introduced two new system-level measures, "overall correctness" and "overall reasonableness"; so, after participants had seen all the items, they were asked to rate the system as a whole, on both measures.

8.1. Experiment 3: Method

8.1.1. Participants

One hundred and eighty-four people were randomly assigned to the 8 groups in the study with roughly equal numbers in each condition (Ns = 23). As before, all participants were aged over 18, native English speakers, and lived in the USA, UK, or Ireland. Exclusion criteria were participation in previous studies by the lab and inattentive answering (1 person excluded). Using GPOWER [84], the power analysis showed this N to be appropriate for a moderate effect-size. This study was also passed in a review by the university's ethics board (ref. LS-E-19-148-Kenny-Keane).

8.1.2. Materials

Twenty-five materials were generated from the CNN classifier using the MNIST dataset as before: 4% error-rate (had 1 wrong classification), 12% error-rate (had 3 wrong classifications), 20% error-rate (had 5 wrong classifications), and 28% error-rate (had 7 wrong classifications). All errors were alternate-labelling errors in which the model gave a close but incorrect classification (see Fig. 4). As in the previous experiments, explanations used the three nearest-neighbor images, from the MNIST training-set (see Fig. 3 and Fig. 4 for examples).

8.1.3. Procedure, Measures, and Analyses

The procedure was identical to that used the previous experiments. As before, measures were correctness and reasonableness ratings (on 5-point scales). After rating all of the presented classifications, participants filled out the DARPA trust (8 questions) and satisfaction surveys (8 questions). In this experiment two new system-level measures were used; before completing these surveys, participants also rated the system's overall correctness and reasonableness on the presented items. This measure was designed to assess people's system-level sense of the model on correctness and reasonableness. As in Expt. 2, MANOVAs were computed for the independent variables (Explanation and Error-Rate) involving the dependent variables of (i) right/wrong classifications for correctness ratings, (ii) right/wrong classifications for reasonableness ratings, (iii) overall system-level correctness/reasonableness ratings, (iv) 8 trust-question ratings, and (v) 8 satisfaction-question ratings.

8.2. Experiment 3: Quantitative Results

The results replicate the patterns of results found for the Explanation and Error-Rate variables on the correctness and trust measures (in Expt. 1 and 2) and, again, showed few/no effects for the reasonableness and satisfaction measures. However, the ratings of the system's overall correctness and overall reasonableness provided interesting converging evidence.

8.2.1. Correctness and Reasonableness Ratings

The MANOVA analyses of correctness ratings revealed significant effects for Explanation, $F(2, 175) = 4.12$, $p < .05$, Wilks' $\Lambda = 0.955$, partial $\eta^2 = 0.045$, and Error-Rate, $F(6, 350) = 3.89$, $p = .001$, Wilks' $\Lambda = 0.879$, partial $\eta^2 = .063$. The interaction was not statistically significant. As before, all of these effects occurred in people's ratings of wrong classifications, not of right classifications. Univariate main effects for the wrong classifications were found for the Explanation ($p < .01$; see Table 2) and Error-Rate variables ($p < .001$; see Table 2). But, no significant effects were found for these variables on right classifications (all $p > .30$). For example, overall people in the Explanation-present condition rated wrong classifications as more correct ($M = 1.40$, $SD = 0.41$) than those in the Explanation-absent condition ($M = 1.24$, $SD = 0.37$), but both conditions were almost identical for right classifications. Similarly, the Error-Rate variable's effects all occur in wrong classifications ($p < .001$; see Table 2), not right classifications, $F(3, 176) = 1.23$, $p > .30$, partial $\eta^2 = 0.02$. Indeed, this analysis echoed the result from Expt. 2; namely, that the more errors people see, the more they rated them as being correct. *Post-hoc* Tukey HSD comparisons revealed a significant 0.31 mean increase ($p = .001$) from the 4% to the 20%-error conditions, as well as a 0.27 mean increase ($p < .05$) from the 4% to 28%-error conditions. There was also a significant 0.27 mean increase ($p < .01$) from the 12%-error to the 20%-error condition (see Fig. 9). MANOVA analyses of reasonableness revealed no reliable effects.

			Explanation			Error-Rate		
	Measure	Classification Error Type	*df*	*F*	partial η^2	*df*	*F*	partial η^2
Expt.1	Correctness	Alternate-Labelling	1, 100	***11.13	0.10	-	-	-
		Majority-Voting	1, 100	***9.15	0.08			
	Reasonableness	Alternate-Labelling	1, 100	3.68	0.04	-	-	-
		Majority-Voting	1, 100	0.94	0.01	-	-	-
	Trust		1, 100	0.79	0.01	-	-	-
	Satisfaction		1, 100	1.39	0.01	-	-	-
Expt.2	Correctness		1, 159	*4.91	0.03	2, 159	***31.61	0.28
	Reasonableness		1, 159	3.40	0.02	2, 159	2.51	0.03
	Trust		1, 159	0.00	0.00	2, 159	***15.95	0.17
	Satisfaction		1, 159	0.13	0.00	2, 159	0.84	0.01
Expt.3	Correctness		1, 176	**8.24	0.05	3, 176	***7.11	0.11
	Reasonableness		1, 176	1.41	0.01	3, 176	0.15	0.00
	Global Correctness		1, 175	1.47	0.01	3, 175	***15.34	0.21
	Global Reasonableness		1, 175	0.04	0.00	3, 175	***10.87	0.16
	Trust		1, 176	1.00	0.01	3, 176	***6.89	0.11
	Satisfaction		1, 176	2.79	0.02	3, 176	0.05	0.00

*$p < .05$; **$p < .01$; ***$p < .001$. Italicized measures are item-level measures of misclassifications, all others being system-level measures.

Table 2: Expt. 1-3 Results Summary: ANOVA Analyses for misclassifications, for all measures tested on the Explanation and Error-Rate Variables [n.b., for Expt. 1 the correctness and reasonableness ratings cover the three classification-types used in that experiment].

8.2.2. System-Level Measures: Overall Correctness and Overall Reasonableness

The MANOVA analyses of overall correctness and reasonableness ratings revealed significant effects for Error-Rate, $F(6, 348) = 9.43$, $p < .001$, Wilks' $\Lambda = 0.74$, partial $\eta^2 = 0.14$. No statistically significant effects were found for Explanation or the interaction. The Error-Rate effect was significant both for the overall correctness score ($p < .001$; see Table 2); as well as for the overall reasonableness score ($p < .001$; see Table 2). However, most of this effect can be attributed to the 4%-error condition. *Post-hoc* Tukey HSD comparisons revealed a higher overall correctness and reasonableness ratings for the system in the 4%-error condition versus all others. So, this system-level measure of correctness (like the trust measure) shows that the item-level effects for Explanation do not persist into a system-level evaluation of correctness. The Error-Rate 's variable was also found to impact overall correctness and overall reasonableness; see Table 2), showing people are sensitive to changes in error-rates from 4% to 12% and 28%.

8.2.3. Trust and Satisfaction

The MANOVA analyses of trust ratings of the overall system also revealed a significant effect for Error-Rate, $F(24, 490) = 1.85$, $p = .009$, Wilks' $\Lambda = 0.78$, partial $\eta^2 = 0.08$. No statistically significant effects were found for Explanation or an interaction (for ANOVA analyses see Table 2). Univariate main effects for Error-Rate were found in for half of the 8 questions (1, 2, 4, 6). Overall, people in the 4%-error condition reported the highest level of confidence in the system ($M = 4.43$, $SD = 0.54$), found the system to be more reliable ($M = 3.24$, $SD = 1.02$) and reported trusting the system to make accurate classifications ($M = 3.93$, $SD = 0.8$) than for other Error-Rate levels. In contrast, the MANOVA analyses of satisfaction ratings show no main effects or interaction, although there were pair-wise differences for specific questions under the Explanation variable. Specifically, the Explanation-Present group rated Question 1 ("From the explanation, I understand how the program works"), Question 3 ("This explanation of how the program works has sufficient detail"), and Question 5 ("This explanation of how the program works tells me how to use it") significantly higher than the Explanation-Absent group, all $p < .05$.

9. General Discussion and Conclusions

In this paper, we have presented an end-to-end treatment of a model-agnostic solution to the problem of Explainable AI (XAI) involving *post-hoc* explanations. Specifically, we have advanced a new computational method for finding example-based explanations, within a twin-systems approach, implementing this method in a CNN-CBR twin using the MNIST dataset. Then, we systematically tested for the impact of these explanations on right and wrong classifications across three user-studies, where we also varied the error-rates encountered by users. As such, we hope that this work provides a potential methodological blueprint for how to evaluate XAI techniques. To summarize, the main contributions are:

- User's mental models can be differentiated into sub-models of the domain, AI-system, and explanation strategy. These can then be used to "parse" an experimental paradigm to better understand where mental model development is occurring.
- A feature-weighting technique – COLE-HP – can capture an ANN's feature-weights and use them in a CBR-system to find *post-hoc* explanatory examples for a twinned, ANN-CBR system (n.b., [22] have applied this technique to many different datasets and domains).
- These *post-hoc* example-based explanations impact people's mental models of misclassifications made by a black-box AI model; people judge individual errors to be more correct when explanations are given, an item-level effect that is replicable across three experiments (albeit with small-to-medium effect sizes). Hence, people view the AI model to be *acting in a way that is more correct* (or, at least, *less incorrect*) when given an explanation (which makes sense, as often it is the data which is in error, not the model). In short, there is evidence (qualitiative and quantitive) that these explanations impact people's understanding of the causal role the model plays in misclassifications.

- The present evidence shows that explanations mainly impact people's judgements of errors, not correct items. This result could be taken to suggest that correct classifications do not require explanation. However, we do not think such a conclusion is warranted; the effect we see here may be due to people's high-expertize with the domain and the measures used. To put it another way, we believe there may well be circumstances in which these explanations play a role with correct items too; though these circumstances remain to be demonstrated.

- Error-rates also impact people's perceptions of the overall system's competence (in system-level measures of trust, overall correctness, and overall reasonableness); as error-rates increase, anywhere above a low threshold of 3-4%, people trust the system less and discount its overall correctness/reasonableness (see Fig. 9). Indeed, the effects of error-rates look more like a step-function than a trend across when one views the relative changes in key measures (see Fig. 9).

- The item-level effects of these explanations (on correctness judgments of errors) do not aggregate into overall improvements in people's system-level evaluations of the model (e.g., in trust or satisfaction); indeed, a qualitative analysis of people's verbal reports on the model suggest that being given explanatory-examples may simply serve to highlight weaknesses in the model's operation (i.e., they blame-the-system more).

- Reasonableness and Satisfaction measures appear not to be as informative as the other measures; both seem to engender very different interpretations in the user population. Although several satisfaction questions reached a significant difference between groups, the findings were not replicated across experiments.

In conclusion, the present studies present a rich set of findings for wider consideration of the dynamics of user interactions with explanation-strategies and error-rates in XAI research. This work also adds significantly, to the (now) small pool of carefully-controlled user studies on how explanations and error-rates impact AI systems. It also suggests a number of lessons for future work that we try to extract and generalize in the next sub-section.

9.1. Lessons Learned For User Tests in XAI

From the present concerted effort to develop a new XAI technique and then evaluate it in a series of user studies, a number a lessons present themselves.

Excellent Computational Explanations May Not Be Good Psychological Explanations. At present, a whole slew of XAI techniques are being developed and, without user studies, it is not at all clear whether any of them are psychologically valid. Here, we have tested a long-heralded XAI technique, using *post-hoc* explanation-by-example, and found that these explanations have quite circumscribed impacts on people's mental models. Just as vaccines go through successive primary, secondary, and tertiary trials before being used, it might be good to have similar stages in assessing XAI methods (perhaps with health warnings on trust-appropriateness and fairness).

Item-level Impacts May Not Sum to System-level Impacts. In the present work, a distinction was made between item-level impacts and system-level impacts that deserves some consideration; it is generally assumed that the item-level impacts of explanation provision (e.g., speed-up for a user on an item or changed perception of an item) aggregate into some overall system-level impact (e.g., overall performance is faster or overall trust in the system is better); however, we have seen that item-level impacts may not necessarily aggregate to system-level effects. Many current studies in XAI show item-level effects or explore a small number of items in tests, without necessarily assessing whether the effects found aggregate to overall system-level impacts; assessments at both levels are clearly needed.

People's Tolerance of Failure is Low. The overwhelming evidence from this work (and previous studies) is that people's tolerance of failure in automated systems is low (aka *algorithmic aversion*). This growing body of evidence suggests that the bar for using predictive AI models is very high; basically, if the model is not near-perfect then it is highly likely that users will have trust issues or an adverse response (though, the influence of expertize in the domain being used, needs to be assessed).

Testing as Close to the "Real" Task-Context is Critical. Earlier, we reviewed a threefold taxonomy of XAI evaluation from application-grounded, to human-grounded, to functionally-grounded evaluations [27] where from first-to-last one moves further away from the user and the "real" use-case for the system. Given the observed complexities of the interactions seen here between task demands, presented classification types, relative proportions of errors, and the presence/absence of explanations, it is hard to see how diverging from the "real" use-case could tell one much about the utility of a given XAI-technique/model. Recently, Buccinca *et al.* [20] made similar arguments about what they call, "proxy" user-testing scenarios (i.e., ones less like the human-grounded evaluations), after showing that they generate different behaviour in users than "real" use-case scenarios, making them questionable for evaluation purposes (though there were some methodological issues with their tests).

Paradigm Templates Would Help. People's understanding of these AI systems is based on a multi-faceted mental model with at least three component sub-models (i.e., for the domain, the AI method, and explanation strategy, wrapped in a task-goal context) and user tests need to control and/or carefully vary interventions to impact these models; here, we advanced a paradigm in which the domain and explanation sub-models were held relatively constant allowing us some insight into how people's mental model of system was impacted. XAI needs a set of template, benchmark paradigms of this sort to guide future testing (e.g., [21] provide another good template where they vary the bias of the dataset systematically and also compare multiple explanation strategies on the same task).

Robust Benchmark Measures Are Needed. This research area needs to identify reliable and robust benchmark measures for assessing the impact of explanation strategies; we have seen that correctness ratings seem to consistently assess people's evaluations of model predictions, whereas reasonableness ratings did not (perhaps because people have different understandings of what is reasonable). There are probably a relatively small number of benchmark measures that work in these user studies that should be identified by the field and then re-used (e.g., the DARPA trust questions used here seem quite robust, whereas the the satisfaction questions seemed less informative).

9.2. Future Directions

There are at least four main directions in which future research could be taken. Firstly, the present twin-systems only build explanations based on factual cases (a.k.a., nearest neighbors); yet, we know that people find semi-factual and counterfactual cases very compelling [41, 46, 40, 86]. Considering this, the current twin-system framework could be extended to include these other forms of case-based explanations (see e.g., Keane and Smyth [45]). Secondly, we have shown how twin-systems can be applied to one type of deep learning method; other deep learning techniques could be explored (e.g., BERT-CBR for NLP and GAN-CBR twins for generative tasks). Thirdly, the twinning approach may have some potential to provide global explanations for CNNs by summarizing feature activation maps [22] for a class similar to other approaches [87]. Lastly, there is a larger research program of user testing of all XAI techniques needed to determine whether many popular current methods hold up to an examination of their psychological validity.

Acknowledgments

This paper emanated from research funded by (i) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics (12/RC/2289-P2), (ii) SFI and DAFM on behalf of the Government of Ireland to the VistaMilk SFI Research Centre (16/RC/3835), and (iii) the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

References

[1] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law 7 (2) (2017) 76–99.

[2] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI magazine 38 (3) (2017) 50–57.

[3] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica, May 23 (2016) 2016.

[4] D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2.

[5] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2016, pp. 1135–1144.

[6] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140.

[8] N. Frosst, G. Hinton, Distilling a neural network into a soft decision tree, arXiv preprint arXiv:1711.09784.

[9] C. Chen, O. Li, A. Barnett, J. Su, C. Rudin, This looks like that: deep learning for interpretable image recognition, arXiv preprint arXiv:1806.10574.

[10] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai, arXiv preprint arXiv:1902.01876.

[11] Z. C. Lipton, The mythos of model interpretability, arXiv preprint arXiv:1606.03490.

[12] Z. C. Lipton, The mythos of model interpretability, Queue 16(3) (2018) 30.

[13] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), Vol. 8, 2017, p. 1.

[14] F. Sørmo, J. Cassens, A. Aamodt, Explanation in case-based reasoning–perspectives and goals, Artificial Intelligence Review 24 (2) (2005) 109–143.

[15] A. J. Johs, M. Lutts, R. O. Weber, Measuring explanation quality in xcbr, ICCBR 2018 (2018) 75.

[16] D. Leake, D. McSherry, Introduction to the special issue on explanation in case-based reasoning, The Artificial Intelligence Review 24 (2) (2005) 103.

[17] D. Leake, Cbr in context: the present and future. case based reasoning experiences-lessons and future experiences. d. leake (1996).

[18] M. T. Keane, E. M. Kenny, How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems, in: International Conference on Case-Based Reasoning, Springer, 2019, pp. 155–171.

[19] C. Nugent, P. Cunningham, D. Doyle, The best way to instil confidence is by being right, in: International Conference on Case-Based Reasoning, Springer, 2005, pp. 368–381.

[20] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 454–464.

[21] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, C. Dugan, Explaining models: an empirical study of how explanations impact fairness judgment, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 275–285.

[22] E. M. Kenny, M. T. Keane, Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ann-cbr twins for xai, in: Twenty-Eighth International Joint Conferences on Artifical Intelligence (IJCAI), Macao, 10-16 August 2019, 2019, pp. 2708–2715.

[23] Y. LeCun, C. Cortes, C. Burges, Mnist handwritten digit database, ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist 2.

[24] A. Bäuerle, H. Neumann, T. Ropinski, Training de-confusion: An interactive, network-supported visual analysis system for resolving errors in image classification training data, arXiv preprint arXiv:1808.03114.

[25] K. J. . A. D. Glickenhaus, B., Darpa xai phase 1 evaluations report, in: DARPA XAI Program, Report, 2019.

[26] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608.

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

[28] M. W. Eysenck, M. T. Keane, Cognitive psychology: A student's handbook, Taylor & Francis, 2005.

[29] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 93.

[30] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[31] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, in: Proceedings of the 2018 CHI conference on human factors in computing systems, ACM, 2018, p. 582.

[32] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: 2007 IEEE 23rd International Conference on Data Engineering Workshop, IEEE, Istanbul, Turkey, 2007, pp. 801–810.

[33] G. H. Harman, The inference to the best explanation, The philosophical review 74 (1) (1965) 88–95.

[34] B. C. Van Fraassen, et al., The scientific image, Oxford University Press, 1980.

[35] W. C. Salmon, Scientific explanation and the causal structure of the world, Princeton University Press, 1984.

[36] F. C. Keil, Explanation and understanding, Annu. Rev. Psychol. 57 (2006) 227–254.

[37] L. H. Gilpin, C. Testart, N. Fruchter, J. Adebayo, Explaining explanations to society, arXiv preprint arXiv:1901.06560.

[38] E. M. Kenny, E. Ruelle, A. Geoghegan, L. Shalloo, M. O'Leary, M. O'Donovan, M. T. Keane, Predicting grass growth for sustainable dairy farming: A cbr system using bayesian case-exclusion and post-hoc, personalized explanation-by-example (xai), in: International Conference on Case-Based Reasoning, Springer, 2019, pp. 172–187.

[39] E. M. Kenny, E. Ruelle, A. Geoghegan, L. Shalloo, M. O'Leary, M. O'Donovan, M. Temraz, M. T. Keane, Bayesian case-exclusion and personalized explanations for sustainable dairy farming.

[40] E. M. Kenny, M. T. Keane, On generating plausible counterfactual and semi-factual explanations for deep learning, arXiv preprint arXiv:2009.06399.

[41] R. M. Byrne, Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.

[42] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[43] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 279–288.

[44] A. Lucic, H. Haned, M. de Rijke, Contrastive explanations for large errors in retail forecasting predictions through monte carlo simulations, arXiv preprint arXiv:1908.00085.

[45] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: Under Review for – International Conference on Case-Based Reasoning, 2020.

[46] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, Journal of Intelligent Information Systems 32 (3) (2009) 267–295.

[47] G. A. Klein, Do decision biases explain too much, Human Factors Society Bulletin 32 (5) (1989) 1–3.

[48] M. S. Cohen, J. T. Freeman, S. Wolf, Metarecognition in time-stressed decision making: Recognizing, critiquing, and correcting, Human Factors 38 (2) (1996) 206–219.

[49] J. H. Park, K. H. Im, C.-K. Shin, S. C. Park, Mbnr: case-based reasoning with local feature weighting by neural network, Applied Intelligence 21 (3) (2004) 265–276.

[50] C. K. Shin, S. C. Park, Memory and neural network based expert system, Expert Systems with Applications 16 (2) (1999) 145–155.

[51] C.-K. Shin, U. T. Yun, H. K. Kim, S. C. Park, A hybrid approach of neural network and memory-based learning to data mining, IEEE Transactions on Neural Networks 11 (3) (2000) 637–646.

[52] K. H. Im, S. C. Park, Case-based reasoning and neural network based expert system for personalization, Expert Systems with Applications 32 (1) (2007) 77–85.

[53] R. Caruana, H. Kangarloo, J. Dionisio, U. Sinha, D. Johnson, Case-based explanation of non-case-based learning methods., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 1999, p. 212.

[54] M. T. Keane, E. M. Kenny, The twin-system approach as one generic solution for XAI: An overview of ANN-CBR twins for explaining deep learning, arXiv preprint arXiv:1905.08069.

[55] B. Kim, C. Rudin, J. A. Shah, The bayesian case model: A generative approach for case-based reasoning and prototype classification, in: Advances in neural information processing systems, 2014, pp. 1952–1960.

[56] M. T. Keane, E. M. Kenny, How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems, arXiv preprint arXiv:1905.07186.

[57] P. Cunningham, D. Doyle, J. Loughrey, An evaluation of the usefulness of case-based explanation, in: International Conference on Case-Based Reasoning, Springer, 2003, pp. 122–130.

[58] D. Doyle, P. Cunningham, D. Bridge, Y. Rahman, Explanation oriented retrieval, in: European Conference on Case-Based Reasoning, Springer, 2004, pp. 157–168.

[59] C. Nugent, P. Cunningham, A case-based explanation system for black-box systems, Artificial Intelligence Review 24 (2) (2005) 163–178.

[60] S. K. Biswas, M. Chakraborty, H. R. Singh, D. Devi, B. Purkayastha, A. K. Das, Hybrid case-based reasoning system by cost-sensitive neural network for classification, Soft Computing 21 (24) (2017) 7579–7596.

[61] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3145–3153.

[62] N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, arXiv preprint arXiv:1803.04765.

[63] J. Kolodner, Case-based reasoning, Morgan Kaufmann, 2014.

[64] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, arXiv preprint arXiv:1904.07451.

[65] D. Gunning, D. W. Aha, Darpa's explainable artificial intelligence program, AI Magazine 40 (2) (2019) 44–58.

[66] C. J. Cai, J. Jongejan, J. Holbrook, The effects of example-based explanations in a machine learning interface, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 258–262.

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

[68] J. W. Burton, M.-K. Stein, T. B. Jensen, A systematic review of algorithm aversion in augmented decision making, Journal of Behavioral Decision Making 33 (2) (2020) 220–239.

[69] R. R. Hoffman, M. Johnson, J. M. Bradshaw, A. Underbrink, Trust in automation, IEEE Intelligent Systems 28 (1) (2013) 84–88.

[70] P. de Vries, C. Midden, D. Bouwhuis, The effects of errors on system trust, self-confidence, and the allocation of control in route planning, International Journal of Human-Computer Studies 58 (6) (2003) 719–735.

[71] T. Inagaki, N. Moray, M. Itoh, Trust, self-confidence and authority in human-machine systems, IFAC Proceedings Volumes 31 (26) (1998) 431–436.

[72] J. Lee, N. Moray, Trust, control strategies and allocation of function in human-machine systems, Ergonomics 35 (10) (1992) 1243–1270.

[73] S. M. Merritt, D. R. Ilgen, Not all trust is created equal: Dispositional and history-based trust in human-automation interactions, Human Factors 50 (2) (2008) 194–210.

[74] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International journal of human-computer studies 58 (6) (2003) 697–718.

[75] B. J. Dietvorst, J. P. Simmons, C. Massey, Algorithm aversion: People erroneously avoid algorithms after seeing them err., Journal of Experimental Psychology: General 144 (1) (2015) 114.

[76] D. A. Cafarelli, Effect of false alarm rate on pilot use and trust of automation under conditions of simulated high risk, Ph.D. thesis, Massachusetts Institute of Technology (1998).

[77] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, in: European Conference on Computer Vision, Springer, 2016, pp. 3–19.

[78] A. S. Ross, F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Thirty-second AAAI conference on artificial intelligence, 2018.

[79] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400.

[80] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[81] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[82] C. Wang, B. Han, B. Patel, F. Mohideen, C. Rudin, In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction, arXiv preprint arXiv:2005.04176.

[83] T. Le Nguyen, S. Gsponer, I. Ilie, M. O'Reilly, G. Ifrim, Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations, Data Mining and Knowledge Discovery 33 (4) (2019) 1183–1222.

[84] E. Erdfelder, F. Faul, A. Buchner, Gpower: A general power analysis program, Behavior research methods, instruments, & computers 28 (1) (1996) 1–11.

[85] R. Hoffman, Personal communication to m.t. keane.

[86] L. Yang, E. Kenny, T. L. J. Ng, Y. Yang, B. Smyth, R. Dong, Generating plausible counterfactual explanations for deep transformers in financial text classification, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6150–6160.

[87] A. Ghorbani, J. Wexler, J. Y. Zou, B. Kim, Towards automatic concept-based explanations, in: Advances in Neural Information Processing Systems, 2019, pp. 9277–9286.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: