Towards a Formalization of Explanations for Robots' Actions and Beliefs[1]

Felix Lindner
Institute of Artificial Intelligence, Ulm University

Abstract. A robot's capacity to self-explain its behavior is a means of ensuring trust. This work presents a preliminary formal characterization of explanations embracing the distinction between explanations based on counterfactuality and those based on regularity. It also distinguishes generative and instrumental explanations. The formalization will guide future work on explanation generation, explanation sharing, and explanation understanding in human-robot interaction.

Keywords. Explainable AI, Ontology, Human-Robot Interaction

1. Introduction

Robots are complex systems composed of software components for perception, motion control, and decision making. Humans often fail to correctly interpret the behavior of robotic systems, cf., [1]. This lack of interpretability of robot behavior may lead to a loss of trust and thus hinders acceptance and adoption of new technology.

One way of mitigating trust loss are explanations [1,2]: Robotic systems should be able to self-explain their beliefs and behavior and to self-introduce their capabilities. Current approaches to explanation generation in robotics are often tailored to specific AI components. For example, explanation of kinematic capabilities are explained by sensoric and motor components [3], explanation of image classifications are explained by highlighting important portions of the image [4], explanations of actions are explained by their contribution to maximizing some metric [5,6] or their contribution to the achievement of specific goals [7], and a robot's ethical judgments can be explained by morally relevant aspects of an ethical problem [8].

The emerging field of explainable AI is developing fast and generates a highly diverse landscape of solutions. A strongly required research agenda consists in conceptual work on systematizing and unifying the concept of explanation, and eventually coming up with a format for formally representing individual explanations for the sake of communication between systems and systems, as well as systems and humans.

For the sake of this paper, I will focus on explanations for *why* a robot executes an action and for *why* a robot maintains some belief. I anticipate that a user of a robotic system will be interested in knowing why a robot exhibits some observable behavior. The explanation most likely will consist in statements of belief, such as "because the

robot believes doing so is morally permissible", "because the robot believes this is the shortest plan", or "because the robot believes you want something to drink", and the user will sometimes be interested in explanations for why the robot believes in these beliefs. Generally, there are also explanations for *how* something works. Explanations as a response to how questions are relevant to human-robot interaction, as well: When the robot explains itself can bring about some state of affairs, the user might ask for an explanation of how the robot is going to achieve that. For the time being, however, I will only make an attempt to formally characterize explanations for why questions.

2. Background and Related Work

There is a significant body of research on explanations in various disciplines as diverse as philosophy of science, psychology, linguistics, and artificial intelligence. Given limited space, I will concentrate on recent work on explanation in AI.

Research on ontology modeling is interested in supporting modelers by automatically providing explanations for why the modeled ontology entails a formula (the explanandum). The solution amounts to identifying a minimal subset (the explanans) of the set of axioms in the ontology such that the explanans entails the explanandum [9]. The explanans is also called a *justification*. The concept of a *repair* is dual to justification and denotes subsets of the ontology to be removed such that the explanandum is not entailed by the altered ontology. Hence, justifications are *sufficient* conditions for the explanandum and repairs are *necessary* ones.

In the explainable AI planning community [5,7], the idea that goals and performance metrics are explanations of actions is prevalent: According to this view, explaining why an action (explanandum) was executed amounts to identifying the goals (explanans) the action contributed to. Alternatively, also performance metrics under which the action was preferred, or prior beliefs that necessitate the execution of the action can be explanans. In machine ethics, one is interested in moral reasons for actions [8]. Here the goal is to identify aspects of a situation that are necessary or sufficient conditions for an action to be rendered permissible by an ethical principle. In machine learning, explanation generation methods are designed to explain black-box classifiers. These approaches generate sets of features that were necessary [10] or sufficient [11] for the observed classification.

There already exists a proposal for a formal characterization of explanation by Tiddi and colleagues [12]. The authors present an ontology design pattern and show how it is instantiated by various explanation concepts in different research fields such as psychology, philosophy, and computer science. The ontological formalization is in large part similar to the one I propose. A difference is that my formalization stresses the contrastive nature of explanation [13,14]. Based on contrastivity, I define the distinction between explanations based on counterfactuality and those based on regularity. Moreover, by connecting explanation to causality, the distinction between generative and instrumental explanations will be defined.

3. Formalization

The following axioms are meant to characterize the concept of explanation and to define four sub-concepts: counterfactual explanations, regularity explanations, generative

explanations, and instrumental explanations. These distinctions result from an attempt to classify current work on explainable AI. Determining the empirical significance of these distinctions to human-robot interaction remains an open question.

3.1. Explanandum and Explanans

An explanation has an explanandum (*Exum*) and an explanans (*Exans*) (1). The explanandum is the entity to be explained, and the explanans is the entity that is meant to explain the explanandum. The entities that play the role of explanandum and explanans play this role in the context of an explanation, Axioms 2 and 3.

$$\forall x(Ex(x) \rightarrow \exists y,z(Exum(x,y) \wedge Exans(x,z))) \tag{1}$$

$$\forall x,y(Exum(x,y) \rightarrow Ex(x)) \tag{2}$$

$$\forall x,y(Exans(x,y) \rightarrow Ex(x)) \tag{3}$$

The axioms leave open which kinds of entities play the role of explanandum and explanans. In fact, the fillers may be as diverse as events, facts, situations, theories, phenomena, or objects.

3.2. Origin and Contrast

The explanans is part of a situation, which is called here *the origin of the explanation* (Axioms 4 and 5). For instance, if the robot explains "I am swerving left because people are standing ahead of me", the origin is the situation made of everything that is currently perceived by the robot including the people standing ahead of. When explaining the output of a neural network classifier, the origin is the whole input vector fed into the neural network. The origin *gives rise to* (*Grt*) the explanandum (Axiom 6). The gives-rise-to relation is three-placed relating the explanation, the origin, and the explanans. The reason is that in principle the origin can give rise to many different things. Consider the case when the origin is the mereological sum of predicates describing a person. In the context of a loan classifier, this description may give rise to the decision that the loan is rejected. In the context of a personality classifier, the same origin may give rise to the decision that the person is agreeable. In a given explanation context, only a fraction of these relations is usually relevant (in most cases only one).

$$\forall x,y(Exans(x,y) \rightarrow \exists z(P(y,z) \wedge Origin(x,z))) \tag{4}$$

$$\forall x,y(Origin(x,y) \rightarrow Ex(x)) \tag{5}$$

$$\forall x,y,z((Origin(x,y) \wedge Exum(x,z)) \rightarrow Grt(x,y,z)) \tag{6}$$

In line with Lipton [13] and Miller [14], explanations are taken to be contrastive. If the autonomous car explains "I turn left because this is the fastest route", then this implies the existence of a contrast situation. The contrast may either be a situation in which turning left is not optimal ("If it were not optimal, I would not turn left.") or another situation in which turning left is also optimal ("Whenever turning left is optimal, I turn left."), see Section 3.4 for more on this distinction. The contrast relation is also

three-placed. This is necessary because the contrast must—besides being a plausible contrast to the origin—respect the context of the explanation. Consider how to respond to the question "Why does the robot go into the kitchen?" The contrast could either be a situation which gives rise to "Lisa goes into the kitchen" or another situation which gives rise to "the robot goes into the dining room." Which one of these options is a valid contrast depends on the nature of the explanation, viz., if it explains why the robot rather than Lisa goes into the kitchen, or why the robot goes into the kitchen rather than into the dining room.

Axiom 7 asserts the existence of a contrast, and Axiom 8 ensures that contrasts are relative to explanations and relate two different entities.

$$\forall x, y(Origin(x,y) \rightarrow \exists z(Contrast(x,y,z))) \tag{7}$$

$$\forall x, y, z(Contrast(x,y,z) \rightarrow (Ex(x) \wedge y \neq z)) \tag{8}$$

3.3. Regularity and Counterfactuality

Hume [15] gives two definitions of causality when he writes: *[. . .] we may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed.* The first part of the quote requires a relation of regularity between the cause and the effect, while the second part of the quote requires a relation of counterfactuality. Both these views on the relationship between cause and effect can also be observed to apply to explanation. One example from the literature on black-box explanations in AI is the difference between explanations based on anchors [11] and those based on counterfactuals [10]. Anchors are *sufficient* sets of predicates that give rise to the explanandum, while counterfactual explanations highlight *necessary* features. As stated above, in the research area of ontology debugging, there is the distinction between *justifications* and *repairs* [9]. Justifications are minimal subsets of the ontology that are individually sufficient to entail the explanandum, while repairs are minimal subsets of the ontology that are necessary for the entailment of the explanandum.

As for explanations based on necessity, Axiom 9 defines counterfactual explanations. An explanation counts as a counterfactual explanation if and only if the origin and the contrast give rise to different outcomes. Consider again the case of the robot swerving left because it has perceived people ahead. A counterfactual explanation could be framed like this: "I swerve left because there are people ahead. If there were no people ahead, I would move straight." The origin of this explanation is the situation where people are ahead, and this situation gives rise to swerving left. The contrast is a situation where the people are not present (or standing somewhere else), and this situation gives rise to moving straight. The explanans is thus a necessary condition for the explanandum.

Next, one can ask what is the difference between the origin and the contrast. To denote this difference, a function *Minus* is used that computes the mereological difference (cf., [16]) between the origin and the contrast. For counterfactual explanations, the explanans must be part of this difference, as expressed by Axiom 10. In the robot example, the origin may be the mereological sum of people standing ahead of and a plant being on the left. The contrast may be just the fact that a plant is on the left. The mereological difference between origin and contrast is the fact that people are ahead, and this is

already the explanans. Usually, one will require that contrasts are minimal perturbations of origins. However, this minimality constraint is not formalized here.

$$\forall x(CEx(x) \leftrightarrow \exists y,z,y',z'(Origin(x,y) \wedge Contrast(x,y,z) \wedge Grt(x,y,y') \wedge Grt(x,z,z') \wedge y' \neq z')) \tag{9}$$

$$\forall x,y,y',z,z'((Origin(x,y) \wedge Contrast(x,y,z) \wedge Exans(x,x') \wedge Grt(x,y,y') \wedge Grt(x,z,z') \wedge y' \neq z')) \rightarrow P(x',Minus(y,z)) \tag{10}$$

If one wants to explain why one does or believes something in some situation, one can also refer to another situation which is stable with respect to the explanans and state that one does or believes the same. Instead of giving a counterfactual explanation to explain swerving left, the robot from the swerving example could explain its behavior by citing a regularity: "I swerve left because I have perceived people ahead. Also in other situations when people are ahead, I use to swerve left." This time, the origin and the contrast give rise to the same behavior, which is seen as definitory for a *regularity explanation* as defined by Axiom 11. In the case of regularity explanations, the explanans is part of the mereological product [16] of the origin and the contrast, see Axiom 12. In the robot example, the origin again is the situation that there are people ahead and there is a plant to the left. The contrast situation might be the sum of the facts that there are people ahead, a plant to the left, and that the sun is shining. The product then is the sum of the facts that there are people ahead and a plant to the left, and clearly, the explanans is part of this product.

$$\forall x(REx(x) \leftrightarrow \exists y,z,y',z'(Origin(x,y) \wedge Contrast(x,y,z) \wedge Grt(x,y,y') \wedge Grt(x,z,z') \wedge y' = z')) \tag{11}$$

$$\forall x,x',y,y',z,z'((Origin(x,y) \wedge Contrast(x,y,z) \wedge Exans(x,x') \wedge Grt(x,y,y') \wedge Grt(x,z,z') \wedge y' = z')) \rightarrow P(x',Product(y,z)) \tag{12}$$

Nothing prevents an explanation from being counterfactual and regular at the same time. Indeed, often very good candidates for explanans are both necessary and sufficient [8]. In this case, an origin has two contrasts one of which gives rise to something different and a second one giving rise to the same.

3.4. Generativity, Instrumentality, and Causality

Besides the distinction between counterfactual and regularity explanations, there is another distinction to observe in explainable AI literature: I call it the distinction between *generative explanations* and *instrumental explanations*.

Again consider the robot swerving left. The generative type of explanation may read "I swerve left because I perceive people ahead. If there were no people ahead, I would not swerve left." In this case, the explanans "I perceive people ahead" causes the explanandum "I swerve left". Contrast this to the explanation "I swerve left because this way I avoid bumping into the people. If I did not swerve left, then I would bump into the group of people." This time the explanandum "I swerve left" causes the explanans "I avoid bumping into the people". Definitions 13 and 14 take the direction of causality to

be definitory for the distinction between generative explanations (GEx) and instrumental explanations (IEx). One could also say that in the case of generative explanations, the explanans grounds the explanandum, i.e., the explanandum holds *in virtue of* the explanans. Conversely, in the case of instrumental explanations, the explanandum holds *for the purpose of* the explanans. Moreover, in the case of instrumental explanations, it is required—via Axiom 15—that the origin gives rise to something which is preferred over what is given rise to by the contrast. In the robot navigation example, avoiding the people is preferred over bumping into the people.

$$\forall x GEx(x) \leftrightarrow \exists y,z(Exum(x,y) \wedge Exans(x,z) \wedge Causes(z,y)) \tag{13}$$

$$\forall x IEx(x) \leftrightarrow \exists y,z(Exum(x,y) \wedge Exans(x,z) \wedge Causes(y,z)) \tag{14}$$

$$\forall x(IEx(x) \rightarrow \exists y,y',z,z'(Origin(x,z) \wedge Contrast(x,y,z) \wedge Grt(x,y,y') \wedge Grt(x,z,z') \wedge z' > y')) \tag{15}$$

As the robot swerving example shows both types of explanations are possible when a robot seeks to explain its actions. Both types of explanations are also possible when a robot seeks to explain its beliefs: One can slightly change the above why-question: Instead of asking why the robot swerves left, one asks why the robot believes that it must swerve left. The above explanations still work. However, this might be the case just because this belief essentially is about an action. Consider a robot believing in the afterlife. A generative explanation of this belief may be given by other beliefs, e.g., the robot may say it believes in everything written in the bible. An instrumental explanation, on the other hand, would cite some value this belief has to the robot. For instance, it could say that believing in the afterlife makes itself more optimistic overall, or that this belief helps the robot to console those who have lost loved ones. Even if one hesitates to ascribe beliefs of this kind to robots, a robot still should have the capability to represent and reason about such beliefs because humans the robot will interact with are likely to give explanations of these kinds. Moreover, in application areas where a robot simulates human reasoning to adapt its behavior, it can be of value to simulate such beliefs.

4. Conclusions

There has already been quite some work on generating explanations in AI. Much less work has been done on representing and communicating explanations—neither between systems and humans nor between different components within a system. The preliminary formalization presented in this paper is meant as a first step towards filling this gap. A conceptual distinction between four types of explanations, viz., counterfactual, regular, generative, and instrumental, has been defined. Future work will target at deepening the analysis of explanation and linking it to formalizations of causality and preferences. Having a richer formal theory of explanation will also guide the development of algorithmic solutions to explanation generation, explanation fusion, explanation verbalization, and explanation understanding. These problems are considered subjects of future work with practical significance to human-robot interaction, viz., robots representing and exchanging explanations for actions and beliefs with humans, and also robots understanding and reasoning about humans' explanations.

References

[1] Tolmeijer S, Weiss A, Hanheide M, Lindner F, Powers TM, Dixon C, Tielman ML. Taxonomy of trust-relevant failures and mitigation strategies. In Proceedings of HRI'20, 2020.

[2] de Graaf M, Malle B. How people explain action (and autonomous intelligent systems should too). AAAI Fall Symposium Series, 2017.

[3] Kunze L, Roehm T, Beetz M. Towards semantic robot description languages. In Proceedings of 2011 IEEE International Conference on Robotics and Automation, 2011, pp. 5589–5595.

[4] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[5] Fox M, Long D, Magazzeni D. Explainable planning. In IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), 2017.

[6] Krarup B, Krivic S, Lindner F, Long D. Towards contrastive explanations for comparing the ethics of plans. In ICRA 2020 Workshop Against Robot Dystopias: Thinking through the ethical, legal and societal issues of robotics and automation (AGAINST-20), 2020.

[7] Broekens J, Harbers M, Hindriks K, Van Den Bosch K, Jonker C, Meyer JJ. Do you get it? User-evaluated explainable BDI agents. In German Conference on Multiagent System Technologies, 2010, pp. 28–39.

[8] Lindner F, Möllney K. Extracting reasons for moral judgments under various ethical principles. In KI 2019: Advances in Artificial Intelligence, 2019, pp. 216–229.

[9] Horridge M. Justification based explanation in ontologies. Ph.D. thesis, University of Manchester, UK (2011)

[10] Sandra W, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology 1(2), 2018.

[11] Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In AAAI Conference on Artificial Intelligence, 2018.

[12] Tiddi I, D'Aquin M, Motta E. An ontology design pattern to define explanations. In K-CAP 2015: Proceedings of the 8th International Conference on Knowledge Capture, 2015, pp. 1–8.

[13] Lipton P. Inference to the best explanation. London and New York: Routledge (1991)

[14] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 2019.

[15] Hume D. An Enquiry Concerning Human Understanding (1748)

[16] Varzi A. Mereology. Stanford Encyclopedia of Philosophy, 2016.