A PROOF OF CONCEPT ON DIALOGUE GAMES FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

ILIA STEPIN
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Departmento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain
ilia.stepin@usc.es

ALEJANDRO CATALA
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Departmento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain
alejandro.catala@usc.es

JOSE M. ALONSO-MORAL
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Departmento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain
josemaria.alonso.moral@usc.es

Abstract

Recent years have witnessed a groundbreaking number of accurate artificial intelligence-based algorithms. However, their oftentimes obscure nature is known to prevent end users from a safe and responsible use or leads to decreased trustworthiness in their predictions or decisions. In this work, we present a novel dialogue game that serves as an explanatory dialogue model to communicate contrastive, selected, and social explanations from an interpretable rule-based classifier to an end user. In addition, we show how it can address the problem of diversity for such explanations via an empirical human evaluation study.

1. Introduction

Explanations are essential for understanding behaviour of complex Artificial Intelligence (AI)-based decision-making systems and placing trust in them. Ribeiro et al. (2016) state that "if the users do not trust a model or a prediction, they will not use it". It is therefore of crucial importance to enhance their output with automatically generated explanations justifying their reasoning in an efficient and comprehensive manner. Further, eXplainable AI (XAI) is considered to be "essential for users to understand, appropriately trust, and effectively manage this emerging generation of artificially intelligent partners" (Gunning, 2021).

Adadi and Berrada (2018) distinguish four main reasons why state-of-the-art AI-based decision-making systems need explanations. First, explanations justify system's decisions. Second, explanations enable developers to detect errors and debug AI-based systems. Third, explanations make it easier to understand the overall behaviour of the system and therefore facilitate the process of continuous improvement of software in accordance with users' suggestions. Finally, explanations allow the explainee (i.e., the end user) to learn and/or infer new facts about the system's behaviour and therefore gain further knowledge about it.

However, only a limited number of explanation generation algorithms are able to efficiently communicate them to end users, partly due to "the lack of a well-defined protocol for evaluating interactive explanations and the challenging process of assessing their quality and effectiveness" (Sokol et al., 2020). In this work, we propose a transparent argumentative explanatory dialogue model that governs bi-directional interaction between an AI-based agent (hereinafter, the system) and the person responsible for making a decision based on the system's decision guided with the corresponding pieces of explanation (hereinafter, the user). We frame the overall process of explanation communication as a dialogue game between the system and the user. The corresponding dialogue protocol includes turntaking rules, a set of requests and replies for the user and the system, respectively, and a set of transition rules that allow for dialogue state switching to ensure effective communication of automated explanations. In order to assess the utility of the proposed explanatory dialogue model, we carry out a human evaluation study. As a result, we show that explainees find the proposed dialogue model satisfying, specifically, in information-seeking dialogue settings. Further, they particularly appreciate the ability to inquire alternative explanations in the course of iterative explanatory dialogue.

2. Explanation and argumentation

Miller (2019) claims that efficient explanations in the context of XAI are contrastive, selected, and social. The property of contrastiveness implies that the explainee expects to receive a justification for the given decision in terms of hypothetical, non-occurring alternative decisions (e.g., "Why P [instead of Q [and/or $Q_1, Q_2, \ldots, Q_n$)]]?" where P is the fact (i.e., the actually produced decision to be explained) and Q, Q_1, Q_2, …, Q_n are foils (i.e., non-made decisions)). In addition, only the most relevant factors leading to the given decision should be included in the explanation, making it selected. Last but not least, explanation is regarded as a social process, i.e., information exchange between the explainer and explainee.

In this section, we exemplify the aforementioned properties of explanation and briefly discuss how argumentation can inherently link them and therefore increase their efficiency.

2.1. Contrastive explanation

Intuitively, an automated explanation for a decision-making system's output justifies the system's decision in terms of the reasons (causes) that led to it. Let us, for example, consider a loan application for a bank client whose monthly income is 1250 euros where the system makes a negative decision upon inspecting the applicant's profile. Example 1 illustrates a piece of explanation justifying the system's decision.

Example 1. *Your loan application is rejected because your monthly income ranges between 1200 and 1500 euros.*

As the explanation from Example 1 indicates only those reasons that motivate the actually made decision (i.e., the fact), we refer to such an explanation as *factual*.

Alternatively, the same decision can be explained contrastively in terms of (one or more) hypothetical non-made decisions (or foils, following the terminology introduced by Lipton (1990)). In the context of XAI, the property of contrastiveness is often captured by means of the so-called *counterfactual* explanations (Stepin et al., 2021). Contrarily to factual explanations, counterfactual explanations oppose two distinct decisions, the fact and a/the foil, while maximising their relevance to the explanandum. Furthermore, Example 2 shows that counterfactual explanations do not only further explain the fact but also offer recommendations on how the system's decision can be changed in favour of the desired alternative.

Example 2. *Your loan application would be approved if your monthly income ranged from 2000 to 3000 euros and if you had less than two active loans.*

2.2. Selected explanation

Automated explanations should offer explainees only a limited number of causes or reasons that led to the decision made (Miller, 2019). Further, all such reasons should be of predominant importance w.r.t. the decision under consideration when it is explained either factually or counterfactually. The number of reasons justifying the automatic decision as well as their contents (e.g., "monthly income is less than 1500 euros") determine the relevance of such an explanation to the decision in question. Indeed, the explanation from Example 1 may sufficiently yet more accurately and concisely motivate the loan rejection decision than, e.g., that from Example 3 for the same client irrespective of how many active loans he or she has got.

Example 3. *Your loan application is rejected because your monthly income is less than 2000 euros and you have more than two active loans.*

Remarkably, the explanations for the system's decision can include imprecise information, i.e., vague linguistic quantifiers (Stepin et al., 2022). Whereas expert users of AI-based systems may require finely-granulated explanations for an automated decision, a wider non-expert audience may be satisfied with those containing (possibly, in part) imprecise information. Example 4 shows how the factual explanation from Example 1 can be approximated with a piece of imprecise information.

Example 4. *Your loan application is rejected because your monthly income is too low.*

Similarly, Example 5 shows how the counterfactual explanation from Example 2 can be simplified using imprecise linguistic quantifiers.

Example 5. *Your loan application would be approved if your monthly income were high and if you had few active loans.*

Examples 1-5 include illustrative rule-based explanations. In such explanations, the consequent represents the system's (possibly, hypothetical) decision that is explained by means of the explanatory features found in the antecedent of a given rule. In this regard, the antecedent-consequent schema of the exemplified explanations may be deemed equivalent to the premise-conclusion schema commonly found in argumentation theories. Further, this representation of rule-based explanations allows us to treat them as arguments following Hempel's theory of explanation, which defines explanation as "an argument to the effect that the phenomenon to be explained, the explanandum phenomenon, was to be expected in virtue of certain explanatory facts" (Hempel, 1965).

In the remainder of this work, we fuse the notions of explanation and argument following Hempel's theory of explanation. In addition, we assume that we have got access to contrastive(-counterfactual) selected

explanations generated automatically to explain an arbitrary decision of an interpretable rule-based AI system.

2.3. Social explanation

In general, any phenomenon can be explained in numerous different ways. In the context of XAI, various distinct explanations can be offered for the same (possibly, non-made) decision. Altogether, sets of explanations for all possible (including those non-made) decisions are said to make up an explanation space.

As shown above, distinct explanations may have different degrees of relevance to the phenomenon being explained (see Examples 1 and 3). In his conversational model of explanation, Lipton (1990) stresses the importance of estimating relevance when evaluating explanations. Indeed, (some of) the most relevant explanations (among those generated by the system) may not be found relevant enough by the end user. Instead, the user may be offered the opportunity to examine the explanation space iteratively, i.e., during his or her interaction with the system. To do so, both parties can be engaged in an argumentative explanatory dialogue where the user explicitly or implicitly evaluates explanations presented by the system.

In such an explanatory dialogue, the system's decision can be considered a claim supported by a rule-based explanation whose features from the antecedent serve as premises for that claim. Subsequent system's responses can then be treated as arguments in favour of its own claim while user's requests attack it. Furthermore, every piece of explanation offered to the end user may be regarded as an argument attacked, as the user rejects the previously offered explanations while exploring the explanation space.

3. Explanatory dialogue game

In this work, we address the social aspect of explanation by modelling explanatory dialogue in form of a dialogue game (Prakken, 2005) between an AI-based system and its user. In this section, we outline essential rules that constitute the dialogue protocol that is proposed to govern interaction. The AI-based system is assumed to be capable of generating high-level textual factual (see Example 4) or counterfactual (see Example 5) explanations or their low-level counterparts (see Examples 1 and 2, respectively). On the one hand, low-level explanations are pieces of aggregated information about the decision made in terms of numerical intervals. On the other hand, high-level explanations are produced after applying the so-called "linguistic approximation" to the low-level explanations (Stepin et al., 2022).

3.1. Turntaking

An explanatory dialogue between the system and the user is said to start taking place when the system makes a decision (i.e., a claim) and communicates it to the user. Therefore, the first dialogue move (m_1) is always made by the system. Each dialogue participant is allowed to make only one move at a time. The dialogue presupposes a sequence of request-response pairs by the user and the system. Only the user is allowed to send queries to the system whereas only the system is authorised to respond to them. Therefore, every even dialogue move (m_2, m_4, ...) is made by the user. Complementarily, every subsequent odd move (m_3, m_5, ...) is made by the system. The dialogue terminates when the user makes an informed decision w.r.t. the system's claim – whether it should be accepted or rejected. Further, the user is allowed to end the dialogue at any time. As a result, an explanatory dialogue is said to contain three main building blocks – claim, explanation, and termination – explanation constituting the principal element of the dialogue.

3.2. Requests and responses

Given poor explanatory capacities of many of the state-of-the-art AI algorithms, it becomes essential to find the balance between the information that the system can exploit when explaining its decisions ("supply") and the anticipated user's requests ("demand"). Provided that the system is equipped with an explanation generation module that is able to offer textual rule-based (factual and counterfactual) explanations upon request, it has to be able to explain not only its own decision but also all the components that constitute its explanations (i.e., the features and the corresponding values found in the antecedent of the related rule).

Following the taxonomy of dialogue moves for earnings conference calls that was introduced by Budzynska et al. (2014, p. 22), we address this challenging problem by proposing the following explanation-related user's requests:

1) *The requests of explanation*: these include general requests for factual (e.g., "Why is my loan application rejected?", see Example 1) and counterfactual explanations (e.g., "What can I do to have my loan application approved?", see Example 2);

2) *The request of clarification*: this request is sent when a definition of a feature is needed (e.g., "What do you mean by *monthly income*?" (see Example 4));

3) *The request of detailisation*: this request aims to specify a feature value if it contains an imprecise quantifier, allowing to switch from a high-level explanation to its low-level equivalent (e.g., "Could you specify how low is *too low*?" (see Example 4));

4) *The request of alternative explanation*: this request allows the user to further explore the explanation space with the aim of finding the piece of explanation that is the most relevant to his or her needs (e.g., "Could you offer me another (counter-)factual explanation?") by rejecting or attacking the pieces of explanation that are found unsatisfactory.

In summary, the set of user's requests includes the request of factual explanation ("why-explain"), the request of counterfactual explanation ("why-not-explain"), the request of clarification ("what-is"), the request of detailisation ("what-details"), the request of an alternative factual explanation ("why-alternative"), and the request of an alternative counterfactual explanation ("why-not-alternative").

Possible system's responses mirror all the user's requests. For each type of requests, the system provides either of the two types of responses: *positive* if the system is able to adequately respond to user's request (i.e., to generate a piece of (alternative) explanation, retrieve the corresponding numerical intervals for the given feature, or provide a definition for it) or *negative* – otherwise. The set of system's responses is therefore said to include the following items: the positive factual explanation response ("explain-f"), the negative factual explanation response ("no-explain-f"), the positive counterfactual explanation response ("explain-cf"), the negative counterfactual explanation response ("no-explain-cf"), the positive clarification response ("clarify"), the negative clarification response ("no-clarify"), the positive detailisation response ("elaborate"), the negative detailisation response ("no-elaborate"), the positive alternative factual explanation response ("alter-f"), the negative alternative factual explanation response ("no-alter-f"), the positive alternative counterfactual explanation response ("alter-cf"), and the negative alternative counterfactual explanation response ("no-alter-cf").

All in all, the proposed requests and responses are considered to sufficiently communicate all the components of rule-based explanations. Further, the alternative explanation requests enable the user to interactively evaluate explanations offered to him or her during iterative exploration of the available explanation space.

3.3. Dialogue state transitions

While the user may desire to have explanations for a (non-)made automatic decision, he or she is by no means obliged to request and subsequently receive any. Furthermore, recent worldwide legal regulations concerning AI (e.g., the European Union's General Data Protection Regulation – GDPR – or the AI Act) aim to ensure the user to have the "right to explanation" (Wachter, 2018, p. 860). Therefore, the explanation-related dialogue block is made optional in all cases.

Given the turntaking rules defined in Section 3.1, the initial dialogue move is made by the system. Recall that the system formulates the claim

in favour of its decision at this stage. The user is subsequently allowed to accept the claim, reject it, or express his or her doubts over it by requesting a factual explanation. Upon receiving a factual explanation, the user is allowed to demonstrate disagreement by asking for (an) alternative factual explanation(-s) or inspecting counterfactual explanations for other (non-made) decisions. Importantly, processing the given piece of explanation is considered finalised when a request for explanation of a different (possibly, non-made) decision is made.

Table I presents all possible dialogue state transitions. The dialogue proceeds from user's request to system's response (a move from the left-most to the central column). Subsequently, the dialogue is processed in a loop. The user can choose a follow-up request (a move from the central to the right-most column). Once selected (a move from the right-most to the left-most column), the request is processed by the system following the same scenario (a move from the left-most to the central column). The dialogue continues until the termination state (i.e., accept or reject) is reached.

Table I. User's requests, possible system's responses, and user's follow-up requests.

User's request	*Possible system's response*	*Possible user's follow-up request*
why-explain	explain-f	why-not-explain, what-details, what-is, why-alternative, accept, reject
	no-explain-f	why-not-explain, accept, reject
why-not-explain	explain-cf	why-not-explain, what-details, what-is, why-not-alternative, accept, reject
	no-explain-cf	why-not-explain, why-alternative, accept, reject
what-details	elaborate	what-details, what-is, why-not-explain, why-alternative, why-not-alternative, accept, reject
	no-elaborate	
what-is	clarify	
	no-clarify	
why-alternative	alter-f	what-details, why-not-explain, what-is, why-alternative, accept, reject
	no-alter-f	what-details, why-not-explain, what-is, accept, reject
why-not-alternative	alter-cf	what-details, why-not-explain, what-is, why-not-alternative, accept, reject
	no-alter-cf	what-details, why-not-explain, what-is, accept, reject

Table II provides the reader with a generic template-based example of an explanatory dialogue modelled following the principles described above.

Table II. An illustrative example of the modelled explanatory dialogue.

Move	*Agent*	*Utterance*
m_1	System	The test instance is of class y_1.
m_2	User	Could you explain why you think so?
m_3	System	It is of class y_1 because <feature$_1$> is <value$_1$>.
m_4	User	What do you mean by <feature$_1$>?
m_5	System	<feature$_1$> is <definition for feature$_1$>.
m_6	User	But why is the test instance not of class y_2?
m_7	System	It would be of class y_2 if <feature$_1$> were <value$_2$> and <feature$_2$> were <value$_3$>.
m_8	User	Could you specify how <feature$_1$> is defined?
m_9	System	<feature$_1$> is defined to be <value$_2$> because it is found in the interval <[value$_{2min}$, value$_{2max}$]>.
m_{10}	User	But why is the test instance not of class y_3?
m_{11}	System	It would be of class y_3 if <feature$_1$> were <value$_3$> and <feature$_3$> were <value$_4$>.
m_{12}	User	I am not quite satisfied with your explanation. Could you offer me another one?
m_{13}	System	Sure! It would be of class y_3 if <feature$_1$> were <value$_5$> and <feature$_4$> were <value$_2$>.
m_{14}	User	Ok, now I trust your decision.

4. Insights from the human evaluation study

4.1. Experimental settings

In order to validate the utility of the proposed dialogue model, we carried out a human evaluation study. The participants were first to choose one of the three given use cases: basketball player position classification[1], beer style classification[2], or thyroid disease diagnosis[3]. In all the cases, the study participants were given numerical characteristics of an object following the selected use case (i.e., those of a basketball player, an instance of beer, or hormone levels of a hypothyroid patient, respectively). The details of the objects are outlined in Table III. The system's decision determined the position of the basketball player, the style of the beer instance, or the type of the thyroid disease that the patient had (if any).

1 https://gitlab.citius.usc.es/jose.alonso/basketballplayers-dataset
2 https://dx.doi.org/10.13140/RG.2.2.20313.67680
3 https://doi.org/10.24432/C5D010

Table III. The characteristics of the objects upon which the system made a decision (i.e., classification).

Use case	*Characteristics of the object*	*System's decision*	*Alternative decisions*
Basketball player position	Height = 1.85; minutes played = 21.19; points scored = 9.2; two-points field goals percentage = 43.1; three-points field goals percentage = 40.0; free throws percentage = 81.9; rebounds = 1.9; assists = 3.8; blocks = 0.0; turnovers = 0.7; global assessment = 8.8	Point-guard	Shooting-guard, Small-forward, Power-forward, Center
Beer style	Colour = 2; bitterness = 18; strength = 0.049	Blanche	Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian Strong Ale
Thyroid disease	Thyroid-stimulating hormone (TSH) = 4.6; triiodothyronine (T3) = 1.2; total thyroxine (TT4) = 48; thyroxine utilisation rate (T4U) = 0.89; free thyroxine index (FTI) = 54	Secondary hypothyroid	No hypothyroid, Primary hypothyroid, Compensated hypothyroid

The study participants were placed in the information-seeking dialogue settings (following the dialogue type classification by Walton and Krabbe (1995)). They interacted with the system without having any prior knowledge about the system or the dataset used. Instead, they were only given the object's characteristics outlined in Table III and the system's decision. The participants interacted with the system until they could make an informed decision on whether the system's decision was credible enough. Notably, the system's decision was correct in all the cases. However, the study participants had not been informed about it prior to their interaction with the implemented dialogue system.

Decision trees were used as classifiers representing the AI-based system, with branches of trees translated into IF-THEN rules for the purpose of explanation generation. Multiple counterfactual explanations were generated using the XOR algorithm for counterfactual explanation generation for rule-based classifiers (ibid). Only continuous features were used for training the system in all the use cases.

Overall, we collected 60 explanatory dialogue transcripts: 14 participants selected the basketball player position use case, 37 – the beer

style classification scenario, 9 – the thyroid disease diagnosis scenario. All the collected data were anonymous and obtained strictly after receiving an explicit informed consent from the study participants.

4.2. Request utility

All the proposed requests were extensively used by the study participants in all the use cases. Table IV shows the absolute numbers of all the types of requests made by the users as well the relative numbers to the overall number of explanation-related requests in all the use cases.

Table IV. Numbers of explanation-related requests submitted to the system by the study participants.

User's request	*Use case*		
	Basketball player position	*Beer style*	*Thyroid disease diagnosis*
Factual explanation	12 (17.91%)	36 (15.32%)	9 (31.04%)
Counterfactual explanation	18 (26.87%)	50 (21.28%)	8 (27.59%)
Detailisation	15 (22.39%)	78 (33.19%)	6 (20.69%)
Clarification	13 (19.40%)	46 (19.57%)	3 (10.34%)
Alternative counterfactual explanation	9 (13.43%)	25 (10.64%)	3 (10.34%)
In total	**67 (100%)**	**235 (100%)**	**29 (100%)**

As it can be observed from Table IV, a majority of the study participants requested (at least, factual) explanations to the system's decision. Furthermore, most of such study participants accepted the system's decision at the end of their interaction with the system (91.67% in the basketball player and the beer style scenarios, 55.56% in the thyroid disease scenario).

Remarkably, responses to all the types of requests are found to trigger users' final decisions. In the case of the basketball player position scenario, 38.46% of the participants accepted the system's decision once an alternative counterfactual explanation was offered to them. In the case of the beer style scenario, 29.41% of the participants accepted the decision when high-level counterfactual explanations as well as low-level explanations of either kind were presented to them. Finally, 40.00% of the study participants accepted the system's decision once their clarification requests were responded to in the thyroid disease scenario.

4.3. Free-form user feedback

Upon completion of the experimental task, the study participants were asked to optionally leave free-text responses to the following questions and/or suggestions: (Q1) "If you could add other types of requests to the system, what would those be?"; (Q2) "Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?"; (Q3) "If you have any other comments for us, please leave them in the textbox below". Below we summarise the most informative comments.

As for Q1, the study participants would like to extend the actual dialogue model so that it could further inform them about the domain knowledge available to the system as well as the technical details on how the classification was obtained. In addition, some users wished to have access to the previously processed explanations.

As for Q2, a number of commentators expressed their satisfaction with the offered explanations. Further, the automated explanations were largely deemed convincing w.r.t. the system's claim. In addition, some study participants positively assessed the ability to query the system for counterfactual explanations and further details and clarifications. Finally, the study participants positively commented on the ability to request counterfactual explanations for hypothetical system's decisions.

As for Q3, some study participants pointed to the following limitations of the proposed explanatory dialogue model. First, the system's responses appeared unnaturally fast. Second, a use of visualisation tools was desired to support the textual explanations. Finally, the difference between the detailisation and clarification requests seemed rather unclear to one study participant.

5. Conclusion

In this work, we designed a dialogue game that serves the task of communicating automatically generated rule-based explanations for an AI-based decision-maker. We showed that the proposed rule-based dialogue protocol represents a transparent mechanism of factual and counterfactual explanation communication.

The designed dialogue game has multiple potential applications. Due to the modularity of the protocol, it is flexible enough to be adapted for running experiments on the trustworthiness, satisfaction, and/or persuasive capability of automatically generated explanations. In addition, it can be used as a benchmark for evaluating the effectiveness of automatic rule-based explanation generation algorithms, as it operates on the full explanation space.

Both quantitative and qualitative results of the human evaluation study carried out to validate the dialogue model confirm the necessity in all the proposed requests and responses for explanatory information-seeking human-machine dialogue. In addition, user free-form feedback shows that the proposed dialogue model is, in principle, found to be an efficient explanation communication tool to support collaborative human-machine decision-making.

The present dialogue game-based framework opens the door for several prospective lines of research. First, the dialogue protocol needs to be enhanced with the capability to handle explanations making use of non-numerical (e.g., categorical) features that cannot be quantified to ensure the protocol's operability for any decision-making scenario. This implies necessary changes in the set of dialogue state transitions, as detailisation requests may need to be made unavailable for categorical features whose meaningful numerical interpretations are unavailable (e.g., gender). In addition, further experiments are necessary with classifiers whose feature space is poorly interpretable (if at all). Second, the explanatory dialogue settings as well as user's expectations may impose the requirement of tackling visual or multi-modal explanations. It therefore appears important to integrate mechanisms of communicating textual explanations with those of non-textual nature. Finally, further human evaluation experiments need to be designed to validate the aforementioned dialogue protocol improvements.

Acknowledgements

Ilia Stepin is an *FPI* researcher (grant PRE2019-090153). This work was supported by the Spanish Ministry of Science and Innovation (grants PID2021-123152OB-C21, and TED2021-130295B-C33) and the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19). All the grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

Adadi, A., & Berrada M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Budzynska, K., Rossi, A., & Yaskorska, O. (2014). Financial dialogue games: A protocol for earnings conference calls. *Proceedings of the Conference on Computational Models of Argument*. IOS Press, 19-30. https://doi.org/10.3233/978-1-61499-436-7-19

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2(4):e61. https://doi.org/10.1002/ail2.61

Hempel, C. G. (1965). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press.

Lipton, P. (1990). *Contrastive explanation.* Royal Institute of Philosophy Supplement, 27, 247-366. https://doi.org/10.1017/S1358246100005130

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence, 267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Prakken, H. (2005). Coherence and Flexibility in Dialogue Games for Argumentation. Journal of Logic and Computation, 15(6), 1009-1040. https://doi.org/10.1093/logcom/exi046

Ribeiro, M. T., Singh, S., & Guestrin C. (2016). "Why Should I Trust you?": Explaining the Predictions of Any Classifier. *Proceedings of the International Conference on Knowledge Discovery and Data Mining,* 1135-1144. https://doi.org/10.1145/2939672.2939778

Sokol, K. & Flash, P. (2020). One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI – Künstliche Intelligenz, 34,* 235-250. https://dx.doi.org/10.1007/s13218-020-00637-y

Stepin, I., Alonso, J.M., Catala, A., & Pereira-Fariña, M. (2021). A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access, 9,* 11974-12001. https://doi.org/10.1109/ACCESS.2021.3051315

Stepin, I., Alonso, J.M., Catala, A., & Pereira-Fariña, M. (2022). An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences, 618*, 379-399. https://doi.org/10.1016/j.ins.2022.10.098

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology, 31(2),* 841-887. https://dx.doi.org/10.2139/ssrn.3063289

Walton, D., & Krabbe, E. C. (1995). *Commitment in dialogue: Basic concepts of interpersonal reasoning*. New York: SUNY Press.