# Cases are King: A User Study of Case Presentation to Explain CBR Decisions

Lawrence Gates, David Leake, Kaitlynne Wilkerson

Luddy School, Indiana University, Bloomington IN 47408, USA
{gatesla,leake,kwilker}@indiana.edu

**Abstract.** From the early days of case-based reasoning research, the ability of CBR systems to explain their decisions in terms of past cases has been seen as an important advantage. However, there have been few studies on the factors affecting the effectiveness of explaining CBR decisions by cases. This paper presents results from a human subjects study that examined how alternative retrieval processes (one-shot or conversational) and case presentation approaches affect the perceived goodness of case-based explanations for explaining system behavior, their convincingness, and the trust they engender. The study corroborates that cases are well received as explanations, with some benefit for providing information to support similarity comparison, and suggests that elucidating the retrieval process has little effect on explanatory effectiveness.

**Keywords:** Case-based reasoning, conversational case-based reasoning, counterfactuals, explanation, interfaces, XAI, human subjects study

## 1 Introduction

The growing deployment of AI systems for high-impact tasks, coupled with governmental policies such as the EU GDPR regulations providing the right to "meaningful information" about AI system decisions, often called "right to explanation" [31], have spurred much interest into eXplainable AI (XAI) [9]. Case-based reasoning (CBR) has long been seen as well-suited to explanation because it is intrinsically interpretable, in that a CBR system can account for its decisions by presenting the cases on which its solutions are based [20]. Much research has studied methods to explain the CBR process and to leverage CBR to facilitate explanation of other AI methods, leading to an active CBR explanation community and a series of workshops on XCBR (e.g., [27]).

In both XAI and XCBR, much of the research focus has been technical, aimed at the development of methods to provide AI systems with new explanatory capabilities. However, human subjects studies to assess the response of users to those capabilities have been less widespread. Keane and Kenny's [14] survey of research on twinning CBR with neural networks for explanation described the "embarrassment of user testing," noting that only "a handfull" of the works that they surveyed on twinning included user tests; a survey of other aspects of explanation using CBR by Gates and Leake [8] found similarly sparse coverage.

In seminal work, Cunningham et al. [3] performed a human subjects study that supported the convincingness of cases as explanations, compared to rules. However, despite the importance of that work, important questions remain for knowing how cases can be used most effectively as explanations. One question is how best to present explanatory cases to users. Explanatory cases may be presented to users in different ways, including different types of contextualizing information; this raises the question of what forms of case presentation and contextual information may enhance the explanatory value of cases. If certain forms are more effective, knowledge of which to apply can provide "low-hanging fruit" for enhancing the explanatory benefit of CBR systems: system designers can adapt the case presentation interface accordingly.

A second question concerns the effect of different modes of interaction with a CBR system. Interactions between CBR systems and their users are commonly managed in one of two ways. In traditional "one-shot" CBR systems, a problem is presented to the system, which then presents its solution. In conversational case-based reasoning (CCBR), users provide information incrementally, guided by questions the system provides, with questions aimed at identifying the most similar case in the case base rapidly [1]. The conversational interaction can be seen as making the case retrieval process transparent, which might be seen as an additional implicit explanation of case relevance, which might increase the user's sense of understanding of the system process. If either interaction type affects user perception of explanation quality, designers could use that information to guide decisions of which interaction to select when both are applicable.[1]

A third question concerns the effects of case presentation on user perceptions of different measures of explanation quality. Cunningham et al. [3] focused on convincingness, the ability of the explanation to convince users that the decision was correct. In addition to convincingness, this paper considers effects on the *goodness* of the explanations for explaining current system decision-making and the *trust* they engender for future system decisions.

Our human subjects study tests the effects of three types of CBR system interactions—standard CBR, CCBR with question retrieval based on information gain, and CCBR retrieval based on individual feature importance—and of four explanation presentation designs. For each, we measure impact on the convincingness, goodness for understanding system behavior, and future trust of the system. We also test whether successive exposures to each explanation design affect the reported scores. Finally, we test whether scores are affected when incorrect solutions provided to users and by the level of similarity between the problem case and solution case presented to the user.

Analysis of our results suggests that the form of CBR interaction, one-shot or CCBR, was not important—only explanation type played a role in the observed goodness of, convincingness of, and trust in explanations. It also supported that the similarity of the prior case to the new situation had an impact on convincingness and trust. This suggests that the key aspect determining the usefulness

[1] When the user does not have a full description of the problem, using CCBR may be necessary to guide problem elaboration.

of explanations by cases is the case provided to explain the system decision. This provides support for the common CBR intuition that case presentation carries the primary burden for explaining CBR decisions. Our results also suggest that simply presenting the current problem and most similar case is an effective explanation approach, and that presentation of supplementary information about similarity in tabular form is helpful for all three criteria. Supplementing the nearest neighbor with the counterexample of the nearest unlike neighbor (NUN) was expected to improve explanations by helping users assess scope, but surprisingly was found to be detrimental for goodness, convincingness, and trust compared to the most similar case alone. However, no presentation variants were ranked negatively.

## 2 Related Work

*Metrics for Assessing Explanations:* XAI research has used a variety of metrics to assess the explanations generated by AI systems. Two of the most common metrics are trust and goodness, where goodness takes various forms [28]. It has been defined based on a wide variety of aspects of the explanation and the context in which it is presented, as well as in relationship to the explainer's purpose (e.g. [19, 32]), and its presentation. A central point is how the information and format of an explanation are received by a human [11], which has been assessed by criteria such as whether the explanation was easy to understand, and satisfying and useful towards understanding the domain and/or the AI system's reasoning [11, 28]. Another metric, convincingness, can be linked to discussions of goodness and trust. Previous studies examining the convincingness of explanations have relied on subjects assessing its common parlance meaning [3, 29], and we follow that approach.

Another common evaluation criterion is trust. Trust can be conceptualized in terms of user vulnerability and the extent to which the user accepts the risk present in an interaction [12, 23, 28]. In our study, we again rely on the user understanding the common parlance meaning of trust. The task domain we will use in our study, blood alcohol content estimation—which is used for breathalyzer tests of whether it is safe to drive—is one for which in principle risks could result from wrong predictions, which gives implicit stakes for the trust assessment.

*Human Subjects Studies on Explaining by Cases:* In a landmark human subjects study, Cunnningham et al. [3] compared case-based and rule-based explanations of whether the blood alcohol levels of drinkers was over or under the limit, given information about pub visits such as visit duration and number of drinks consumed. Subjects were presented with predictions in three conditions, with a case as explanation, with a rule-based explanation, or with no explanation, and asked the convincingness of the prediction. In their results case-based explanation decisively outperformed rule-based explanation.

A later study by Doyle et al. [6] considered explanation by cases for three domains, hospital admission and discharge decisions for bronchiolitis, an e-clinic

domain, and the blood alcohol domain. Subjects were provided with a system decision (e.g., a recommendation about admission or discharge), an explanation case (not necessarily the nearest neighbor but selected by a hand-coded utility function to be closer to the decision boundary [5]), justification text, and a confidence value. The justification text presented pros and the cons for the recommendation. Cases as explanations were useful for all domains, but with a split in the cases found most useful: cases selected by utility-based criteria were favored for bronchiolitis but not for the blood alcohol or e-clinic domains. Doyle et al. hypothesized that the difference was due to the increased complexity of the bronchiolitis domain, which made the directional effects less apparent.

Presentation of counterexamples enables a "compare and contrast" process to determine case applicability [2, 17, 21]. Doyle at al. also assessed the effects of presenting nearest unlike neighbors. Including this counterexample was found to be useful when subjects considered recommendations incorrect but overall had little effect.

Lamy et al. [18] performed a small scale (N=11) human subjects study of cases as explanation for a breast cancer domain, with primary focus on the benefit of visualization of similarity to explain case relevance; they reported a very positive response. Massie, Craw, and Wiratunga [24] studied the benefit of a visualization interface for explaining problem and solution similarities in a tablet formulation domain. A small-scale human subjects study (N=2) supported the usefulness of the visualizations of similarity and a preference for visualizations over text. McSherry [26] notes the subtlety that for prediction tasks, similarity between features that mitigate against the outcome may not strengthen the conclusion, and proposes an evidential approach that presents supporting and opposing features.

Kenny and Keane [15] conducted large-scale user studies (N=184) of the effects of post-hoc case-based explanations of black box systems. In their results,the primary impact of explanations was on mental models of misclassifications.

Warren et al. [33] conducted a human subjects study that compared counterfactual and causal explanations in a blood alcohol domain. It showed that counterfactuals were slightly more effective than causal explanations in improving user knowledge of the operation of an AI system (in terms of predictive accuracy), and which also raised concerns for a possible human tendancy to overestimate their own causal understanding.

## 3 Methods

The study protocol was approved by Indiana University's ethics review board (Indiana University IRB: 16546).

### 3.1 Participants

We recruited 110 participants via flyers, mailing lists, and word-of-mouth. They were paid $8 for participation in the 20-30 minute study. 89 participants cor-

rectly completed all 12 trials. We discarded the data of participants who did not correctly complete all trials. Each participant was randomly assigned to one of the system design groups, corresponding to one-shot CBR (26 participants) and to either of two CCBR groups, one with question order guided by information gain (30 participants) and the other with an alternative question ordering (33 participants). Section 3.2 provides more details on the CCBR systems.

Participants were generally young and well educated. Approximately 50% were born between 1994 and 2003, approximately 70% held a Bachelor's degree or higher. 90% had background in STEM. Concerning depth of understanding of AI mechanisms, 37% reported that they could write AI code, 29% could program but could not write AI code, 15% worked with AI-powered systems, and 19% had only heard about AI. Participants generally reported being interested in (70%) and excited (60%) about the progress being made in AI, while also being concerned about the prevalence of AI (53%) and how AI systems arrive at conclusions (97%). The majority of participants reported being neutral (41%) to open (40%) to trusting the information provided by AI systems. Given the homogeneity of our respondents backgrounds, our results may not be representative of individuals outside of these demographics.

### 3.2 Materials

**Case data set** Case data was a Blood Alcohol Content (BAC) data set of 85 cases collected from people leaving a pub [4], available online with documentation[2]. Cases included the categorical features Gender, Frame Size, Amount Consumed, Meal Consumed, and Duration, which were used to predict BAC. Approximately 52% of the cases were over the limit (0.8).

Each participant was presented with scenarios based on the same 12 cases[3]. The 12 were chosen to obtain a representative distribution of feature values and an even distribution of similarity levels between the problem case and the nearest neighbor that would be used in the explanation, to be able to assess whether similarity level affected goodness, convincingness, and trust. Features were weighted equally for similarity calculations. The set of selected cases included approximately 50% for which leave-one-out testing would generate an incorrect prediction, to assess whether system error affected participant judgments of goodness, convincingness, and trust. The user was not informed when predictions were incorrect. Ideally, explanation methods would tend to reveal likely errors and support higher ratings for correct solutions than for erroneous ones.

Participants interacted with one of the three CBR system variants whose results were presented in four explanation design templates. The systems performed a Blood Alcohol Content (BAC) prediction task to predict whether an individual's BAC level would be over or under the legal limit for driving.

[2] GitHub Link: https://github.com/gateslm/Blood-Alcohol-Domain

[3] The cases were presented in the following order (based on the case number in the dataset): 2, 8, 9, 12, 19, 33, 45, 48, 58, 82, 4, 54.

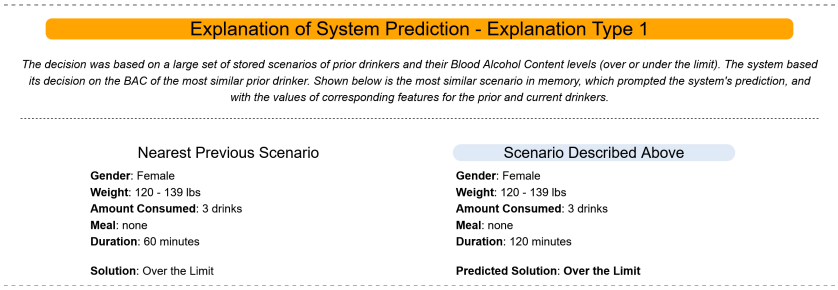
**System types** The study tested response to one of three CBR systems to assess how the interaction type and system process may effect observed levels of goodness, convincingness, and trust. The systems are a traditional CBR system and two versions of CCBR system, differing in question ordering:

- **CBR**: Traditional CBR; This system provides data fields to enter information; when all information is entered the system provides a prediction.
- **CCBR-IG**: CCBR - Information Gain (IG); Question ordering was determined by generating a prediction decision tree and ordering questions based on their first appearance in the tree, with the question process terminating when a unique best-match case is identified. (There were no ties in the examples used in the study.)
- **CCBR-CF**: CCBR - Combined Features (CF); The first questions asked concern the two features most predictive individually for the case base: amount of alcohol consumed and meal consumed. The other features were asked in the arbitrary order gender, frame size, and duration.

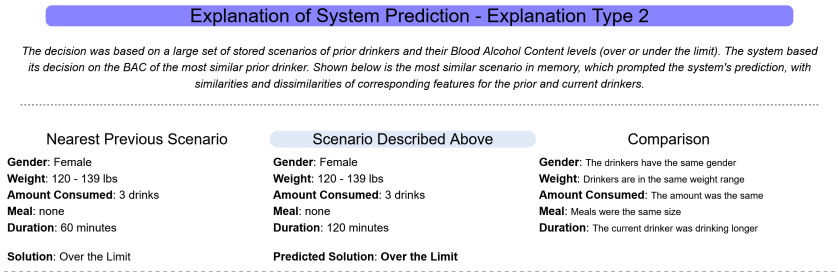
**Explanation designs** For each of the three systems, four types of explanations were tested for system predictions: Nearest Neighbor (NN), Nearest Neighbor+Similarity in two variants, one presenting similarity in tabular form and the other in textual form (NN+Sim:tab and NN+Sim:txt), and nearest neighbor with NUN as counterexample (NN+CE):

- **NN**: Nearest Neighbor (Fig. 1a); This presents the most similar case and its solution without any other information.
- **NN+Sim:tab**: Nearest Neighbor + Similarity Tabular Form (Fig. 1b); This augments the NN information with a brief tabular summary of similarities and differences between the nearest neighbor and the current problem.
- **NN+CE**: Nearest Neighbor + Counterexample (Fig. 1c); This presents two cases for comparison: the most similar case over the limit and the most similar case under the limit. This relates to the counterexample presentation studied by Doyle et al. [6] but differs in presenting the nearest unlike neighbor rather than selecting by their utility function. It parallels the "bracketing case" approach of Leake et al. [22].
- **NN+Sim:txt**: Nearest Neighbor + Similarity Textual Form (Fig. 1d); This presents information about the most similar case and its similarity to the current problem in textual form. Text passages were generated using a simple template-based generator.

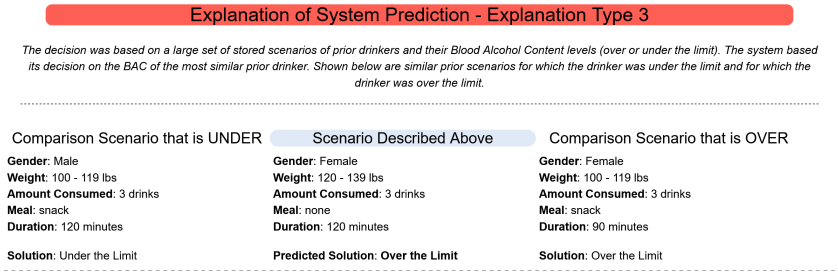
**Experimental data collection process** Data collection was online. The need to provide the interactive experience of a CCBR dialogue precluded using survey tools that simply present questionnaires. We used psiTurk, an interactive tool that can handle data collection while running Python code [7, 10].



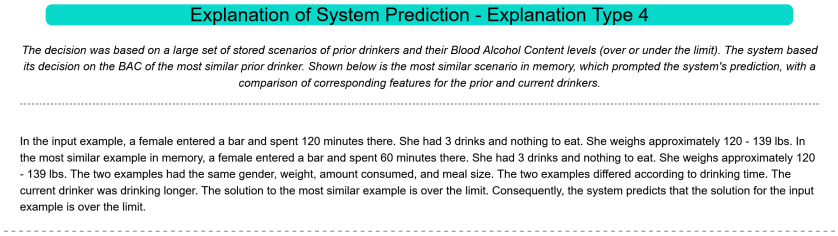

(a) Nearest Neighbor (NN)




(b) Nearest Neighbor + Similarity Tabular Form (NN+Sim:tab)




(c) Nearest Neighbor + Counterexample (NN+CE)




(d) Nearest Neighbor + Similarity Textual Form (NN+Sim:txt)

Fig. 1: Sample screen images for the four presentation types tested

## 4 Experimental Design

Our experiments compare the four types of explanation presentation and three system types: CBR, CCBR-IG and CCBR-CF.

### 4.1 Procedure

Participants were randomly assigned to three groups, each interacting with a different system for the entire experiment. Participants completed 12 trials. The entire experiment lasted approximately 30 minutes. During each trial, the participant was provided with information from a case about a person's time in a bar and was asked, based on system group, either to fill in the features of the bar visit description (one-shot CBR) or to answer a sequence of system-selected questions (CCBR). The system then provided its prediction for whether the visitor was over or under the limit and one of the four types of explanations. The explanation type was randomly assigned using counterbalancing. Each participant saw the same explanation content, but with different explanation formats. Over the 12 trials, each participant encountered each explanation type three times. For each case, participants were asked to assess the system's decision-making and explanation along three dependent variables: goodness, convincingness, and trust. Goodness was assessed by asking "Does the explanation provide good information for assessing the system's decision making?". Convincingness was assessed by asking "Is the provided explanation convincing?" Trust was assessed by asking "Based on the provided information, would you expect to trust the system's future decisions?" The questions were answered on a 5-point Likert scale. After the 12 trials participants provided demographic data (year of birth, highest education level attained, background in STEM, and familiarity with and opinions of AI) based on questions from the literature [13, 16, 30].

### 4.2 Hypotheses

We divide our hypotheses into three types: system effects, explanation effects and interactions.

**System Effects**

- **S1**: Conversational systems will have a positive impact on the observed levels of goodness, convincingness and trust, due to the increased transparency on the system retrieval process.
- **S2**: The average scores of the two CCBR systems will differ according to the question selection strategy used.

**Explanation Effects**

- **E1**: NN+Sim:tab, NN+Sim:txt, and NN+CE designs will have more positive levels of goodness, convincingness and trust than NN because NN+CE provides more information to the user, while NN+Sim:tab and NN+Sim:txt make explicit similarity comparisons between solution and problem.

**Interactions**

- **I1**: CCBR systems using the NN+Sim:tab, NN+Sim:txt, and NN+CE designs will score better on each measure compared to the non-conversational system using the same designs due to increased system process transparency.

- **I2**: Correctness of solutions will have an effect on scores.
- **I3**: The level of similarity between the problem case and the cases used in explanation will have an impact on the scores.

We also analyzed whether perceptions changed with increasing exposure to a given explanation type. However, we generated no hypotheses for this.

### 4.3 Analysis

To assess the hypotheses stated above (excluding I2 and I3), we used three mixed model, repeated measures ANOVAs with one between-subjects factor (System Type) and two within-subjects factors (Explanation Type and Exposure) for goodness, convincingness, and trust. As discussed in Section 4.1, our study considered three system types and four explanation types. Each explanation type was seen by participants 3 times, enabling consideration of effects of exposure. A mixed model, repeated measures ANOVA was used to assess whether significant differences exist in average scores for system type, explanation type, exposure or some combination of the three. We used a pairwise comparison of means to parse any significant results and determine which values of each were particularly influential. Significance was set at 0.05.

To assess hypotheses I2 and I3, we used three mixed model, repeated measures ANCOVAs with two time-varying covariates: similarity level and incorrect responses. Each test was structured like the ANOVAs with the exception of the addition of the covariate variables. The ANCOVAs allow controlling for certain factors that may have influenced the results, and illuminate whether the listed covariates had an influence in the results obtained from the ANOVA. A pairwise comparison of means was also used on the significant results obtained by this test. The comparison provides corrected average scores for each factor value along with mean differences between factor values. Significance was set at 0.05.

## 5 Results

### 5.1 ANOVA Results

As discussed in the previous section, a mixed model, repeated measures ANOVA was run to assess whether system, explanation, exposure, or some combination of these factors had an impact on assessments of goodness, convincingness, and trust. The test showed statistically significant results for explanation type for goodness ($F = 6.937$, $p < 0.001$, $\eta^2 = 0.071$), convincingness ($F = 5.02$, $p = 0.002$, $\eta^2 = 0.055$) and trust ($F = 4.749$, $p = 0.004$, $\eta^2 = 0.052$). No other statistically significant main effects or interactions were found.

Explanation was found to be a significant factor for all three measurements, but the ANOVA results do not tell us how individual explanation types contributed to this result. To find out, we ran a pairwise comparison of means and found a statistically significant difference in the average scores between NN+Sim:tab and NN+CE for goodness ($p = 0.002$, 95% C.I. = [0.115, 0.681]),

| | NN | NN+CE | NN+Sim:tab | NN+CE | NN+Sim:txt | NN+CE |
|---|---|---|---|---|---|---|
| Goodness | | | | | | |
| CBR | 42% | 23% | 54% | 23% | 46% | 27% |
| CCBR-IG | 45% | 24% | 58% | 24% | 55% | 27% |
| CCBR-CF | 50% | 20% | 57% | 23% | 57% | 23% |
| Convincingness | | | | | | |
| CBR | 58% | 15% | 50% | 27% | 46% | 19% |
| CCBR-IG | 61% | 18% | 58% | 24% | 55% | 21% |
| CCBR-CF | 50% | 30% | 57% | 27% | 43% | 43% |
| Trust | | | | | | |
| CBR | 46% | 31% | 46% | 23% | 38% | 35% |
| CCBR-IG | 61% | 12% | 52% | 27% | 55% | 21% |
| CCBR-CF | 43% | 27% | 60% | 27% | 47% | 27% |

Table 1: For pairs of explanation types with statistically significant differences, percentages of participants who preferred each type over the other. Not shown are the percentages for "draws" between the pair (for each pair, wins and draws add up to 100%)

convincingness ($p = 0.004$, 95% C.I. = [0.088, 0.659]) and trust ($p = 0.004$, 95% C.I. = [0.073, 0.559]). NN+Sim:txt and NN+CE had statistically significant differences for goodness ($p = 0.002$, 95% C.I. = [0.106, 0.624]) and trust ($p = 0.018$, 95% C.I. = [0.030, 0.474]). NN and NN+CE for convincingness ($p = 0.026$, 95% C.I. = [0.024, 0.579]) and trust ($p = 0.032$, 95% C.I. = [0.015, 0.509]). Figure 2 shows confidence intervals for each of these significantly different pairs.

To illuminate how pairwise differences corresponded to explanation type preferences, we calculated the percentage of times that each explanation type "won" over the other in our data (Table 1). This was assessed by averaging all of the explanation type scores for one participant, grouping the scores by system type and comparing which had a higher score. Identical scores were counted as a "draw." To obtain the percentages, the number of raw win values was divided by the total number of participants for that system group. Generally, the type in each pair with a significantly greater average score also had a higher percentage of wins than their pairwise counterpart, which is to be expected. However, these wins percentages were not always above 50%, such as NN+Sim:txt when paired with CBR and measured with goodness. This may suggest that certain type, system, and measurement pairs are less effective, but as no interaction was found we draw no conclusion.

### 5.2 ANCOVA Results

When considering both similarity level and incorrect response as time-varying covariates, a mixed model, repeated measures ANCOVA showed significant results for both covariates and explanation type for convincingness (explanation: ($F = 5.981$, $p < 0.001$), similarity level: ($F = 9.586$, $p = 0.002$), incorrectness: ($F = 19.980$, $p < 0.001$)) and trust (explanation: ($F = 5.646$, $p < 0.001$), similarity

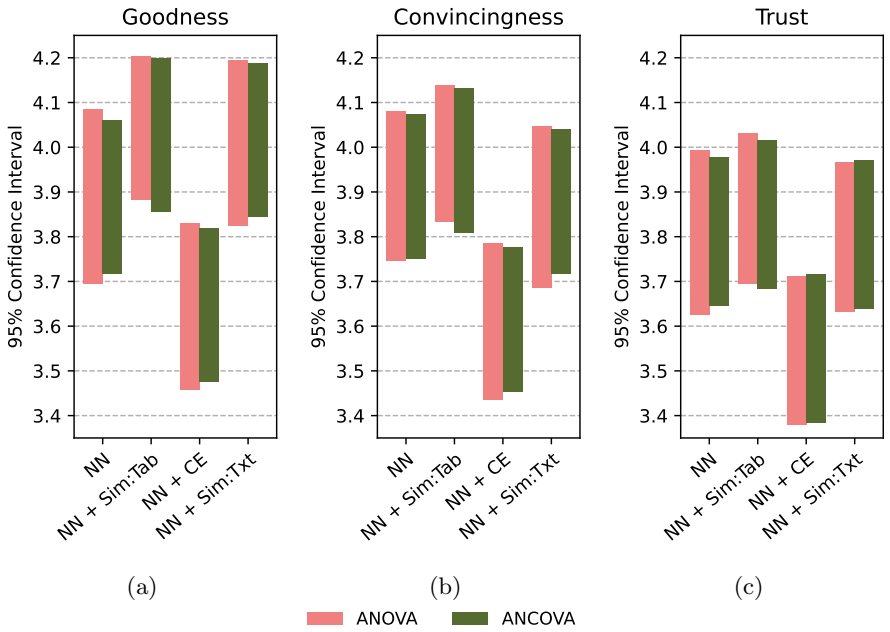



Fig. 2: 95% Confidence Intervals for Explanation Type Average Scores.

level: (F = 8.223, p = 0.004), incorrectness: (F = 10.511, p = 0.001)). Only explanation type and incorrectness were significant for goodness (explanation: (F = 9.571, p < 0.001), incorrectness: (F = 18.762, p < 0.001)).

Because incorrectness, and under certain conditions similarity level, were found to be influential in the responses participants had to the explanations. We ran a pairwise comparison of means to assess whether and what differences between explanation types existed when these factors were controlled for. We found significantly greater average scores for NN & NN+CE (goodness: (p = 0.023, 95% C.I. = [0.022, 0.471]), convincingness: (p = 0.01, 95% C.I. = [0.049, 0.547]), trust: (p = 0.012, 95% C.I. = [0.039, 0.487])), NN+Sim:tab & NN+CE (goodness: (p < 0.001, 95% C.I. = [0.159, 0.603]), convincingness: (p = 0.001, 95% C.I. = [0.106, 0.605]), trust: (p = 0.003, 95% C.I. = [0.075, 0.523])), and NN+Sim:txt & NN+CE (goodness: (p < 0.001, 95% C.I. = [0.148, 0.592]), convincingness: (p = 0.031, 95% C.I. = [0.015, 0.514]), trust: (p = 0.016, 95% C.I. = [0.031, 0.48])) for all three dependent variables.

The major difference between the ANOVA and ANCOVA pairwise comparisons is that NN & NN+CE was now significant for goodness and NN+Sim:txt & NN+CE for convincingness when they were not originally so. As with the ANOVA results, the wins data for these comparisons can help illuminate the magnitude of the differences in the data (1). The wins percentages for NN are 50% or less across all three system types, while NN+Sim:txt had wins no greater than 55%. These relatively lower percentages of wins suggest that while a dif-

ference was detected, it may not be as strong as differences that were originally detected by ANOVA.

## 6 Discussion

The ANOVA results showed that only explanation type had a significant impact on observed levels of goodness, convincingness and trust. System type and exposure, as well as an interaction between any of these factors, was not found to play a significant role in user responses. This implies that these have little impact on how a human user experiences the explanations in relation to each measure. For hypotheses S1 and S2, both concerning differences in CCBR systems compared to the other systems tested, both failed to be supported by the data. Although this result was surprising, it may be convenient for CBR practice: system developers need only apply CCBR when the domain requires it (e.g., for diagnosis), without consideration of whether to include it for explanation purposes.

Likewise, the result suggests that only the retrieved case, rather than details of the CCBR retrieval process, is likely to be important for explanation. CCBR retrieval often focuses on distinguishing the target case rapidly from other cases, which might have raised concerns that its question sequence could be unintuitive to a user who unaware of details of the process and case base contents. The primacy of case and similarity suggest that CCBR question order is unlikely to reduce goodness, convincingness, or trust assessments.

Hypothesis I1, regarding CCBR systems scoring better than CBR systems when using NN+Sim:tab, NN+Sim:txt and NN+CE, was also not supported. This again suggests that the case itself and its presentation are the primary concern for explanation, rather than how it was found.

Concerning how to present explanations to a user, no explanation type was consistently worse than all others, but pairwise comparison of means showed clear preferences between certain pairs according to the quality criterion of greatest interest, as described below.

*NN+Sim:tab Surpassed NN+CE for all Three Measures:* NN+Sim:tab had statistically greater average scores than NN+CE for all measures. This suggests that for this scenario, the most similar supporting case was most compelling, and that presentation of the closest conflicting case did not have the expected effect of increasing explanation quality by helping to delineate the applicability of the current case. Thus in this context, the NN+Sim:tab was a more appropriate explanation type than NN+CE for maximizing observed levels of goodness, convincingness, and trust. However, we note that in a domain that requires more expertise, the observed measurements for NN+CE might have differed. It is possible that in such a domain experts would make more use of the counterexample case to determine the decision boundary and consider that in their assessment, as has been hypothesized in Doyle et al. [5] and Leake et al. [21].

*NN+Sim:txt and NN Surpassed NN+CE for Some Measures:* NN+Sim:txt and NN types produced consistently statistically greater average scores than NN+CE

for goodness and trust and convincingness and trust, respectively. These results are somewhat consistent with hypothesis E1. We believed we would see NN+Sim:tab, NN+Sim:txt and NN+CE score better than NN. NN+Sim:tab and NN+Sim:txt performed better than NN+CE, which performed worse than all other presentations. No statistically significant differences existed between scores for NN+Sim:tab, NN+Sim:txt, and NN.

*Similarity Level and Incorrect Responses Influence Participant Assessments:* Based on the ANCOVA results, explanation type had an impact across the board and the similarity level of the nearest neighbor used in the explanation affected convincingness and trust. Whether the system was providing an incorrect answer affected the goodness of, convincingness of, and trust in the same explanations. These results support hypotheses I2 and I3, both stating that similarity level and incorrect solutions would have an impact on scores. These results appear suggestive of low similarity between problem and solution cases and incorrect solutions resulting in lower scores on each of the relevant measurements. However, we cannot definitively state this and leave it for further study.

Interestingly, controlling for these factors generalized the relations (i.e., made all three significantly different pairs significant for each measure) between NN & NN+CE and NN+Sim:txt & NN+CE to all three dependent variables, where originally NN & NN+CE were only significantly different when considering convincingness and trust and NN+Sim:txt & NN+CE were only significantly different when considering goodness and trust. This essentially extends the options available for choosing certain explanation types over others in certain contexts.

## 7 Ramifications: Cases are King

Explanation type alone is only part of the equation for good, convincing, and trustworthy explanations. The results show that aspects of the explanation case are important as well. Whether the system presents a decision supported by an incorrect solution case, and in certain circumstances, the level of similarity between the problem case and the retrieved case, have an impact on observed levels of goodness, convincingness, and trust.

Furthermore, the results did not suggest a significant difference in subjects' perceptions of the quality of case-based explanations when they knew how cases had been retrieved (comparing case presentation alone with both of the CCBR conditions). This suggests that, at least in this commonsense domain, the result case presented to the user is a key factor. That cases had primary importance is encouraging for the use of CBR-Neural Network hybrid systems that use learned similarity judgments (e.g., [25, 34]): Having an opaque similarity process may not decrease user perceptions that the decision of a CBR system is well explained.

Similarly, the absence of significant difference when similarity was highlighted suggests that in this commonsense domain, subjects are comfortable doing their own similarity judgments and do not need component explanation. For complex domains in which similarity may be hard to assess without support we still expect

that explanations of similarity would generally be useful, as supported by Massie, Craw and Wiratunga [24]. This remains a topic for further study. However, that explanations of similarity do not affect perceived explanation quality in this domain is consistent with early intuitions that CBR decisions are well explained simply by presenting the cases on which they are based (e.g., [20]).

## 8 Conclusions

A human subjects study was run to assess the impact of system type, explanation type, and exposure to explanation types on goodness and convincingness of and trust, when presenting explanations based on cases to the user. It was found that only explanation type played a significant role in the observed scores of each measurement, which suggests that users are less influenced by system interactions during case retrieval and that exposure to a given explanation type over the course of the study not change perceptions. Furthermore, all other explanation types were found to be preferred over explaining with an example and counterexample (NN+CE). When controlling for correct vs. incorrect solutions presented to users and the similarity between the problem and solution cases, explanation type was still significant, but similarity and incorrectness played a role in the scores obtained. This underlines the importance of case base competence and having sufficiently similar cases. It also suggests that the benefit of cases as explanations is fairly robust to potential presentation variants for the tested scenarios.

As in results by Doyle et al. [6], our study did not find benefit for presenting counterexamples. An interesting future research path would be to compare user preferences for the four explanation types for more complex domains, exploring Doyle et al.'s hypothesis that explanations involving counterexamples may be preferred in that context. Given the high proportion of participants with STEM backgrounds, it would also be interesting to examine whether these results hold among highly skeptical, low trust individuals and among those with other backgrounds. The effects of the explanation types on user models, as explored in some other work [15], would be another interesting subject for future study.

## 9 Acknowledgments

This work was funded by the US Department of Defense (Contract W52P1J2093009), and by the Department of the Navy, Office of Naval Research (Award N00014-19-1-2655).

## References

1. Aha, D., Munoz, H.: Interactive Case-Based Reasoning, vol. 14. Kluwer (2001), special issue of *Applied Intelligence*
2. Ashley, K.: Modeling legal argument: reasoning with cases and hypotheticals. MIT Press, Cambridge (1990)

3. Cunningham, P., Doyle, D., Loughrey, J.: An evaluation of the usefulness of case-based explanation. In: Case-Based Reasoning Research and Development: Proceedings of the Fifth International Conference on Case-Based Reasoning, ICCBR-03. pp. 122–130. Springer, Berlin (2003)
4. Doyle, D.: A knowledge-light mechanism for explanation in case-based reasoning. Ph.D. thesis, University of Dublin, Trinity College. Department of Computer Science (2005), available at http://www.tara.tcd.ie/handle/2262/847
5. Doyle, D., Cunningham, P., Bridge, D., Rahman, Y.: Explanation oriented retrieval. In: Funk, P., González Calero, P.A. (eds.) Advances in Case-Based Reasoning. pp. 157–168. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
6. Doyle, D., Cunningham, P., Walsh, P.: An evaluation of the usefulness of explanation in a case-based reasoning system for decision support in bronchiolitis treatment. Computational Intelligence **22**(3-4), 269–281 (2006)
7. Eargle, D., Gureckis, T., Rich, A.S., McDonnell, J., Martin, J.B.: psiTurk: An open platform for science on Amazon Mechanical Turk (Jan 2020), https://doi.org/10.5281/zenodo.3598652
8. Gates, L., Leake, D.: Evaluating CBR explanation capabilities: Survey and next steps. In: ICCBR Workshops. pp. 40–51 (2021)
9. Gunning, D., Aha, D.W.: DARPA's explainable artificial intelligence program. AI Magazine **40**(2), 44–58 (2019)
10. Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, J.B., Chan, P.: psiTurk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods **48**, 829–842 (2016)
11. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
12. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 624–635. FAccT '21, Association for Computing Machinery, NY, USA (2021)
13. Jin, W., Fan, J., Gromala, D., Pasquier, P., Hamarneh, G.: EUCA: the End-User-Centered Explainable AI framework. arXiv preprint arXiv:2102.02437 (2021)
14. Keane, M.T., Kenny, E.M.: 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: Case-Based Reasoning Research and Development: ICCBR-19. pp. 155–171. Springer, Berlin (2019)
15. Kenny, E.M., Ford, C., Quinn, M., Keane, M.T.: Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. Artificial Intelligence **294**, 103459 (2021)
16. Knapič, S., Malhi, A., Saluja, R., Främling, K.: Explainable artificial intelligence for human decision support system in the medical domain. Machine Learning and Knowledge Extraction **3**(3), 740–770 (2021)
17. Kolodner, J., Leake, D.: A tutorial introduction to case-based reasoning. In: Leake, D. (ed.) Case-Based Reasoning: Experiences, Lessons, and Future Directions, pp. 31–65. AAAI Press, Menlo Park, CA (1996)
18. Lamy, J.B., Sekar, B., Guezennec, G., Bouaud, J., Séroussi, B.: Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. Artificial intelligence in medicine **94**, 42–53 (2019)
19. Leake, D.: Goal-based explanation evaluation. Cognitive Science **15**(4), 509–545 (1991)

20. Leake, D.: CBR in context: The present and future. In: Leake, D. (ed.) Case-Based Reasoning: Experiences, Lessons, and Future Directions, pp. 3–30. AAAI Press, Menlo Park, CA (1996)
21. Leake, D., Birnbaum, L., Hammond, K., Marlow, C., Yang, H.: Integrating information resources: A case study of engineering design support. In: Proceedings of the Third International Conference on Case-Based Reasoning. pp. 482–496. Springer, Berlin (1999)
22. Leake, D., Birnbaum, L., Hammond, K., Marlow, C., Yang, H.: An integrated interface for proactive, experience-based design support. In: Proceedings of the 2001 International Conference on Intelligent User Interfaces. pp. 101–108 (2001)
23. Lee, M.K.: Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society **5**(1), 2053951718756684 (2018)
24. Massie, S., Craw, S., Wiratunga, N.: A visualisation tool to explain case-base reasoning solutions for tablet formulation. In: Proceedings of the 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer-Verlag, Berlin (2004)
25. Mathisen, B.M., Aamodt, A., Bach, K., Langseth, H.: Learning similarity measures from data. Progress in Artificial Intelligence (10 2019). https://doi.org/10.1007/s13748-019-00201-2
26. McSherry, D.: Explaining the pros and cons of conclusions in CBR. In: Proceedings of the Seventh European Conference On Case-Based Reasoning. pp. 317–330. Springer, Berlin (2004)
27. Minor, M. (ed.): Proceedings of XCBR: Case-Based Reasoning for the Explanation of Intelligent Systems, Workshop at the 26th International Conference on Case-Based Reasoning. Stockholm, Sweden (2018), URL https://iccbr18.com/wp-content/uploads/ICCBR-2018-V3.pdf
28. Mueller, S.T., Veinott, E.S., Hoffman, R.R., Klein, G., Alam, L., Mamun, T., Clancey, W.J.: Principles of explanation in human-AI systems. arXiv preprint arXiv:2102.04972 (2021)
29. Nugent, C., Cunningham, P.: A case-based recommender for black-box systems. Artificial Intelligence Review **24**(2), 163–178 (2005)
30. Sarwar, S., Dent, A., Faust, K., Richer, M., Djuric, U., Van Ommeren, R., Diamandis, P.: Physician perspectives on integration of artificial intelligence into diagnostic pathology. Digital Medicine **2**(1), 28 (2019)
31. Selbst, A., Powles, J.: Meaningful information and the right to explanation. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 48–48. PMLR (2018), https://proceedings.mlr.press/v81/selbst18a.html
32. Sormo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning—perspectives and goals. Artificial Intelligence Review **24**(2), 109–143 (2005)
33. Warren, G., Byrne, R.M.J., Keane, M.T.: Categorical and continuous features in counterfactual explanations of AI systems. In: Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI-23). p. 171–187. ACM, New York (2023)
34. Ye, X., Leake, D., Crandall: Case adaptation with neural networks: Capabilities and limitations. In: Case-Based Reasoning Research and Development. pp. 143–158. Springer, Cham (2022)