Paving the way towards counterfactual generation in argumentative conversational agents

Ilia Stepin, Alejandro Catalá, Jose M. Alonso
Centro Singular de Investigación en
Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago
de Compostela, Spain
{ilia.stepin,alejandro.catala,
josemaria.alonso.moral}@usc.es

Martín Pereira-Fariña
Departamento de Filosofía
e Antropoloxía,
Universidade de Santiago
de Compostela, Spain
martin.pereira@usc.es

Abstract

Counterfactual explanations present an effective way to interpret predictions of black-box machine learning algorithms. Whereas there is a significant body of research on counterfactual reasoning in philosophy and theoretical computer science, little attention has been paid to counterfactuals in regard to their explanatory capacity. In this paper, we review methods of argumentation theory and natural language generation that counterfactual explanation generation could benefit from most and discuss prospective directions for further research on counterfactual generation in explainable Artificial Intelligence.

1 Introduction

Automatic decision-making systems using black-box machine learning (ML) algorithms are now widely used in various complex domains from legislation (Greenleaf et al., 2018) to health care (Gargeya and Leng, 2017). However, such systems cannot be trusted blindly as their output often comes unexplained to end users (Rudin, 2018). As a result, there exists a lack of confidence in such automatic decisions caused by a low degree of their interpretability (Ribeiro et al., 2016).

The need for intelligent systems to explain their decisions has driven a decent amount of research in the past decades (Biran and Cotton, 2017). However, advances in social sciences impose novel challenges on explainable agents. For example, recent findings from cognitive science testify that the key feature of explanations is their contrastiveness (Miller, 2019), that is the ability to reflect on alternative scenarios of actually happened events. Whereas little research has been performed on generation of such counterfactual explanations, we believe that enabling virtual assistants and recommendation systems with the ability to generate them should increase greatly their acceptance among end users.

In this paper, we briefly review prospective methods for addressing the problem of counterfactual explanation generation. Subsequently, we aim to further shape the line of research devoted to counterfactual analysis for explainable Artificial Intelligence (AI) by pointing to the existing field-specific theoretical foundations and potential directions of its algorithmic design. As a result, this work supports a discussion on prospective methods for argumentative conversational agent development.

The rest of the manuscript is organised as follows. Section 2 inspects definitions of a counterfactual explanation and reviews existing generation approaches to counterfactual explanations. Section 3 describes the most prominent formal argumentation frameworks as a theoretical basis for counterfactual analysis. Section 4 discusses the classification of and recent advances in developing argumentative conversational agents in the context of counterfactual generator implementation. Finally, we conclude with outlining open challenges relevant for counterfactual explanation generation in section 5.

2 Counterfactual explanations

Explanations are argued to be contrastive (Miller, 2019). According to Miller, people are not satisfied with mere direct explanations in form of causal relations between the antecedent and consequent but also require to know why an alternative (or opposing) event could not have happened. Furthermore, Pearl and Mackenzie (2018) argue that it is the ability to produce such contrastive statements, referred to as *counterfactuals*, that lies on top of human reasoning.

In ML, a counterfactual explanation describes

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 20–25,
Tokyo, Japan, October 29 2019.

an alternative (hypothesised) situation which is as similar as possible to the original event in terms of its feature values while having a different outcome prediction ("the closest possible world") (Molnar, 2019). When searching for a suitable counterfactual explanation, the distance between a given piece of factual information and its counterpart is to be minimised while the outcome is different so that the counterfactual presumes only the most relevant alterations to the original fact. In addition, counterfactuals capture contextual information as they describe "a dependency on the external facts that led to a decision" (Wachter et al., 2018). As a result, explanations supported by counterfactuals are likely to gain acceptability by end users.

While the general understanding of the concept of counterfactuals is shared among researchers, there exist several interpretations of this phenomenon. As counterfactuals are generally assumed to have a clear connection with causation (Pearl and Mackenzie, 2018), they are often viewed as non-observable potential outcomes that would have happened in the absence of the cause (Shadish et al., 2002). In terms of causality, they are informally defined as conditional statements in the form: "If event X had not occurred, event Y would not have occurred" (Lewis, 1973). However, Wachter et al. (2018) propose a causation-free definition of an unconditional counterfactual statement based on the idea of subject's disbelief in a given hypothetical situation. On the other hand, counterfactuals are also sometimes referred to as "conditional connectives" in conditional logic (Besnard et al., 2013).

In recent years, there have been several attempts to approach the problem of counterfactual explanation generation. Wachter et al. (2018) suggested an approach for calculating counterfactuals based on the use of the Manhattan distance. Sokol and Flach (2018) adopted this approach to implement a counterfactual explanation generator for a decision tree-based AI system. In addition, Hendricks et al. (2018) proposed a model where candidate counterfactual pieces of evidence are selected from a set of all the noun-phrases of the corresponding textual descriptions of input images. Such evidence is then verified to be absent in the original image so that it can be used in the output counterfactual explanation. A rule-based system is then used to generate fluent negated explanations. Later, Birch et al. (2019) introduced an arbitrated dispute tree model arguing that the explanations generated by their model are indeed contrastive in accordance with the principles proposed by Miller (2019) as opposite outcomes are presented for all cases. Furthermore, the corresponding features and stages are explicitly found for cases opposing to the focus case.

As has been shown above, the problem of counterfactual explanation generation is concerned with several topics from philosophy, (computational) linguistics, and AI. While this leaves room for developing novel synergistic methods and algorithms that would combine insights from all the relevant fields, potential challenges when developing such tools are multiplied. For example, the fact that certain types of counterfactual explanations are preferred over their counterparts (Byrne, 2019) places further restrictions on newly developed frameworks as in designing heuristics for reducing the search space of the most relevant counterfactual explanations in accordance with such additional restrictive criteria.

In conclusion, counterfactual explanations are likely to enrich conversational interfaces of any system to be considered explainable. However, counterfactuals produced directly from ML algorithm predictions show a lack of coherence and appear unreliable from the ethical point of view (Kusner et al., 2017). Moreover, they usually do not involve a user in an extensive dialogic interaction, which makes them self-explanatory only in a limited number of cases. Therefore, we hypothesise that going deeper with their formalization is likely to overcome these weaknesses.

3 Formal argumentation

Formal argumentation (Baroni et al., 2018) provides practitioners with a natural form of counterfactual explanation formalization. Indeed, argumentation is claimed to mimic human reasoning (Cerutti et al., 2014). As such, it offers a set of tools that have become widely applicable to interpreting the output of ML algorithms. Formal argumentation embraces a wide range of theoretical frameworks from *argumentation schemes* (Walton et al., 2008) to *dialogue games* (Carlson, 1985), among others. In this paper, we focus on *abstract argumentation* (AA) frameworks as a prospective theoretical basis for counterfactual explanation generation.

While disregarding the internal structure of ar-

guments, AA frameworks primarily deal with relations between arguments. The AA framework introduced in Dung (1995) is a pioneering theoretical framework, which has become well known. This AA framework is a directed graph (also referred to as "argument graph") formally defined as a pair $AA = (A, R)$ where A is a set of arguments, $R \subseteq A \times A$ being a set of binary *attack* relations between pairs of arguments $(a, b) \in R$. In these settings, argument a is assumed to attack argument b. The acceptability of arguments is defined through numerous semantics in form of extensions over a conflict-free set of arguments, which is defined as a subset of all arguments that do not attack each other.

Due to its seeming simplicity, Dung's framework only presents the very basic argumentative constructs. Indeed, a number of extensions address this handicap. For example, some models attempt to extend the original Dung's argumentation framework by refining the concept of attacks between arguments allowing attack-to-attack relations (Modgil, 2007; Baroni et al., 2011). In contrast, a significant body of research aims to complement the nature of relations between arguments by incorporating supportive relations (Verheij, 2002; Amgoud et al., 2008).

It is worth noting that variants of AA have already been employed to address the problem of explanation generation. For example, Amgoud and Serrurier (2008) use the AA framework to resolve a binary classification task and motivate the outcome with arguments constructed, subsequently compared against each other, and ranked according to their strength. Šešelja and Straßer (2013) augment AA with explanatory features for scientific debate modelling. However, none of these works embodies counterfactual explanations.

Dung et al. (2009) proposed a conceptually novel instance of the AA framework which is known as the assumption-based argumentation (ABA) framework. Thus, ABA operates on a set of assumptions deducted via inference rules and reconsiders attack relations defined now as contraries to assumptions supporting the original argument. Following this approach, Zhong et al. (2019) implements an ABA multi-attribute explainable decision model that generates textual explanations on the basis of dispute trees. Notice that this model is claimed to be an argumentation-based framework to generate textual explanations for decision-making models. Nevertheless, while justifying why a particular decision is preferred over its counterpart, the model does not offer counterfactual explanations for rejected decisions.

Despite a rising interest towards counterfactual explanation generation in recent years, little work has been done in the direction of applying formal methods (including argumentation) to generation of counterfactual explanations. While most existing counterfactual frameworks make use of elements of causal inference, we find counterfactual statements naturally integrated into conditional logic-based (Besnard et al., 2013) as well as abstract argumentation (Sakama, 2014) frameworks. However, none of these frameworks governs any existing counterfactual explanation generation system so far.

4 Argumentative conversational agents

Argumentative frameworks can be embedded directly into chatbots or conversational agents to interact with end users. In terms of practical implementation, conversational agents are broadly divided into two main groups: retrieval-based and generative agents (Chen et al., 2017). On the one hand, a retrieval-based agent aims to select the most suitable response from the set of predefined responses that it contains given user's inquiry (Rakshit et al., 2017; Bartl and Spanakis, 2017). This kind of agents is based on the use of templates and produces grammatical utterances in all cases. However, such template-based text generators are expensive to develop and maintain due to immense expert labour resources required. On the other hand, generative models can form previously unseen utterances as they are trained from scratch without any templates in store (Li et al., 2016; Shao et al., 2017). Nevertheless, their generic responses limit their applicability to explainable AI problems.

The need for explanability of complex ML-based systems imposes additional requirements on conversational agents. Thus, automatically generated explanations are expected to be convincing enough in order to increase user's confidence in system's predictions with respect to the given task. This is hypothesised to lead to an indispensable shift of attention towards development of argumentative conversational agents (or argumentative dialogue systems) operating on a set of arguments as responses to user's inquiries. Further-

more, such argumentation-based agents are considered to push the boundaries of the present-day conversational agents towards more human-like interaction (Dignum and Bex, 2018). In combination with recent advances in deep learning and reinforcement learning, the use of argumentation as a theoretical basis for conversational agents opens prospects for a new era of generative conversational agents (Rosenfeld and Kraus, 2016; Rach et al., 2019).

Finally, the issue of evaluation of argumentation-based conversational agents merges with those coming directly from the field of natural language generation (NLG) and explainable AI. At present, there is no unifying agreement on a set of evaluation metrics to be used neither within the NLG community (Gatt and Krahmer, 2018) nor within the explainable AI community (Adadi and Berrada, 2018). While common objective (automatic) and subjective (human-oriented surveys) metrics used for NLG evaluation are found in the literature on conversational agents and dialogue systems, novel metrics are regularly introduced for instances of argumentative chatbots (e.g., distinctiveness, as in (Le et al., 2018)) and counterfactual generators (e.g., accuracy with counterfactual text and phrase-error, as in (Hendricks et al., 2018)). Thus, a direct comparison between analogous agents becomes a particularly challenging task. As a possible solution, a combination of subjective and objective metrics is believed to be a reasonable starting point for a discussion on the choice of evaluation techniques. At the same time, automatically generated explanations are expected to be accurate, consistent, and comprehensible. As the perception of these properties is highly subjective, they cannot be measured (and therefore evaluated) directly and require further investigation.

5 Concluding remarks

Our literature review has revised the foundations of current approaches to counterfactual explanation generation. The limitations found call for some potential areas for improvement on the development of explainable AI systems.

First, there is no single definition of a counterfactual explanation. While counterfactuals have various interpretations in the literature, we find it particularly important to suggest a uniform definition that would not only capture all the properties of counterfactual explanations but also allow for designing a universal domain-independent framework for their generation.

Second, existing argumentation-based explanation generation models do not fully solve the problem of counterfactual explanation generation. While some of such models do not offer consistent explanations in textual form, others do not output contrastive explanations. Therefore, a more holistic counterfactual generation framework should be developed to close this gap.

Third, formal argumentation is rarely considered in present-day conversational agents. To the best of our knowledge, such argumentation-based agents do not consider incoming dialogic information received from the direct interaction with the user to contextualise their counterfactual explanations. However, processing such information may help to improve the quality of the offered counterfactual explanations making them more personalised. Therefore, capturing such contextual information presents another noteworthy line of research.

The aforementioned issues, along with others not discussed due to space limitations, show that the generation of counterfactual explanations is a timely but complex problem. In the future, we plan to address these issues by designing an argumentation-based dialogue protocol and developing a conversational agent ready to make use of the protocol to output accurate and consistent counterfactual explanations.

Acknowledgments

Jose M. Alonso is *Ramón y Cajal* Researcher (RYC-2016-19802). This research was also funded by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29 and "accreditation 2016-2019, ED431G/08"). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable arti-

ficial intelligence (XAI) . *IEEE Access*, 6:52138–52160.

Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and Pierre Livet. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093.

Leila Amgoud and Mathieu Serrurier. 2008. Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209.

P. Baroni, D. Gabbay, and M. Giacomin. 2018. *Handbook of Formal Argumentation*. College Publications.

Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Giovanni Guida. 2011. AFRA : Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 52(1):19–37.

A. Bartl and G. Spanakis. 2017. A retrieval-based dialogue system utilizing utterance and context embeddings. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1120–1125, United States.

Philippe Besnard, Éric Grégoire, and Badran Raddaoui. 2013. A conditional logic-based argumentation framework. In *Scalable Uncertainty Management*, pages 44–56, Berlin, Heidelberg. Springer Berlin Heidelberg.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence (XAI)*, pages 8–13.

David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. 2019. Explanations by arbitrated argumentative dispute. *Expert Systems With Applications*, 127:141–156.

Ruth M. J. Byrne. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282.

Laura Carlson. 1985. *Dialogue Games: An Approach to Discourse Analysis*. Reidel.

Federico Cerutti, Nava Tintarev, and Nir Oren. 2014. Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, ECAI'14, pages 207–212. IOS Press.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19:25–35.

Frank Dignum and Floris Bex. 2018. Creating dialogues using argumentation and social practices. In *Internet Science*, pages 223–235. Springer International Publishing.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. 2009. *Assumption-Based Argumentation*, pages 199–218. Springer US, Boston, MA.

Rishab Gargeya and Theodore Leng. 2017. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124:962–969.

Albert Gatt and E.J. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170.

Graham Greenleaf, Andrew Mowbray, and Philip Chung. 2018. Building sustainable free legal advisory systems: Experiences from the history of AI & law. *Computer Law & Security Review*, 34(2):314–326.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. In *Proceedings of the Workshop on Human Interpretability in Machine Learning (WHI)*, pages 95–98.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 4069–4079.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, Brussels, Belgium. Association for Computational Linguistics.

David K. Lewis. 1973. *Counterfactuals*. Blackwell.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Sanjay Modgil. 2007. An abstract theory of argumentation that accommodates defeasible reasoning about preferences. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 648–659. Springer Berlin Heidelberg.

Christoph Molnar. 2019. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*, 1st edition. Basic Books, Inc., New York, NY, USA.

N. Rach, K. Weber, A. Aicher, F. Lingenfelser, E. André, and W. Minker. 2019. Emotion recognition based preference modelling in argumentative dialogue systems. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 838–843.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems (IWSDS)*, pages 45–52.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, NY, USA.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, volume 285, pages 320–328.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. In *Proceedings of the Workshop on Critiquing and Correcting Trends in Machine Learning – 32nd Conference on Neural Information Processing Systems (NIPS)*.

Chiaki Sakama. 2014. Counterfactual reasoning in argumentation frameworks. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA)*, pages 385–396.

Dunja Šešelja and Christian Straßer. 2013. Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190(12):2195–2217.

W. R. Shadish, T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2210–2219.

Kacper Sokol and Peter A. Flach. 2018. Conversational explanations of machine learning predictions through class-contrastive counterfactual statements. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5785–5786.

Bart Verheij. 2002. On the existence and multiplicity of extensions in dialectical argumentation. In *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning (NMR)*, pages 416–425.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law Technology*, 31:841–887.

D. Walton, P.C. Reed, C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. 2019. An explainable multi-attribute decision model based on argumentation. *Expert Systems With Applications*, 117:42–61.