Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210165

On the Rationality of Explanations in Classification Algorithms

Zoe FALOMIR [a,b] Vicent COSTA [c,d,1],
[a] *Universitat Jaume I, ES Tecnologia i Cincies Experimentals*
[b] *Bremen Spatial Cognition Center (BSCC)*
[c] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela*
[d] *Artificial Intelligence Research Institute (IIIA-CSIC)*

Abstract. This paper is a first step towards studying the rationality of explanations produced by up-to-date AI systems. Based on the thesis that designing rational explanations for accomplishing trustworthy AI is fundamental for ethics in AI, we study the rationality criteria that explanations in classification algorithms have to meet. In this way, we identify, define, and exemplify characteristic criteria of rational explanations in classification algorithms.

Keywords. Rationality, rational explanation, explainable AI, classification algorithm, AI ethics, trustworthy AI

1. Introduction

Explainability is a fundamental topic for AI ethics. It increases users' trust in the outcomes of AI systems, eases to clarify the decisions made and whether the system has been trained on a biased/fair view of the world. But to accomplish these ethical goals, explanations need to be rational. In this way, this paper is a preliminary study of the criteria that rational explanations produced by classification algorithms must meet. Furthermore, examples of explanations produced by AI systems are extracted from research works in the state-of-the-art and used for discussion. We envision that these criteria may enable us to give a measure of how rational the explanations produced by AI systems are.

1.1. Trustworthy AI needs rational explainability

Broadly speaking, explainable artificial intelligence (XAI) is the discipline that studies those systems that describe their results, actions, or decisions. Regarding explainable classification algorithms, XAI yields reasons for the classification results obtained. Note that the lack of explainability decreases the trust in the outcomes of the AI systems, it reduces its fairness (without explainability, responsibility cannot be claimed) and usability [1], and makes it more probable to overlook whether they have been trained using a biased view of the world [2,3] (for a reference on biases in AI see [4,5]).

[1] Corresponding Author: vicent.costa@protonmail.com.

In the literature, AI systems that describe their outcomes in natural language are numerous and involve very diverse domains as finance, medicine, or social services. In this way, the demand for accountability regarding undesirable outcomes produced by algorithms requires the use of XAI; but not only that, explainability involves several topics as safety, transparency, or fairness. In addition, the General Data Protection Regulation[2], established by the European Parliament, includes the individual right to explanation, which affects AI systems and their users. Considering all these reasons, it is clear that XAI becomes a mandatory project.

However, not all descriptions of the outcomes appearing in the literature can be considered as adequate explanations. Similarly to human-human interactions, simple/incomplete/irrational descriptions might not be enough to discover bias in the design or training of the AI system, so they will not help to enlighten unethical results of their classification systems. Also, we cannot require trust if users do not find rationality in the explanations provided by algorithms. Therefore, providing rational explanations is a desirable property for AI classification models, and the present paper is a preliminary work on that direction.

1.2. How much does the ball cost? Biased thinking vs. rational thinking

In the literature, there are research studies that distinguish between two types of cognitive processes: those executed quickly with little conscious deliberation and those that are slower and more reflective [6,7,8,9]. These type of processes were called *System 1* and *System 2*, respectively [10]. *System 1* processes occur spontaneously and not require or consume much attention, i.e., recognizing a face, whereas *System 2* processes involve mental operations that require effort, concentration, and probably the execution of some learned rules, i.e., calculating 123x45 without a calculator. Note that the task 123x45 offers no intuitive solution, that is, no number spontaneously comes to your mind as a possible answer. Let us exemplify these two type of processes with an easy question:

A bat and a ball cost 1.10 €. The bat costs 1.00 € more than the ball.
How much does the ball cost? ___ cents.

This is a question in the Cognitive Reflection Test (CRT) [11]. In this case, individuals who answer this question with the first idea that comes to mind usually fail this test[3]. They are also biased reasoners since they are quite convinced that the problem is very easy since they estimate that 92% of the people would solve this problem right [11]. This exemplifies that we people can be mistaken and unaware, specially in situations where we think superficially. That is a flaw in the reflective mind, a failure of our rationality.

Moreover, as [9] mentions, people are supposed to have consistent preferences that they trust since they reflect their interests. But sometimes our decisions do not produce the best possible experience for us. Note that taste and decisions are shaped by memories and the duration of these memories can be neglected/biased. For example, usually people give the same importance/weight to the good and to the bad part of an experience, although the good part have lasted ten times longer than the other.

Thus, AI systems that produce rational explanations may also help individuals to realise other sides of their judgment that might be biased, so that individuals can also see

[2]https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[3]Note that "10 cents" is a wrong answer. If the ball costs 10 cents, then the bat will cost 1 euro, so the difference would be 90 cents.

the situation differently and even learn from it. XAI systems can transmit their knowledge and arguments to the users, so if the system is right, then users can learn from it, or in case the AI system is mistaken, users can correct it or update it with the missing information.

2. Preliminaries: the notions of *rationality* and *explanation*

Rationality is a feature commonly included in the diverse theories on intelligence and recently has been introduced as a research topic into experimental and engineering sciences [12]. In the previous section, we argued that trustworthy AI needs rationality, but note that we could add that since rationality is conceived as a crucial part of intelligence, general AI requires addressing the question of rationality. The notion of *rationality* is usually defined as the quality of being based on or following reason. It applies to a wide range of domains and entities [13]. For instance, it is common to say that some emotions, beliefs, societies, reactions, or people themselves are rational. Many different disciplines, such as philosophy, economics, or psychology, have studied rationality from diverse points of view (for a review of the study of this topic, the reader is referred to [13]). Therefore, it is not surprising that rationality has been conceptualized and modeled in different forms. For instance, mainly four kinds of models of rationality have been considered in cognitive science and psychology (see [14] for a general presentation). And the diversity of proposals is even richer in philosophy [13]. Research in AI has been focused on the study of rational agents, and the design of agents often uses the belief-desire-intention model [15] as a rationality paradigm. The utility-maximizing account of rationality, inspired by the instrumental approach from philosophy, has been predominant in AI, even if it is not reasonable to consider an agent acting in this fashion rational [12].

As a complementary approach, this paper focuses on the question of rationality applied to explainable classification algorithms (we refer the reader to [16] for a description of the notion of classification in AI). However, as it occurs with the notion of rationality, there is no canonical definition of explanation. Diverse philosophical approaches conceptualized explanations in different and quite irreconcilable ways. In the literature, an explanation has been defined as a deductive relation, a probabilistic relation, a provision of causes, or a causal relation (we refer the reader to [13, Ch.19]). So, defining the notion of explanation is still a philosophical problem. In this paper, we consider that an *explanation* is any proposition generated by a classification algorithm and labeled by the system itself as such. Next, we can proceed with the principal issue of this work, that is, seeking the criteria that determine whether an explanation of a classification algorithm is rational. For that, we define rational explanations as those meeting certain criteria.

3. Defining *rational explanations*

In this section, we identify, define, and exemplify the criteria of rational explanations in classification algorithms.

3.1. Human understandable

Trivially, the rationality of an explanation can only be analyzed whether the explanation is understandable by people. Instances of explanations satisfying this criterion are: any

response of the classical expert system PROSPECTOR written by Richard O. Duda [17] (e.g., *On a scale from -5 to 5, my certainty in MVTD is now: .8995*, or the following answer of the XAI Beer Style Classifier presented in [18]: *It is very likely that this beer is Branche, because its color is pale, its bitterness is low, and its strength is session*.

3.2. Conceptual

Concepts are essential to human thought and play a fundamental function in explaining cognitive tasks as decision-making or categorization [19,20]. Furthermore, according to conceptual coherentism, it is impossible to believe, understand or trust in a proposition (in particular, an explanation) without having the concepts that figure essentially in it [13, Ch.1]. Hence concepts play a very significant role in XAI. Indeed, several philosophical theories on concepts (e.g., connectionism, the analogical approach, or the classical-symbolic one) have inspired diverse AI frameworks like neural networks or formal ontologies [19]. Regarding concepts, we propose three conditions for rational explanations:

- Rational explanations have to use concepts to proceed with the categorization and employ them for building the propositions explaining the result. The explanation *This leaf is not like any known leaf species, but it is round with toothed and lobed margin* [21] meets this condition.
- The concepts used by the explanation must be coherent with human perception, that is, AI systems must align their perception (sensor data) to concepts that people can understand and that they usually use to communicate i.e. in natural language. In the literature, approaches that define concepts that are aligned with human perception are for instance: qualitative descriptors based on reference systems [22], conceptual spaces [23], data-driven conceptual spaces [21], ontologies [24], or fuzzy sets [25]. Note that the mere appearance of concepts is not enough to ensure that the explanations displayed are human understandable. For instance, some machine learning algorithms might build meaningless linguistic labels related to concepts. So, later we must aligned these learned labels with human understanding to gather meaning for the users.
- The concepts used to classify an item into a group have to be related to the distinctive and relevant group's traits. Indeed, an algorithm could correctly classify an item into a category using concepts, and associate them with meaningful linguistic labels, but it could be that these concepts have nothing to do with the group's characteristics within which the item is classified. For instance, shortcuts may fail in real-world scenarios far from standard benchmarks and classify the breed of a dog according, for example, to the presence of the snow in the image. This would lead to an explanation of the form *The dog is a Siberian Husky because it is found in a snowed mountain*, which does not include any dog's traits.

3.3. Context adequate

In linguistics, the goal of Referring Expression Generation (REG) is to find a set of properties (minimal, or psychologically plausible) that are all true for the intended referent, but not all true for any distractor [26,27]. So, in this case, the context of the referent object and the distractor objects must be taken into account.

When referring to an object, multiple categorizations are usually possible [28]. For example, assuming a dog has the following properties: *small*, *black and white*, *tall* and *have spots*, someone may refer to it by saying *the tall black and white dog with spots is a Dalmatian*. However, if that dog is in a context where other dogs are also tall, then including the attribute *tall* in the explanation is not cognitively adequate, i.e., people will not say it since it does not help to distinguish it. If there are several tall dogs and only one dog is small, then we can refer to it by the property that distinguishes it from the rest, i.e., *the small dog is a chihuahua*. In this case, if there are no more small dogs, the color and if it has spots or not is not significant according to the context.

3.4. Personalized according to users' background

The rationality of some explanations depend on the user's background, in the sense that one might need specific knowledge to understand them. To exemplify this, let us consider the classification's explanation of a basketball player presented in [29]: *The player is Center because Height is extremely high and Rebounds is medium. There is also a minor chance that it is Small-Forward*. Although the language of this explanation is not the most natural, individuals familiar with basketball may understand it with little effort. However, individuals with no background knowledge in this sport would need an explanation including more suitable terms such as *the player is Center because s/he is the tallest*[4]. Also the notion of *medium* would need more contextualisation, that is, which is the typical number of rebounds in a play, and then what is considered a high/low number of rebound catching for a player. Knowing *high* or *low*, that is the context, we can infer *medium* number of rebounds, but without knowing the thresholds, the meaning of the concept *medium* remains unclear (see Section 3.3 for further details). So, the more adapted to the user's background is an explanation, the more rational the explanation.

3.5. Coherent with observable human reasoning

As argued in [30], decision methods must support known patterns of human reasoning. Similarly, we propose another rationality criteria: classification explanations need to be coherent with observable and rational patterns of human reasoning (these patters depend on the classification problem considered). Otherwise, users could not recognize the rationality of the reasoning behind a classification result. For instance, people classify the breed of a dog using reasoning involving its size, its fur, shape, eye type, etc. Let us consider the chihuahua breed, whose average weight ranges [1.8, 2.7] kilograms. Then, if a classification algorithm explained that *This dog is not a chihuahua because its weight is 2.8 kilograms, which is larger than 2.7 kilograms*, users would find this explanation not rational, since it is not the kind of observable human reasoning we do for classifying dogs. A more cognitive statement would include argumental steps like *If a dog is 1.2 meters tall, then this dog cannot ever be classified as a chihuahua*.

Some theoretical tools used to accomplish this objective are logic aggregators (the common idea is to use inputs and outputs of logic aggregators as the conditions filtering those criteria that might serve in mathematical models of human reasoning). For example, when classifying a picture into a painting style, people may have input percepts of the degrees of the adequacy of the image concerning the distinctive traits of the painting

[4]Note that basketball players are usually ordered by height 1–5 and the center is generally number 5.

styles. Humans would aggregate the degrees corresponding to the different features to form a composite percept, assigning a degree of membership of a picture to the painting styles. Based on this theory [30], in [31] the idempotent aggregation, the noncommutativity, and the non-use of annihilators are identified as characteristic patterns of human aggregative reasoning related to the art painting style categorization, and used to explain the results. The idempotent aggregation is based on the assumption that the membership degree to an art style must be between the lowest and the highest value of the traits of this style. The noncommutative reflects that for each painting style, each color trait may have its degree of importance. The explanation classification algorithm presented in [25] did not consider this pattern. An annihilator is an extreme value of suitability (either 0 or 1 – necessary and sufficient condition, respectively) of a feature that is sufficient to decide the result of aggregation regardless of the values of other inputs. An example of the explanations shown in this work is the following one [31]:*The painting is classified in the Post-Impressionism style. The high contrasts between red and green, and between blue and yellow evidence this style*. In that case, the item has been classified taking into account the three patterns of human aggregative reasoning mentioned.

3.6. Contrastive and counterfactual

Factual explanations are based on the features of the input data instances. In contrast, we find contrastive and counterfactual explanations. The conceptual similarity between these two last kinds of explanations motivates us to present them together (see [32] for a detailed introduction to contrastiveness and counterfactuals).

On the one hand, contrastive explanations in classification algorithms give reasons why an item is not classified differently. This kind of explanation describes a classification result by answering the question *Why was the item classified in P rather than in Q?*. In brief, an explanation is contrastive whenever faces the classification result to one of the possible other classifications. Diverse research argues that contrastive explanations are inherent to human cognition [33,34,35], and thus relevant to XAI. In philosophy, contrastive explanations are claimed to be necessary for moral responsibility [36]. And, furthermore, two of the four types of explanatory questions identified by Van Bouwel and Weber [37] lead to contrastive explanations. In light of this, contrastiveness should be a criterion of rational explanations. In [38] the authors propose a method to use questions of this type to restrain the set of features of machine learning algorithms. Often the result is a contrastive explanation, as the following example shows [38]: *System: The flowertype is Setosa. User: Why Setosa and not Versicolor? System: Because for it to be Versicolor the petal width (cm) should be smaller and the sepal width (cm) should be larger*.

On the other hand, counterfactuals picture alternative scenarios to that occurred in fact. Usually, they are presented as conditionals, where the antecedent represents the alternative case, and the consequent describes the consequences of the antecedent. Thus, regarding classification algorithms, counterfactual explanations answer the question *What would have occurred if the item did not hold the property P?*. In short, a counterfactual explanation reveals how the classification could have been different. As Hume pointed out for the first time, counterfactuals explanations play a crucial role in exploring the causes of an event. And more recent approaches from philosophy agree with this [39]. For a cognitive study of counterfactual reasoning, we refer the reader to [40]. Also, some references on ethical aspects related to counterfactual explanations are [41,42,43].

In this way, rational explanations should also include counterfactuals whenever is possible. For instance, the *what-if* tool[5] from *Google* also permits to obtain counterfactual explanations, and in [44] the authors present the method Counterfactual Local Explanations viA Regression (CLEAR). For example, to the question *What if the person does not have a head in the video?*, the XAI model presented in [45] would answer with the counterfactual explanation *Is it possible for a person to exists without the head*. However, as shown in [46] beliefs about additional conditions take precedence over beliefs about presupposed facts for counterfactuals, and thus counterfactuals are not equally helpful in assisting human comprehension. So, future work will consider the criteria established in [46].

4. Case study: explanations on art painting style categorization

In this section, we use the characteristic criteria of rational explainability defined earlier to study the explanations' rationality provided by the art painting style classification algorithm ℓ-SHE [25]. This classifier integrates qualitative descriptors and t-norm based logics and classifies painting from the Baroque, Impressionism, and Post-Impressionism styles using only color features. The main objective of this part is, thus, to exemplify how the research presented in Section 3 might help not only to analyze the explainability of an algorithm but also to improve the rationality of its explanations.

Let us consider the explanations provided by the ℓ-SHE for Renoir's painting *Le djeuner des canotiers* [25]: *rn3 [Le djeuner des canotiers] is an Impressionist painting. The diversity of qualitative colours evidences the Impressionism style. The variety of hues evidences the Impressionism style. The amount of bluish evidences the Impressionism style. The amount of grey evidences the Impressionism style*. Arises the following question: how rational is this explanation? Let us analyze the explanation using the criteria proposed in the previous section. *Is this explanation ...*

- human understandable? It uses natural language and users can understand it.
- conceptual? It utilizes concepts to proceed with the categorization. The concepts (the diversity of color and hues, and the levels of blue and grey) are coherent with human perception. However, color is not the unique distinctive, and relevant trait related to Impressionism; and not considering other features (e.g., the strokes of the picture, or the painting's theme) might not be considered very rational in some classifications.
- context adequate? The ℓ-SHE algorithm selects one or two reasons (from a larger number of facts) to be the explanation, i.e., highlights the more relevant features that characterize the style obtained. Hence we may affirm that the explanation satisfies the criterion.
- personalized according to users' background? Users do not need to be art experts to understand the explanation provided by the system since it is based in color concepts, and color traits that are common sense, that is, most people would understand them. Therefore, the explanation meets this criterion.
- coherent with observable human reasoning? The ℓ-SHE algorithm does not model observable human reasoning behind art painting style classification, as it is indicated in [31]. For example, the categorizations of the styles used in the ℓ-SHE

[5] https://pair-code.github.io/what-if-tool/.

algorithm use t-norms, which admits annihilators. However, art seems too diverse to justify the use of annihilators. For instance, a high level of darkness is distinctive of the Baroque style but other art styles can also present a high level of darkness (e.g., the Romanticism style). Conversely, the absence of this feature is not enough to dismiss this style: some paintings from the Baroque show a few uses of dark colors. Thus, the explanation analyzed does not meet the coherence criteria.

- contrastive and counterfactual? The explanation does not meet the contrastive and counterfactual criteria. This affects the rationality of the explainability of classifications. The accuracy of ℓ-SHE is around 70%. So, often users would ask why an image has been classified into a style rather than another, seeking a contrastive explanation. Or they would wonder what would have been the classification if, for instance, the level of darkness was higher, etc. Therefore, this is a clear issue to improve the explainability of this algorithm.

According to the criteria presented in this paper, we could say that the rationality of the explanation studied is medium. Indeed, it fully meets half of the criteria, two more criteria are few accomplished, and the remaining one is not satisfied. We also remark that this analysis of rationality guides future improvements of the ℓ-SHE algorithm.

5. Conclusions and future work

In this paper, we highlight that rational explanations for AI systems are fundamental. Especially for preserving the users' trust in AI systems and for showing their fairness. We also show that, although rationality is considered as an inherent human feature, on some occasions our thinking might be biased (e.g., ball-and-bat problem).

The main contribution of this paper is the discussion about the criteria that rational explanations must meet. As far as we are concerned, there are no standards for that, but they are very needed. So, this is a step further in that direction. Any explanation produced by an AI system nowadays has been designed by a human. But, as a larger audience would agree, that does not ensure rational explainability. This is why we propose some criteria here in this paper as a guideline to create rational explanations and which are summarized as follows: an explanation produced by an AI system must be human understandable, use concepts, be adequate to the context of communication, be personalized according to the user's background, be contrastive and be coherent with observable human reasoning.

As future work, we intend to explore further the criteria of rational explanations presented in the paper. In addition, we aim to create a dataset showing which explanations in the literature meet the conditions for rationality proposed in this work. Furthermore, we also consider developing an approach for quantifying and qualifying the rationality of explanations provided by classification algorithms. The results obtained by this approach will be compared to the results of a survey carried out to individuals with different profiles (i.e., experts and non-experts in a topic).

Acknowledgments

We thank the referees for their comments. Z. Falomir acknowledges funds from a Ramon y Cajal (RYC2019-027177-I/AEI/10.13039/501100011033) awarded by the MICINN.

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, R. Rastogi (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144.

[2] W. Samek, K. Müller, Towards explainable artificial intelligence, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Vol. 11700 of Lecture Notes in Computer Science, Springer, 2019, pp. 5–22.

[3] H. Hagras, Toward human-understandable, explainable AI, Computer 51 (9) (2018) 28–36.

[4] L. Devillers, F. Fogelman-Soulié, R. Baeza-Yates, AI & human values - inequalities, biases, fairness, nudge, and feedback loops, in: B. Braunschweig, M. Ghallab (Eds.), Reflections on Artificial Intelligence for Humanity, Vol. 12600 of Lecture Notes in Computer Science, Springer, 2021, pp. 76–89.

[5] R. Baeza-Yates, Bias on the web, Commun. ACM 61 (6) (2018) 54–61.

[6] S. Epstein, Integration of the cognitive and the psychodynamic unconscious., American psychologist 49 (8) (1994) 709.

[7] S. A. Sloman, The empirical case for two systems of reasoning, Psychological Bulletin 119 (1996) 3–22.

[8] D. Kahneman, S. Frederick, Representativeness revisited: Attribute substitution in intuitive judgment., in: T. Gilovich, D. Griffin, D. Kahneman (Eds.), Heuristics & Biases: The Psychology of Intuitive Judgment., New York. Cambridge University Press., 2002, pp. 49–81.

[9] D. Kahneman, Thinking, fast and slow, Farrar, Straus and Giroux, New York, 2011.

[10] K. E. Stanovich, R. F. West, Individual differences in reasoning: Implications for the rationality debate?, Behavioral and Brain Sciences 23 (5) (2000) 645665.

[11] S. Frederick, Cognitive reflection and decision making, Journal of Economic perspectives 19 (4) (2005) 25–42.

[12] T. R. Besold, S. L. Uckelman, Normative and descriptive rationality: from nature to artifice and back, J. Exp. Theor. Artif. Intell. 30 (2) (2018) 331–344.

[13] A. Mele, P. Rawling, The Oxford Handbook of Rationality, Oxford Handbooks, Oxford University Press, 2004.

[14] T. R. Besold, Rationality in/for/through ai, in: J. Kelemen, J. Romportl, E. Zackova (Eds.), Beyond artificial intelligence, Vol. 4, Springer, 2013.

[15] M. Bratman, Intention, plans, and practical reason, Harvard University Press, 1987.

[16] M. Fumagalli, G. Bella, F. Giunchiglia, Towards understanding classification and identification, in: A. C. Nayak, A. Sharma (Eds.), PRICAI 2019: Trends in Artificial Intelligence - 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part I, Vol. 11670 of Lecture Notes in Computer Science, Springer, 2019, pp. 71–84.

[17] R. O. Duda, S. International., G. S. (U.S.)., N. S. F. (U.S.), Development of the Prospector Consultation System for Mineral Exploration: final report, covering the period October 1, 1976 to September 30, 1978 / by Richard O. Duda ... [et al.], SRI International Menlo Park, Calif, 1978.

[18] J. M. Alonso, A. Ramos-Soto, C. Castiello, C. Mencar, Explainable AI beer style classifier, in: K. Martin, N. Wiratunga, L. S. Smith (Eds.), Proceedings of the SICSA Workshop on Reasoning, Learning and Explainability, Aberdeen, Scotland, UK, June 27, 2018, Vol. 2151 of CEUR Workshop Proceedings, CEUR-WS.org, 2018.

[19] M. Fumagalli, R. Ferrario, Representation of concepts in AI: towards a teleological explanation, in: A. Barton, S. Seppälä, D. Porello (Eds.), Proceedings of the Joint Ontology Workshops 2019 Episode V: The Styrian Autumn of Ontology, Graz, Austria, September 23-25, 2019, Vol. 2518 of CEUR Workshop Proceedings, CEUR-WS.org, 2019.

[20] G. L. Murphy, The Big Book of Concepts, MIT Press, 2002.

[21] H. Banaee, E. Schaffernicht, A. Loutfi, Data-driven conceptual spaces: Creating semantic representations for linguistic descriptions of numerical data, J. Artif. Int. Res. 63 (1) (2018) 691742.

[22] K. D. Forbus, Qualitative modeling, Wiley Interdisciplinary Reviews: Cognitive Science 2 (4) (2011) 374–391.

[23] P. Gärdenfors, Conceptual Spaces, A Bradford Book, 2004.

[24] D. Arvor, M. Belgiu, Z. Falomir, I. Mougenot, L. Durieux, Ontologies to interpret remote sensing images: why do we need them?, GIScience and Remote Sensing 56 (2019) 1–29.

[25] V. Costa, P. Dellunde, Z. Falomir, The logical style painting classifier based on horn clauses and explanations (l-she), Log. J. IGPL 29 (1) (2021) 96–119.

[26] E. Krahmer, K. van Deemter, Computational generation of referring expressions: A survey, Computational Linguistics 38 (1) (2012) 173–218.
[27] R. Dale, E. Reiter, Computational interpretations of the gricean maxims in the generation of referring expressions, Cognitive Science 18 (1995) 233–263.
[28] V. Mast, Z. Falomir, D. Wolter, Probabilistic reference and grounding with pragr for dialogues with robots, Journal of Experimental & Theoretical Artificial Intelligence 28 (5) (2016) 889–911.
[29] J. M. Alonso, Explainable artificial intelligence for kids, in: V. Novák, V. Marík, M. Stepnicka, M. Navara, P. Hurtík (Eds.), Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2019, Prague, Czech Republic, September 9-13, 2019, Vol. 1 of Atlantis Studies in Uncertainty Modelling, Atlantis Press, 2019.
[30] J. Dujmović, Soft Computing Evaluation Logic: The LSP Decision Method and Its Applications, Wiley - IEEE, 2018.
[31] V. Costa, The art painting style classifier based on logic aggregators and qualitative colour descriptors (C-LAD), in: S. Rudolph, G. Marreiros (Eds.), Proceedings of the 9th European Starting AI Researchers' Symposium 2020 co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), Santiago Compostela, Spain, August, 2020, Vol. 2655 of CEUR Workshop Proceedings, CEUR-WS.org, 2020.
[32] I. Stepin, J. M. Alonso, A. Catalá, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.
[33] R. M. J. Byrne, Spatial mental models in counterfactual thinking about what might have been, Trends in Cognitive Sciences 6 (10) (2002) 426–431.
[34] S. Chin-Parker, A. Bradner, A contrastive account of explanation generation, Psychonomic Bulletin & Review 24 (5) (2017) 1387–1397.
[35] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[36] N. Elzein, The demand for contrastive explanations, Philosophical Studies 176 (5) (2019) 1325–1339.
[37] J. Van Bouwel, E. Weber, Remote causes, bad explanations?, Journal for the Theory of Social Behaviour 32 (4) (2002) 437–449.
[38] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, M. A. Neerincx, Contrastive explanations with local foil trees, CoRR abs/1806.07470.
[39] J. Woodward, Making things happen: a theory of causal explanation, Oxford University Press, Oxford, 2003.
[40] R. M. J. Byrne, Cognitive processes in counterfactual thinking about what might have been, Psychology of Learning and Motivation 37 (1997) 105–154.
[41] M. J. Kusner, J. R. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 4066–4076.
[42] R. Poyiadzi, K. Sokol, R. Santos-Rodríguez, T. D. Bie, P. A. Flach, FACE: feasible and actionable counterfactual explanations, in: A. N. Markham, J. Powles, T. Walsh, A. L. Washington (Eds.), AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020, ACM, 2020, pp. 344–350.
[43] N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, R. Silva, The sensitivity of counterfactual fairness to unmeasured confounding, in: A. Globerson, R. Silva (Eds.), Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019, Vol. 115 of Proceedings of Machine Learning Research, AUAI Press, 2019, pp. 616–626.
[44] A. White, A. S. d'Avila Garcez, Measurable counterfactual local explanations for any classifier, in: G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (Eds.), ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020, Vol. 325 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2020, pp. 2529–2535.
[45] A. R. Akula, S. Todorovic, J. Y. Chai, S. Zhu, Natural language interaction with explainable AI models, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 87–90.
[46] O. Espino, R. M. J. Byrne, The suppression of inferences from counterfactual conditionals, Cogn. Sci. 44 (4).