Position: The Case Against Case-Based Explanation

Jonathan Dodge

[a]Oregon State University, 1148 Kelley Engineering Center, Corvallis, OR, USA

Abstract

Case-based explanation (*CBE*) goes by many names, and in this paper I will argue why we should tend towards alternative choices when designing XAI systems. My argumentation for the stated claim rests on four broad points: (1) people seem to dislike *CBE*; (2) *CBE* relies on weak semantic linkage; (3) *CBE* is epistemically outmatched; (4) *CBE* is restrictive. This paper expounds on these arguments and concludes with thoughts about characteristics of possible alternatives.

Keywords

Explainable AI, Social Aspects of AI, Social Aspects of Explanation

Introduction

What *is* case-based explanation? Broadly speaking, **case-based explanation** is the notion of providing example(s) from the training data which are "similar" to the instance being explained, for some definition of similar. I have deployed case-based explanation [1], as have other researchers (e.g., [2]). Another name for the same strategy is **example-based explanation** (e.g., [3, 4]). For the rest of this paper, we will use *CBE* to refer to case-based explanation and aliases.

To ground our discussion, here are examples of *CBE*:

> Ex.#1: *"The training set contained 10 individuals identical to Iliana; 6 of them reoffended (60%)"* [1]

> Ex.#2: *"This decision was based on thousands of similar cases.For example, a similar case to yours is a previous customer, Claire. She was 38 years old with 18 years of driving experience, drove 850 miles per month, occasionally exceeded the speed limit, and 25% of her trips took place at night. Claire was involved in one accident in the following year."* [2]

With examples in hand, I'd like to draw a distinction between *CBE* and case-based *reasoning*, a more general process. Case-based reasoning comes very naturally to people, for example, Sarkar, et al. [5] reports participants informally describing it when asked how the system worked. Aamodt and Plaza [6] describe case-based reasoning as being based on a cycle of four processes: Retrieve, Reuse, Revise, Retain. They write:

> *"What we refer to as typical case-based methods also has another characteristic property: They are able to modify, or adapt, a retrieved solution when applied in a different problem solving context."*

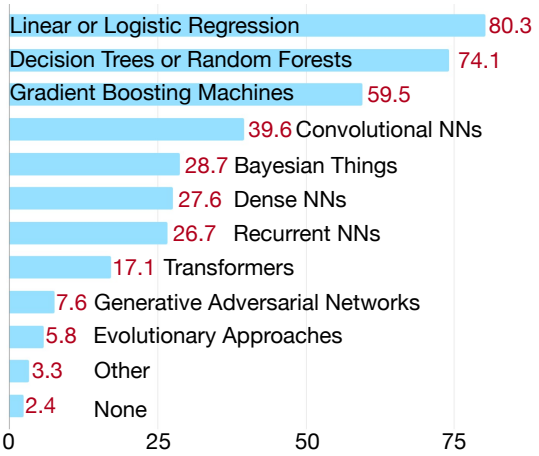

Figure 1: Image credit (redrawn): https://www.kaggle.com/kaggle-survey-2021. Models and Algorithms used in 2021 Kaggle competition submissions. Note that k-NN is not a category, so it represents at most 3.3% of submissions. (Numbers sum $\gg$ 100% because of ensemble methods)

Thus, explanations by perturbation (e.g., Sensitivity from Binns [2]) are a form of case-based reasoning. However, perturbation is not a form of *CBE*, which only consists of Retrieval and Reuse.

k-Nearest-Neighbors (k-NN, see [7], Chapter 5) is not a commonly used classifier, as Figure 1 shows, but it is one of the few circumstances when *CBE* is *sound*[1]. Kulesza et al. [8] proposed the taxonomy of **soundness** and **completeness**, reflecting whether an explanation tells "the whole truth (completeness) and nothing but the truth (soundness)." Those authors describe a possible consequence of low soundness to be "reduced trust in explanations."

Conjecture: It is a *bad idea* to deploy Case-Based Explanations when not using k-Nearest-Neighbors, or similar.

[1]Disclaimer: other criticisms raised in the paper may still apply when using *CBE* for k-NN—but at least it is sound then.

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland

dodgej@oregonstate.edu (J. Dodge)

Low soundness alone might be enough earn a "bad idea" label. One consequence of such labelling is that sometimes bad ideas should remain unused, as Correll [9] ably argues in the context of glyphs. The argumentation there can be applied to explanations, and would start with the assumption that there is a set of N explanations that are the established standards. Now, introducing the $N + 1^{th}$ with appropriate comparison to standard techniques costs increasingly large amounts of experimentation as N increases. Thus, all researchers benefit if the community occasionally weeds out "bad ideas."

Some readers may ask, *"Why do you care enough to write this paper?"* I was presenting an invited talk and during Q&A had the following exchange while answering a question on the explanation types just presented:

> Guest (me): *"...I think case-based explanation is generally a very bad idea. Part of the reason I say this is, first of all, Binns, when they researched case-based explanation, they found it to be the least preferred. We also found that [result]. And then also it has these issues you kind of saw, with that self-refuting explanation. The reason that is happening is because it is not sound. The classifier isn't a K-means clusterer or something, and so when your explanation is assuming that structure to the classifier, you are lying to people"*
> Interviewer (Ali El-Sharif): *"You just shocked me, by the way. I always thought case-based explanations were more intuitive, that they presented something logical[2]."*

Afterwards, I decided that perhaps these thoughts were worth developing further and writing about it. As negative about *CBE* as I was at the time of this exchange, the more I investigated, the more problems I found. As an example, *CBE* is one of the suggested explanation strategies found in Table 1 in the Royal Society of London's policy briefing[3]; meaning that briefers have been, and will be, instructing policymakers that *CBE* is satisfactory. I do not think it is satisfactory, so here are my four reasons for the stated conjecture; let's see if I can shock you too.

1. Users seem to dislike *CBE*

As mentioned, Binns, et al. [2] found *CBE* to be the worst of their four proposed templates for textual explanation (Demographic, Sensitivity, Input-Influence, and Case). More specifically, in those authors' words:

> *"Tukey's post-hoc paired tests showed that case-based explanations result in lower perceptions of appropriateness, fair process perception, and (in the loans case) deservedness, consistently compared to sensitivity based styles and occasionally compared to other styles. This is an effect primarily observed, like most effects in the quantitative part of our study, in the within subject study design, indicating that the act of comparison in a particular scenario is important for these differences to become apparent. Case-based explanations seem to have the most consistent negative impact on justice perceptions when presented alongside alternative explanation styles."* [2]

When we created a program to generate explanations following the same four templates, we observed the same result—*CBE* was the worst of the bunch [1]. More specifically, in those authors' words:

> *"As we found in the quantitative results, case-based explanation was judged to be the least fair—and the qualitative results provided reasons. First, some found it to provide little information about* how *the algorithm arrives at a conclusion. Second, the number of identical cases and the percentage of cases supporting the decision are often considered too small to justify the decision*—"It was unfair for the defendant because she was compared to only 22 other identical individuals... not to mention that only a little over 50% reoffended.*" (CR-61). This observation is consistent with Binns et al. [2], however, our work is based on the actual output of a ML model trained on a real dataset—allowing us to empirically show a limitation of case-based explanation[4]."* [1]

More recently, van der Waa, et al. [4] also found negative results for *CBE*, in those authors' words:

> *"Our results show that rule-based explanations have a small positive effect on system understanding, whereas both rule- and example-based explanations seem to persuade users in following the advice even when incorrect. Neither explanation improves task performance compared to no explanation. This can be explained by the fact that both explanation styles only provide details relevant for a single decision, not the underlying rationale or causality. These results show the importance of user evaluations in assessing the current assumptions and intuitions on effective explanations."* [4]

Here we see a study that found, essentially, no effect from *CBE*—other than misleading users. Their final point about the importance of user evaluation is well taken, but not widely applied.

[2] https://www.youtube.com/watch?v=bD5_q2t-4S8, timestamp ≈37:00

[3] https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

[4] *"We found that 16% of the test data exhibited the failure mode of* contradicting *the claim* ($< 50\%$ *of individuals with identical features share label)."* [1]

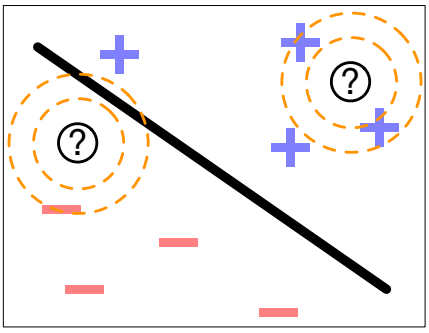

Figure 2: Notional illustration of *CBE*, expanding a circle around instances to be explained (shown with ?). Note that when far from the decision boundary, the neighborhood is more likely to contain cases of the same label. Also, note that the meaning of a distance equal to the radius of that neighborhood varies depending where the ? is in feature space; it could be the difference between changing the decision—or not. The figure shows the neighborhood as a circle, but other geometries would work.

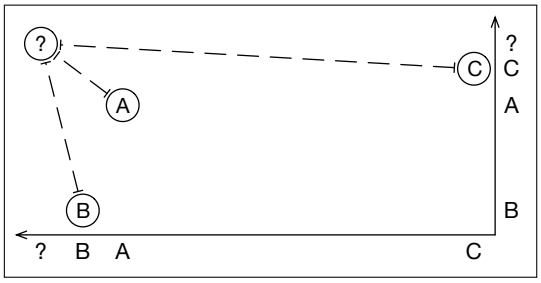

Figure 3: Notional figure intended to show the destruction of distance relationships that occur when projecting a vector space to lower dimension. In the unprojected space, ?'s nearest neighbors are: A, then B, then C. But, we will get BAC as nearest neighbors if we project onto the Y-axis (in essence removing feature X from our assumptions about the world). Similarly, we will get CAB as nearest if we project onto the X axis (by removing feature Y). While this is a contrived example, under the open-world assumption, there are *many* such dimensions being flattened by our inability to include that feature in our modelling.

As evidence that few researchers conduct user evaluations, consider that Keane and Kenny [10] surveyed **1,102** papers on "post-hoc explanation by example" before concluding:

> *"In all the papers we examined we found less than a handful (i.e., <**5**) that performed any adequate user testing of the proposal that cases improved the interpretability of models; this gap needs to be rectified."* [10]

Thus, if an XAI system designer faces a choice between *CBE* and an alternative, to the extent we have evidence at all, most of it seems to suggest that users will prefer the alternative.

2. *CBE* Relies on Weak Semantic Linkage

Under the hood, *CBE* relies on some notion of *distance*, as Sarkar et al. [11] explain:

> *"Since in order to explain the k-NN algorithm's behavior it suffices to represent* proximity*, rather than variation along any particular dimension, we sacrifice concrete interpretations of the spatial axes in favor of expressing "nearness" and "farness"."*

Figure 2 illustrates how *CBE* corresponds to expanding some neighborhood around the instance to be explained. *CBE* has different approaches on how to perform the neighborhood expansion: (1) expand a fixed size, report on distribution of contents (as in [1]); (2) expand until the neighborhood contains k items, regardless of label; (3) expand until the neighborhood contains k items, of a particular label (as in [3]).

So, here the reader asks, what's the problem with distance? Well, distance only means something *up until it doesn't*. Figure 2 illustrates that the same unit distance has drastically different semantic meaning depending where in feature space one considers the displacement. Concretely, in one case moving a distance of the neighborhood radius will *never* change the decision, while in the other case, it may.

The distance problem only gets worse if we assume the feature space is impoverished, which is typically true of decisions involving humans and many other applications which violate the **closed-world assumption**[5]. It seems probable that every explanation style will suffer when deployed under violation of this assumption. However, it also seems that *CBE* will fare particularly badly—due to reliance on distance. Suppose my current representation uses N dimensional features, and a (possibly theoretical) perfect representation uses $N + M$ dimensions. To convert between the representations can be viewed as a vector space projection. However, projections can mangle distances when not done carefully[6], as Figure 3 illustrates.

To be concrete, consider using *CBE* in a domain like chess. This might make perfect sense, because the domain is fully captured by the representation. As a result, one might expect nearby points in feature space to have

[5] https://en.wikipedia.org/wiki/Closed-world_assumption

[6] Earlier, we mentioned Sarkar, et al. [11], who explained k-NN using a 2-D projection which was *distance-preserving* by projecting onto a space defined by the distance function.

a strong semantic linkage, e.g., the *same* board has the same good actions available, a *similar* board might be a successor, predecessor, or sibling.

To contrast, consider Example #2 and implications of the closed world assumption. The explanation contains only some information about driving history, while one could imagine many other features being useful to predict insurance costs, such as risk tolerance, alcohol consumption, etc. This is not to say that such information should be obtained, merely that we should be wary of over-relying on distance when we know *in advance* that the features are incomplete. The reason is that the strength of the semantic linkage of distance *decreases* as the feature set gets further from a total representation.

3. *CBE* is Epistemically Outmatched

Due to unsoundness, *CBE* can generate self-refuting explanations—a form of epistemic *mismatch*. Footnote 4 hints at this effect, Dodge and Burnett [12] previously argued this more thoroughly (see Figure 2 in that paper). To briefly restate that argument: weak evidence occurs when the neighborhood around the instance to be explained contains mixed examples (*"I labelled this an A, 50% of nearby things are As."*); contradictions occur when it contains few or no examples of the same label (*"I labelled this an A, 10% of nearby things are As."*).

By epistemic *outmatch*, I refer to the strength of the claim far exceeding "burden of proof." An assessor consuming explanations might be trying to determine *"If it is fit for the purpose"* [13]. Here, this is a fairly strong claim, and so the legal regime might want a high burden of proof, depicted in Figure 4. Does consuming *CBE* confer "absolute certainty", akin to a mathematical proof?

CBE is definitely *not* proof. To clarify, *CBE* is essentially a form of "proof-by-example," which is a known logical fallacy[7]. Instead of as "proof-by-example", could *CBE* be considered disproof-by-*counter*example, which *is* a valid proof technique[8]? Occasionally, yes; however, many ML/AI systems do not support such formalism, as they are statistical machines. In particular, even a 99.9% accurate classifier will mishandle specific instances. Even the existence of many such examples does not disprove anything—if there are appropriately many *more* correctly handled instances.

If *CBE* isn't proof, and instead is merely *evidence*, what kind of evidence is it? It *describes* the decision, providing potentially important context; but does not justify it, as stated by van der Waa et al. [4]. Because these words have fairly broad meanings, let me clarify my intended meanings with an example, based on a tale about bank robber Willie Sutton [14], and informed by definitions found in Sørmo, et al [15]:

> Willie Sutton, a bank robber, was asked why he robbed banks, and his response was *"That's where the money is."* This **description** just offers a fact about the context of banks—notably the story does not include the explainer. An example **transparent** answer might be: *"I wanted the money in the bank."* The latter example contextualizes the action with respect to why the actor performed it, explaining how the system reached an answer [15]. Note that to **justify** that action, the explanation would need to offer a reason as to why it is *good* for Willie to want to rob banks.

Some explanations do not change based on the instance to be explained, and so are called **global** explanations. As a result of being static in this way, it seems impossible for a single global explanation to offer transparency into *every* decision. However, while this can be taken as evidence that being a local explanation is a necessary condition for transparency, it is not sufficient. In particular, *CBE* is a **local** explanation since each instance will have different neighbors.

Thus, since *CBE* merely describes the decision, offering little transparency to afford the assessor introspection on the system, it amounts to weak evidence attempting to support a strong claim. Case dismissed!

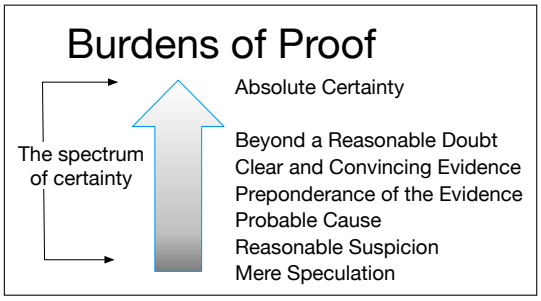

Figure 4: Image Credit: https://ua.pressbooks.pub/criminallawalaskaed/chapter/2-4-the-burden-of-proof/. The various burdens of proof arranged along the spectrum of certainty. While it is unclear what the appropriate burden of proof is for explanation tasks, we can narrow the range. For example, it seems clear that "mere speculation" is too low a standard, but "absolute certainty" is not achievable.

4. *CBE* Is Restrictive

CBE requires two properties to be true of the training data. First, the training data must still be *accessible*; second, the explainer must be allowed to *present* training data

[7] https://en.wikipedia.org/wiki/Proof_by_example
[8] https://en.wikipedia.org/wiki/Counterexample

to the user, possibly in an anonymized form. ML/AI techniques make predictions in many domains where one or the other criteria is violated, which means there are constraints for when researchers can deploy *CBE* responsibly.

Accessing training data is easiest when the training data is *tabular*. Tabular data stands in contrast with unstructured data:

> *"A very naive definition for unstructured data is anything that cannot be put up into traditional row-column or tabular database. The common examples of unstructured data are text or document based data, network or graph data, image data, video, audio, web-based logs, sensor data, etc.*[9]*"*

Concretely, when OpenAI trained GPT-3 [16], they input a vast quantity of unstructured text data—and so the training data would not be tabular. In situation like these, XAI system designers cannot always rely on the training data being cheaply accessible, if at all.

Our second criterion, *presenting* training data to assessors, is easiest when the training data is not private. For example, suppose a robot is delivering objects and that Alice is the current package recipient, but the robot botches the delivery somehow. Should Alice be able to request information about previous deliveries? How about if Alice were instead a developer? The main goal of privacy preserving machine learning (see e.g., [17]) is to prevent leakage of private information. *CBE* stands in tension with that, because even if the explanation is anonymized, it may provide enough features (e.g., *"a combination of gender, race, birth date, geographic indicator, and other descriptors"*[10]) to allow "indirect identification".

What Should We Do Instead?

Having concluded my argumentation for the conjecture, the reader may ask what they should use instead? Personally, I like the visualization strategies, (e.g., search trees, charts, saliency maps), but recognize that not everyone does. Returning to my geometric framing, there are essentially three entities available for explanation:

1. the decision boundary;
2. the instance to be explained;
3. the training data.

Some explanations use the entities directly, or characterize the relationship between them, or some mixture.

CBE relies heavily on the relationship between (2) and (3), and essentially does not use (1). In the introduction

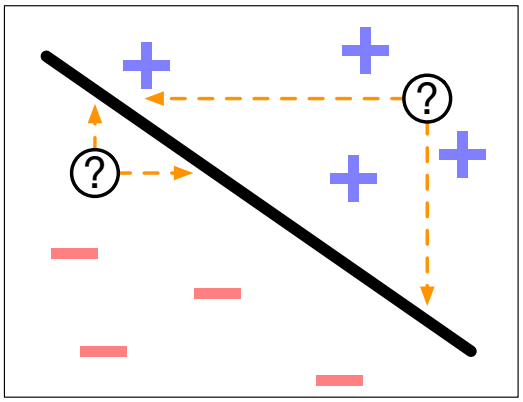

Figure 5: Notional illustration of an alternative to *CBE*: Sensitivity-based Explanation. Here, for each feature we essentially project an axis aligned line from the instance to be explained and report if it collides with the boundary.

we saw three other alternative textual explanation templates from Binns, et al. [2] which appear superior to *CBE*: Demographic, Input-Influence, and Sensitivity. Demographic explanation attempts to use so much of (3) as to insinuate the location of (1). As an example of a fragment of this style:

```
54\% of those with 0 juvenile priors did
NOT reoffend
25\% of those with >0 juvenile priors did
NOT reoffend
```

Meanwhile, Input-Influence explanations directly characterize (1), as we see in this following fragment (features positively correlated to reoffending are shown in +, negative with -, and the strength of the correlation given by the number of symbols; 0 means no effect).

```
priors count :
     0 (------)
     1 to 3 (---)
     4 to 6 (0)
     7 to 10 (+++)
     >10 (++++)
```

Last, as shown in Figure 5, Sensitivity focuses on the relationship between (1) and (2):

```
If the individual was age ‘‘30 to 39’’
they would have been predicted as
likely to reoffend
If the individual had priors count
‘‘7 to 10’’ they would have been predicted as
likely to reoffend
```

In conclusion, decision boundaries may be hard to characterize, but neglecting boundaries in explanation seems to expose consumers to possibly falling prey to logical fallacy.

[9] https://medium.com/analytics-vidhya/sql-to-nosql-4dd15ab121b0

[10] https://www.dol.gov/general/ppii

Acknowledgments

This material is based upon work supported by DARPA #N66001-17-2-4030 and joint support by NSF and USDA-NIFA under #2021-67021-35344.

References

[1] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. 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, IUI '19, ACM, New York, NY, USA, 2019, pp. 275–285. URL: http://doi.acm.org/10.1145/3301275.3302310. doi:10.1145/3301275.3302310.

[2] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, ACM, New York, NY, USA, 2018, pp. 377:1–377:14. URL: http://doi.acm.org/10.1145/3173574.3173951. doi:10.1145/3173574.3173951.

[3] 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, IUI '19, ACM, New York, NY, USA, 2019, pp. 258–262. URL: http://doi.acm.org/10.1145/3301275.3302289. doi:10.1145/3301275.3302289.

[4] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating xai: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404. URL: https://www.sciencedirect.com/science/article/pii/S0004370220301533. doi:https://doi.org/10.1016/j.artint.2020.103404.

[5] A. Sarkar, A. F. Blackwell, M. Jamnik, M. Spott, Teach and try: A simple interaction technique for exploratory data modelling by end users, in: Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on, IEEE, 2014, pp. 53–56. URL: https://www.cl.cam.ac.uk/~as2006/files/sarkar_2014_teach_try.pdfhttp://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6883022. doi:10.1109/VLHCC.2014.6883022.

[6] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI communications 7 (1994) 39–59.

[7] S. J. Russell, P. Norvig, Artificial intelligence - a modern approach, 2nd edition, in: Prentice Hall series in artificial intelligence, 2003.

[8] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W. K. Wong, Too much, too little, or just right? ways explanations impact end users' mental models, in: 2013 IEEE Symposium on Visual Languages and Human Centric Computing, 2013, pp. 3–10. doi:10.1109/VLHCC.2013.6645235.

[9] M. Correll, Ross-chernoff glyphs or: How do we kill bad ideas in visualization?, in: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI EA '18, ACM, New York, NY, USA, 2018, pp. alt05:1–alt05:10. URL: http://doi.acm.org/10.1145/3170427.3188398. doi:10.1145/3170427.3188398.

[10] 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: K. Bach, C. Marling (Eds.), Case-Based Reasoning Research and Development, Springer International Publishing, Cham, 2019, pp. 155–171.

[11] A. Sarkar, M. Jamnik, A. F. Blackwell, M. Spott, Interactive visual machine learning in spreadsheets, in: Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on, IEEE, 2015, pp. 159–163. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7357211. doi:10.1109/VLHCC.2015.7357211.

[12] J. Dodge, M. Burnett, Position: We Can Measure XAI Explanations Better with "Templates", in: IUI Workshops, 2020.

[13] B. Hambling, P. van Goethem, User acceptance testing: a step-by-step guide, BCS Learning and Development, Swindon, 2013. URL: http://cds.cern.ch/record/1619552.

[14] D. Temple, The contrast theory of why-questions, Philosophy of Science 55 (1988) 141–151. URL: http://www.jstor.org/stable/187825.

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

[16] OpenAI, Openai api faq, 2020. URL: https://openai.com/blog/openai-api/.

[17] R. Xu, N. Baracaldo, J. Joshi, Privacy-preserving machine learning: Methods, challenges and directions, CoRR abs/2108.04417 (2021). URL: https://arxiv.org/abs/2108.04417.