Considerations for Applying Logical Reasoning to Explain Neural Network Outputs

Federico Maria Cau[1], Lucio Davide Spano[1], and Nava Tintarev[2]

[1] University of Cagliari, Department of Mathematics and Computer Science,
Cagliari, Italy federicom.cau@unica.it, davide.spano@unica.it
[2] Maastricht University, DKE, Maastricht, the Netherlands
n.tintarev@maastrichtuniversity.nl

Abstract. We discuss the impact of presenting explanations to people for Artificial Intelligence (AI) decisions powered by Neural Networks, according to three types of logical reasoning (inductive, deductive, and abductive). We start from examples in the existing literature on explaining artificial neural networks. We see that abductive reasoning is (unintentionally) the most commonly used as default in user testing for comparing the quality of explanation techniques. We discuss whether this may be because this reasoning type balances the technical challenges of generating the explanations, and the effectiveness of the explanations. Also, by illustrating how the original (abductive) explanation can be converted into the remaining two reasoning types we are able to identify considerations needed to support these kinds of transformations.

Keywords: Explainable User Interfaces · XAI · Reasoning.

1 Introduction

In the last decade, eXplainable AI (XAI) research has made great advances, introducing new explanation techniques like Grad-CAM [18], SHAP [13], LRP [1], LIME [17], LORE [10], RETAIN [6], DeepLIFT [19], a.o. However, much of this previous work does not consider which kind of logical reasoning is presented to users, or how this interacts with characteristics of task, much less individual differences between users. Given the possible implication of the reasoning type in the effectiveness of the XAI system, we inspect the transformation from one kind of reasoning to another. We do this intending to draw possible considerations for selecting between the types of reasoning, starting with neural networks models. While the categorization of the reasoning types applies to explanations for other probabilistic models, the available explanation techniques differ.

The paper is organised as follows: in Section 2 we provide some background on XAI taxonomies, evaluation tasks and reasoning types, in Section 3 we investigate the reasoning types, in Section 4 we establish the guidelines for reasoning transformation, in Section 5 we conclude the paper and we discuss possible ideas for future work.

2 Related Work

In this section, we explore three topics: the first concerns the existing XAI taxonomies, to catalogue existing state-of-the-art techniques that explain neural networks. After that, we will talk about the task types with which explanations techniques are proposed and their importance for evaluation by humans. Finally, we define the reasoning types and their relationship with XAI explanations.

2.1 Explanation Taxonomies

There have been articles that have categorized explanation methods for neural networks. Among them, [24, 3, 12, 20] were very useful to lay the foundations of our research: they describe a comprehensive taxonomy of interpretability methods regarding Deep Neural Networks (DNN), including goals, properties and architecture, together with guiding principles for their safety and trustworthiness. Furthermore, other surveys go beyond the analysis of neural network models, and help us to expand the knowledge on methods of explanation and models used [2, 11]. However, there are two surveys that also focus on the impact of explanations on users [14, 22]. The former supplies a categorization between design goals for interpretable algorithms considering different XAI user groups. The latter introduces a conceptual framework that explains how human reasoning processes informs XAI techniques, which we will deepen in the next sections.

2.2 Types of Tasks

In addition to determining which XAI which techniques to use, another key step is to identify what type of task the user will accomplish. We started studying tasks types from articles [7] and [4], following the distinction present in the latter and taking into account two types of tasks, proxy and real. In studies that use proxy tasks, the user mainly evaluates how well he perceives the AI's explanations and what it has learned, focusing on the AI and on the actual goals the users have in interacting with the system. [16, 25, 21]. Conversely, studies that use real tasks evaluate the cooperation between users and AI: the user has a primary role regarding the decision to make and can decide or not to use the AI advice to complete the task [8, 4, 23]. The paper in [4] also criticises the current evaluation methodology of XAI based on proxy tasks, demonstrating that their conclusions may not reflect the usage of the system on real tasks. Given this discovery, we consider real tasks in the transformations explained in Section 4.

2.3 Types of Reasoning

During the evaluation phase, where the interaction between user and AI takes place, one fundamental factor comes into play: the reasoning type. We started analyzing this subject from the article [22], cited above. The authors highlight that the AI's role is to facilitate the user connection with its decisions, starting from the reasoning expressed through the AI's explanations. Accordingly, a reasonable choice is to deeply explore a subset of the reasoning types, for instance, the logical ones: inductive, deductive and abductive. Here, we consider Peirce's syllogistic theory [9], and note that we can translate between these by exchanging the conclusion (or result), the major premise (the rule) and the minor premise (the cause). We will investigate these reasoning types in the next section.

3 Investigating the Reasoning Types in Explainable Intelligent Interfaces

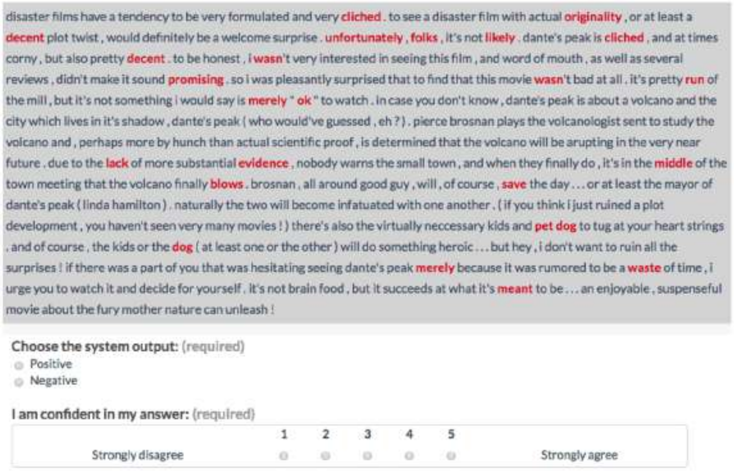
In this section, we investigate the reasoning types previously mentioned borrowing some examples from the literature. Article [4] briefly discusses inductive and deductive reasoning, explaining how to integrate them in a user evaluation context but without an in-depth exploration. Furthermore, abductive reasoning is often (unintentionally) used to compare novel explanation techniques with state-of-the-art techniques where the user role is to identify what is the best-generated explanation during the evaluation. Before starting with the definitions, we introduce the three components identified in all types of reasoning (c.f., Peirce's syllogistic theory [9]): one (or more) Cause (or Case/ Explanation/ Reason), Effect (or Observation/ Result) and Rule (or Generalization/ Theory).

Table 1. Illustrative articles that explain neural network models, divided according to the type of reasoning, network and task.

Year	Article	Reasoning of task	Type/s of network	Type/s of task
2018	[15]	deductive	MLP	proxy
2019	[5]	inductive	RNN	real
2017	[21]	abductive	LSTM, CNN	proxy

When applying this theory to XAI interfaces, it is important to identify whether their representation is *implicit or explicit.* For example, a rule or a cause is implicit when it comes from the user's mental model and not from the AI. Instead, it is explicit when we consider the AI's mental model, i.e. what it has learned in the training process and its explanations on the output's prediction. The reference paper for the concepts we are going to describe is that of [9] (see Table 1). The order of the components described is unimportant, except for the last one: we will use the latter to highlight what is the reasoning to elicit from the user.

Deduction: given a cause and a rule, deduce an effect. This type of reasoning starts with general rules and examines the possibilities to reach a specific, logical conclusion. Deductive reasoning is alternatively referred to as "top-down" logic because it usually starts with a general statement and ends with a narrower, specific conclusion. The article [15] contains an example of this reasoning, as depicted in Figure 1. Cause : The AI's words in red that identify a negative or positive sentiment. Rule : Certain words contribute to the sentiment of text (implicit). Effect : The sentiment prediction.



disaster films have a tendency to be very formulated and very cliched . to see a disaster film with actual originality , or at least a decent plot twist , would definitely be a welcome surprise . unfortunately , folks , it's not likely . dante's peak is cliched , and at times corny , but also pretty decent . to be honest , i wasn't very interested in seeing this film , and word of mouth , as well as several reviews , didn't make it sound promising . so i was pleasantly surprised that to find that this movie wasn't bad at all . it's pretty run of the mill , but it's not something i would say is merely " ok " to watch . in case you don't know , dante's peak is about a volcano and the city which lives in it's shadow , dante's peak (who would've guessed , eh ?) . pierce brosnan plays the volcanologist sent to study the volcano and , perhaps more by hunch than actual scientific proof , is determined that the volcano will be arupting in the very near future . due to the lack of more substantial evidence , nobody warns the small town , and when they finally do , it's in the middle of the town meeting that the volcano finally blows . brosnan , all around good guy , will , of course , save the day . . . or at least the mayor of dante's peak (linda hamilton) . naturally the two will become infatuated with one another . (if you think i just ruined a plot development , you haven't seen very many movies !) there's also the virtually neccessary kids and pet dog to tug at your heart strings . and of course , the kids or the dog (at least one or the other) will do something heroic . . . but hey , i don't want to ruin all the surprises ! if there was a part of you that was hesitating seeing dante's peak merely because it was rumored to be a waste of time , i urge you to watch it and decide for yourself . it's not brain food , but it succeeds at what it's meant to be . . . an enjoyable , suspenseful movie about the fury mother nature can unleash !

Choose the system output: (required)

○ Positive
○ Negative

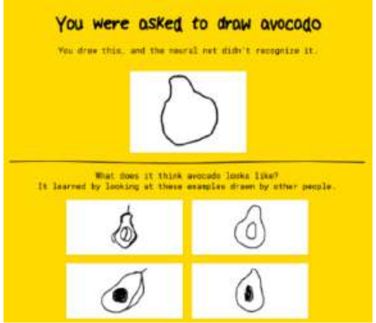
I am confident in my answer: (required)

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Fig. 1. The AI's words in red that identify a negative or positive sentiment are the Cause ; the Rule is implicit (related to the mental model of the user regarding his semantic knowledge and not given by the AI); the user choice about the sentiment prediction (whether positive or negative), is the Effect [15].

Induction: given a cause and an effect, induce a rule. This type of reasoning involves drawing a general conclusion from a set of specific observations. It is alternatively referred to as "bottom-up" logic because it involves widening specific premises out into broader generalizations. Article [5] is an example of inductive reasoning, shown in Figure 2. Cause : The AI's example-based explanations. Effect : The AI was unable to recognize the user's sketch. Rule : Certain properties of sketches represent an object (implicit).

Abduction: given an effect and a rule, abduct a cause. This type of reasoning typically begins with an incomplete set of observation/s and proceeds to the likeliest possible explanation. An example of abductive reasoning is [21], as depicted in Figure 3. Effect : is given by the AI's sentiment prediction. Rule : The chart in the explanation boxes give the user the intuition of the weights the AI uses for computing the valence of the sentence (implicit). Cause : The user selects the weights he considers the best (proxy task).

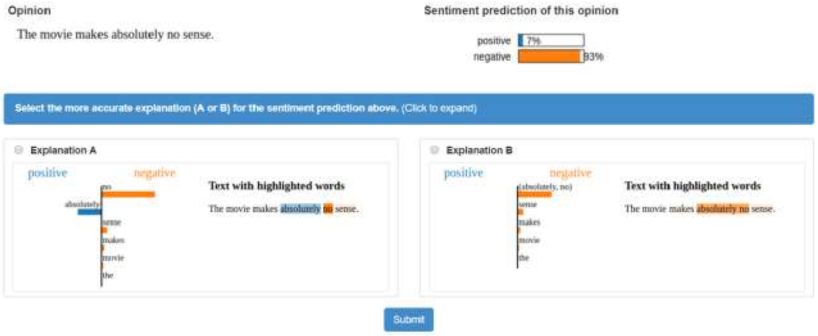


You were asked to draw avocado

You drew this, and the neural net didn't recognize it.

What does it think avocado looks like?
It learned by looking at these examples drawn by other people.

Fig. 2. The example-based explanations shown by the AI identify the Cause ; the Effect is that the AI did not recognize the sketch the user draw; so, the user needs to understand the Rule from the AI's examples. [5].



Opinion

The movie makes absolutely no sense.

Sentiment prediction of this opinion

positive 7%
negative 93%

Select the more accurate explanation (A or B) for the sentiment prediction above. (Click to expand)

○ Explanation A

Text with highlighted words

The movie makes absolutely no sense.

○ Explanation B

Text with highlighted words

The movie makes absolutely no sense.

Fig. 3. The AI's sentiment prediction represents the Effect ; the AI's chart and the sentiment highlight that identify a negative or positive sentiment represents the Rule ; so the user has to find the best Cause , like: "I choose explanation B because it highlights better the sentiment meaning" [21].

4 Transforming Explanations to different Reasoning Types

The goal of transforming the reasoning task is to analyze possible preferences or performance differences on the user during the evaluation phase. Moreover, obtaining all three reasoning types allows us to find underrepresented reasoning types in the literature and to study if they work better with users than the original task's reasoning. Abductive reasoning explanations (Fig. 3) are quite easy to generate and have a good understandability power: this compromise could be the reason why they are the most used reasoning for comparing the quality of explanation techniques. As for inductive explanations (Fig. 2), we can easily generate examples based on data, but the understandability is bounded by the selected examples. Deductive explanations (Fig. 1) are more challenging to generate when we have to create explicit rules, but they are very understandable to the user. As mentioned in Section 2, the resulting transformation's task will be a real one, for avoiding the mistake highlighted by article [4]. This can be achieved by revisiting the task's question from the AI to the user perspective.

Now, let us depict some ideas to formulate the transformation with the three reasoning types described previously, considering that we often have an explicit or implicit rule or cause, and nearly always an effect given by the AI (suggestion).

Deductive to Inductive and Abductive. To adapt to inductive reasoning, we need to replace the cause giving similar or dissimilar examples concerning the data present in the task and based on the output of the AI, thus generating example-based explanations. So, the users grasp the Rule (that becomes implicit) that brought the AI to that result, and draw their conclusion Effect about the given task data. To switch to abductive reasoning, the AI should provide a Cause based on the task data. After that, we need to make the Rule *implicit* to not confuse the reasoning with the deductive one.

Inductive to Deductive and Abductive. To switch to the deductive case, the AI needs to *explicitly* define a Rule, that also includes a Cause. We can accomplish this by leveraging the properties of the AI model or adding a complementary one to obtain a rule. Additionally, sometimes the user may take the decision without obtaining the Effect explicitly from the AI, but letting the user deduce it from the rules and causes. Passing instead from inductive to abductive reasoning, we use the common traits in the inductive examples to create a Cause.

Abductive to Inductive and Deductive. Starting from this reasoning type, we hypothesize to already have an Effect given by the AI. To translate to the inductive case, we need to replace the Cause given by the task's data with that of example-based explanations and transform the Rule to *implicit*. To move to deductive reasoning, we need to *explicitly* define the AI's Rule, that may change the original Cause, and if we want, hide the Effect.

5 Conclusion and Future Work

In sum, we investigated the considerations that arise when transferring between different types of logical reasoning, considering real tasks as the resulting transformation's tasks. We identified the importance of differentiating between implicit and explicit rule representation. We also consider whether the choice of reasoning type balances the technical challenges of generating the explanations, and the effectiveness of the explanations for humans. As future work, we plan to validate these ideas in user evaluations for different reasoning types. Also, we plan to create a taxonomy considering reasoning and task types, in addition to other useful metrics related to the XAI explanation, and further explore logical reasoning on other black-box models beyond neural networks.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (07 2015). https://doi.org/10.1371/journal.pone.0130140, https://doi.org/10.1371/journal.pone.0130140
2. Barredo Arrieta, A., Diaz Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado González, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, V.R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion (12 2019). https://doi.org/10.1016/j.inffus.2019.12.012
3. Buhrmester, V., Münch, D., Arens, M.: Analysis of explainers of black box deep neural networks for computer vision: A survey. ArXiv **abs/1911.12116** (2019)
4. Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L.: Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. Proceedings of the 25th International Conference on Intelligent User Interfaces (Mar 2020). https://doi.org/10.1145/3377325.3377498, http://dx.doi.org/10.1145/3377325.3377498
5. Cai, C.J., Jongejan, J., Holbrook, J.: The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. p. 258–262. IUI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3301275.3302289, https://doi.org/10.1145/3301275.3302289
6. Choi, E., Bahadori, T., Schuetz, A., Stewart, W., Sun, J.: Retain: Interpretable predictive model in healthcare using reverse time attention mechanism (08 2016)
7. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
8. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2015). https://doi.org/10.1109/cvpr.2015.7298754, http://dx.doi.org/10.1109/CVPR.2015.7298754
9. Flach, P., Kakas, A.: Abductive and inductive reasoning: Background and issues (01 2000). https://doi.org/10.1007/978-94-017-0606-3-1
10. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. ArXiv **abs/1805.10820** (2018)
11. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys **51**(5), 1–42 (Jan 2019). https://doi.org/10.1145/3236009, http://dx.doi.org/10.1145/3236009
12. Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., Yi, X.: A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Review **37**, 100270 (Aug 2020). https://doi.org/10.1016/j.cosrev.2020.100270, http://dx.doi.org/10.1016/j.cosrev.2020.100270
13. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (12 2017)

14. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems (2018)
15. Nguyen, D.: Comparing automatic and human evaluation of local explanations for text classification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1069–1078. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1097, https://www.aclweb.org/anthology/N18-1097
16. Rajani, N.F., Mooney, R.J.: Ensembling visual explanations for vqa. In: Proceedings of the NIPS 2017 workshop on Visually-Grounded Interaction and Language (ViGIL) (December 2017), http://www.cs.utexas.edu/users/ai-labpub-view.php?PubID=127684
17. Ribeiro, M., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. pp. 97–101 (02 2016). https://doi.org/10.18653/v1/N16-3020
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (Oct 2019). https://doi.org/10.1007/s11263-019-01228-7, http://dx.doi.org/10.1007/s11263-019-01228-7
19. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences (04 2017)
20. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Towards medical xai (2019)
21. Tsang, M., Sun, Y., Ren, D., Liu, Y.: Can i trust you more? model-agnostic hierarchical explanations (2018)
22. Wang, D., Yang, Q., Abdul, A., Lim, B.: Designing theory-driven user-centric explainable ai (05 2019). https://doi.org/10.1145/3290605.3300831
23. Yin, M., Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. pp. 1–12 (04 2019). https://doi.org/10.1145/3290605.3300509
24. Yu, R., Shi, L.: A user-based taxonomy for deep learning visualization. Visual Informatics **2** (09 2018). https://doi.org/10.1016/j.visinf.2018.09.001
25. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)