Explaining In Time

Meeting Interactive Standards of Explanation for Robotic Systems

THOMAS ARNOLD, Tufts University, USA
DANIEL KASENBERG, Tufts University, USA
MATTHIAS SCHEUTZ, Tufts University, USA

Explainability has emerged as a critical AI research objective, but the breadth of proposed methods and application domains suggest that criteria for explanation vary greatly. In particular, what counts as a good explanation, and what kinds of explanation are computationally feasible, has become trickier in light of oqaque "black box" systems such as deep neural networks. Explanation in such cases has drifted from what many philosophers stipulated as having to involve deductive and causal principles to mere "interpretation," which approximates what happened in the target system to varying degrees. However, such post-hoc constructed rationalizations are highly problematic for social robots that operate interactively in spaces shared with humans. For in such social contexts, explanations of behavior, and, in particular, justifications for violations of expected behavior, should make reference to socially accepted principles and norms.

In this paper we show how a social robot's actions can face explanatory demands for how it came to act on its decision, what goals, tasks, or purposes its design had those actions pursue, and what norms or social constraints the system recognizes in the course of its action. As a result, we argue that explanations for social robots will need to be accurate representations of the system's operation along causal, purposive, and justificatory lines. These explanations will need to generate appropriate references to principles and norms – explanations based on mere "interpretability" will ultimately fail to connect the robot's behaviors to its appropriate determinants. We then lay out the foundations for a cognitive robotic architecture for HRI, together with particular component algorithms, for generating explanations and engaging in justificatory dialogues with human interactants. Such explanations track the robot's *actual* decision-making and behavior, which themselves are determined by normative principles the robot can describe and use for justifications.

CCS Concepts: • **Human-centered computing → Interaction design**;

Additional Key Words and Phrases: explainability, normative HRI, architectural requirements

ACM Reference Format:

Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2010. Explaining In Time: Meeting Interactive Standards of Explanation for Robotic Systems. *ACM Comput. Entertain.* 9, 4, Article 39 (March 2010), 21 pages. https://doi.org/0000001.0000001

1 INTRODUCTION AND MOTIVATION

Explainability has emerged as a critical AI research objective, but the breadth of proposed methods and application domains suggest that criteria for what ought to count as a proper explanation vary greatly. While philosophers of science

Authors' addresses: Thomas Arnold, Tufts University, 200 Boston Ave. Medford, MA, 02155, USA, thomas.arnold@tufts.edu; Daniel Kasenberg, Tufts University, 200 Boston Ave. Medford, MA, 02155, USA, daniel.kasenberg@tufts.edu; Matthias Scheutz, Tufts University, 200 Boston Ave. Medford, MA, 02155, USA, matthias.scheutz@tufts.edu.

Manuscript submitted to ACM

have attempted to establish criteria for explanation that are universal across explanatory domains [Woodward 2017], from deductive-nomologial models (e.g., [Hempel and Oppenheim 1948]), to statistical relevance models (e.g., [Salmon 1971]), to causal mechanical models (e.g., [Salmon 1984]), to unification models (e.g., [Kitcher 1981]), there seems to be evidence for multiple models of explanation in human psychology: "Explanation sometimes engages deductive reasoning and theory-like representations in human psychology, explanation often recruits inductive reasoning and information about causal mechanisms, too. Each explication highlights constraints on what is being explained that correspond to distinct processes and representations in human psychology" [Colombo 2017]. Yet, recent attempts in AI to stipulate what an explanation is deviate significantly from all of the above approaches, neither following logico-statistical models in philosophy of science, nor directly addressing human explanatory demands researched in psychology. Rather, these attempts to stipulate what "explanation" of AI systems ought to be are driven, paradoxically, by the lack of explanation of what it is these systems are doing. Systems like deep neural networks that are trained on data and show high levels of performance on tasks provide little if any insights as to how they are achieved. To get at what such trained "black-box" systems internalized, the notion of "interpretability" functions as a mediator between actual system behavior and a human-understandable approximation thereof, at the risk of erring in cases where the "interpretation" fails to capture the true causality in the system (the latter serving as a complementary criterion of "completeness" [Gilpin et al. 2018]).

While some have argued that black-box systems need to be abandoned in favor of systems that can yield more straightforward explanations of their decisions (e.g., ones based on decision-trees) [Rudin 2019], the field of "explainable AI" is still grappling with laying down the criteria for what ought to count as a good explanation and when certain systems should meet higher standards. AlphaZero, for example, is not designed to explain how to win at chess but just to win at chess [Silver et al. 2017]. Its moves are no less successful for not being explained or made explicit from the design end, and they can be appreciated and analyzed without harm to those who do so. Nonetheless, standards of explanation should not be obfuscated or misrepresented out of accommodation to opaque approaches. Hedging on explanation can resemble the proverbial drunk person searching for lost keys, looking not where the keys were dropped but under a far away lamppost because "that's where the light is."

In the case of robotic systems, especially those that operate interactively in shared spaces with humans, losing track of explanations can be especially fraught. Instead of adjusting "explanation" to cater to trending approaches, we propose to 1) clarify what explaining a robotic system should constitute in context and 2) pursue algorithms and architectural designs that meet those standards. Consequently, we argue that explanations for social robots will need to be accurate representations of the system's *actual* operation and objectives, not *interpreted* representations to buffer the observer's lack of understanding of the system's operation. This means that explanation will often need to address causal, purposive, and justificatory aspects. A social robot's actions can face explanatory demands for how it came to act on its decision, what goals, tasks, or purposes its algorithms had those actions pursue, and what norms or social constraints the system factors in the course of its action.

Part of getting clearer about explanation for robotic systems means anticipating the real world conditions in which an interactive system will have its actions explained. Designing robots to be responsibly explained will require considerations of when and on what terms an explanation is provided. In this paper we devote particular attention to the temporal constraints that lend extra pressure to demands on explanations, and we lay out some preliminary technical approaches to achieve responsive and accurate explanations for real-time interactions. Explaining in time, in real time, is not just an additional design burden for AI. It shows why causal, purposive, and justificatory roles for explaining robotic action are so crucial.

2 BACKGROUND

Explainable AI (XAI) has developed as a research area in large part due to the challenge of understanding how solutions obtained by data-driven machine learning approaches like deep learning, which are increasingly deployed in robotics domains, work. Because the black-box aspect of such systems does not square with everyday and scientific notions of explanation, the proposal is to at least make the operation of such systems more *interpretable* and subject to statistical evaluation [Samek 2019]. Given the many areas in which deep learning applications might feature in critical decisions, from embodied contexts with self-driving cars to disembodied contexts in law and medicine, the notion of "explainability" in AI systems has become a "mediating category": it is clear that explanations are important, but the various models proposed still have difficulty meeting the demands for completeness and inferential accuracy [Došilović et al. 2018; Gilpin et al. 2018; Xu et al. 2019].

This methodological limbo has led to increasing pushback, with XAI being likened to alchemy, or an oracular "42" [Goebel et al. 2018]. The call for "rebooting AI" cites this kind of deficiency in explicit, logical inferences and causal understandings on the part of deep learning [Marcus and Davis 2019]. Rudin's is just one recommendation to steer clear of deep learning in favor of systems that show how a decision is made [Rudin 2019] . The "logicist" approach of Bringsjord aims for explicit, inferential tracking of ethically related decisions, rather than post-hoc analyses of what may or may have not affected a system's output [Bringsjord et al. 2006].

A notable overview by Miller et al. [Miller 2018] argues that social science research reveals many more sides to explanation than what has made its way into so-called XAI. Mittlestadt et al. [Mittelstadt et al. 2019] have presented this work as a basis from which policy makers can better lead discussions of AI accountability, transparency, and interpretability. Still, the search for direct explanation for purposes of social accountability has accommodated the basic machine learning approaches more than question them at the root of design. The "data sheets" approach of Gebru et al. [Gebru et al. 2018] and related "model card" approach of Mitchell et al. [Mitchell et al. 2019] concentrate mostly on the provenance of data and the performance outcomes of systems, respectively. They do not scrutinize the internal architecture of the ML system itself as an object for design methodology. This is what leads Robbins to call the idea of "explicability," in terms that apply quite well to explainability, a "misdirected principle" for AI: if in a given domain the outcomes of a black-box system are unacceptable without a factual, direct explanation, then black-box systems should not be used for that domain [Robbins 2019]. The whole point of black-box systems, Robbins argues, lies in us not having preset considerations as criteria for their pattern discernment. If such criteria are too important to lose (the prevention of racial bias in processing loan applications, for example), then black-box systems should not decide loan applications. Kroll et al. push back against the idea that algorithms are out of reach of accessible explanations at a broader level (and hence not amenable to testing for measures of fairness) [Kroll et al. 2016], but this leaves open on how broad an explanation can be and still offer enough of a grasp on discrete outputs and actions.

The social science review that Miller provides stresses that explainability in AI has struggled both because of uninterpretable models and because social interaction for AI has receded into the background. To reassert the social and practical dimensions of explanation, Miller forwards four characteristics of explanation that XAI might better incorporate: 1) questions why something happened have *implicit contrasts* (i.e., why something else did not happen), 2) explanations will be *selective*, carving out causal stories according to various biases, 3) effective explanations are generally *not statistical*, and 4) explanation are *social* (i.e., they represent an explainer conveying understanding to an explainee, in part based on idea of what the explainee needs to understand). If AI systems are to be explainable, according to Miller, these dimensions need to be integrated into how contextual explanation is, and where a given explainer and explainee are situated in a particular environment will make a difference in what causal accounts suffice.

For robotic systems, the turn toward social context is nothing new. Miller's overview hardly scratches the surface of what embodied systems may face socially. It is notable that human-robot interaction scholarship already grasps explainability as not just a theoretical evaluation but a practical feature of joint human-robot action. Explanations, from that vantage, can be a means toward improving human-robot performance. It has been couched, for example, as "plan reconciliation," wherein human-robot teams might share an understanding through explanation that enables and presupposes future collaboration [Chakraborti et al. 2019]. Hayes et al. see the value of explanation in achieving robot controller transparency [Hayes and Shah 2017]. When put into context of broader, more open-ended social interaction, however, explanation will involve more than joint work, and will demand explanation to be more than a means of improved task efficiency. In other contexts and use cases, the right explanation may be an end in itself for deeming a robot appropriate, effective, and useful. The ability, need, and decision to explain its actions might just be what renders a robot normative. And indeed, as we shall discuss more below, robotic action, unexplained, might be judged for violating certain social and moral norms [Malle 2016]. Models that are more logic-based and symbolic, in line with the work in expert systems that originally launched "explainability" as a goal, provide more explicit features to an architecture to track how a system will arrive at its decision [Goebel et al. 2018]. Giving accurate and detailed explanations thus converges with calls for more explicit reasoning on the part of AI systems for why they do what they do [Pearl and Mackenzie 2018]. To get a better purchase on how design and explanation should relate for robotic systems, it is worth exploring what explanatory demands might weigh on robotics systems. In particular, given the issues of transparency and accountability to which explanations are often tied, it is worth separating "explainable" as a general, diffuse value of a system to the standards of an actual explanation that is being asked of a system. It is worth getting clear, in other words, whether it is better to acknowledge explanation's implicit standard in context rather than adjust what counts as explainable.

3 THREE CRITERIA OF EXPLANATION FOR INTERACTIVE ROBOIC SYSTEMS

3.1 Causality

Explanations in practice cannot encompass every causal factor in the world that led to an agent's action. Some causes are more relevant than others to single out, depending on the interests of those analyzing the system. Still, accuracy about the causes that an explaination singles out is still paramount. When it comes to explaining why an embodied system took the action it did, an explanation will be an orchestration of both (1) a causal account that features an explicit description of the processes that occurred through the system's own design, and (2) a representation of the context and world on the basis of which the system acted as it did. It is important to stress here that causality, including basic counterfactuals that elucidate the system's decision-making, can serve as the basis for people's subsequent actions based on the ability to count on the system to do what it is meant to do.

A possible illustration of what happens when there is no explanatory account of robotic action occurred recently near Los Angeles. A robot that was intended to patrol a public boardwalk and advise residents not to litter was, in the wake of a fight between two people nearby, sought after as an emergency connection to police [Flaherty 2019]. Not only did the buttons a distressed bystander pushed not call the police, but the robot proceeded on its trash-fighting way with no further engagement of the distressed onlooker.

One might say that in this case such a robot should have even less interactive capability than its buttons showed. Still, if robots like this are to operate interactively, one can imagine different explanations as being properly contrastive relative to a bystander's expectation. If a person had an idea the robot could respond to spoken questions, they might ask "Why won't you call the police?" One explanation it could offer, even before being asked, would be a statement of what its general role is and what tasks it is designed to perform. It might also articulate rules or regulations that its operation is designed to follow – perhaps how it is not supposed to leave the sidewalk to go on sand. As for counterfactual aspects of a contrastive explanation, a robot could explain that only an authorized city official would be able to redirect its course. These types of explanation would have their practical limits, of course, just as they do not speak to every possible concept that might be aimed at them. But the accountability of the explanation, one can see, can hinge on how its different elements find correspondence in the actual decision-making process the system instantiates.

What an explanation of a system practically evokes, in other words, are conditional relations about what it does and why. As expert systems research has long appreciated, explanation ties into plans, hierarchies of tasks, and accurate understanding of how those are being managed by an agent. Representing tasks as falling under a principle or guideline, or as violating a guideline violating a guideline, is not just a scripted attempt to assuage or reassure one seeking an explanation – they are claims about how the system works, how it is designed to arrive upon its course of action.

The causal importance of explanations lies as much or more in failure and breakdowns as it does in detailing a system's successful executions. Without some notion of an explicit plan or concept that was being attempted, there is no way to parse a mishap as doing something mistakenly vs. operating with incorrect belief vs. just physically fumbling the task. Only with some degree of faithful representation can a system's execution be evaluated along the norms and practices with which people judge similar actions in the environment. Trying to pick something up and dropping it by accident is very different from tossing an object or dropping it out of a mistaken belief that a user said to do so. Explanation is enmeshed with interpretability but not co-extensive, since it is explanation that can identify what condition obtained or might have obtained: that is, when a counterfactual element is relevant (e.g., if my grippers had not failed I would have held the object).

From the philosophical literature on explanation we can see more clearly why statistical correlation or inference from a pattern is not the same as a causal account. Several philosophers have cited the example of a flagpole that, given a certain angle of light from the sun, casts a shadow of a certain length [Woodward 2017] . It is possible, if one knows the angle of the sun, and the length of the shadow, to infer the height of the flagpole (barring some interfering factors like an object that comes to rest on top of it, or a taller flagpole behind it whose shadow coincides with it perfectly). However, Salmon points out the asymmetry of causality between the flagpole and the shadow [Salmon 2006]. The flagpole explains the shadow by helping to create it, but the shadow does not create the flagpole or its height. This helps demonstrate that a correlation may yield certain predictable relationships without offering a causal account for why a system acts the way it does. Its performance may correlate with certain inputs, but the inputs do not create the design of the system. It is the architecture of the system that, like the flagpole, helps creates the output.

This is important when robotic system stands out in the real world as an agent to which people need to react, when its presence is more like a flagpole than a shadow. In simulated or virtual environments an AI system's performance might be cultivated through an overall utility function applied to that environment as a whole (e.g. in a video game where maximizing a score is the benchmark). The line between agent and environment is less relevant there than what unfolded in the environment as a whole. But when a robot travels down a street, there will be practical lines drawn by people there between what difference the robot's actions made vs.what would have happened on the street regardless of its movements (including the independent decision of people there). The causal question moves into the robot's decision-making, and what architectural flagpole cast the shadows of its action.

3.2 Purpose and Prospective Action

If a robot's action is to be explained as a response to its environment, it is also the case that its actions can and will be viewed prospectively. The question of what a robot did often relies on identifying what it was attempting to do, whether that be moving toward a sought-after space or manipulating an object usefully. Robotic motion in the course of working on tasks is, not surprisingly, where a good deal of HRI research into interpretability and planning concentrates. For ongoing action, though, a causal account carries increasingly prospective implications. Unless otherwise clarified, it carries implicit commitments to what will happen as the robot continues to operate in this context. Chakraborti et al., for example, have couched explanation as "plan reconciliation" in order for human-robot teams to share an understanding through explanation that enables and presupposes future collaboration [Chakraborti et al. 2019].

Some of this purposive explanation can be thought of as a learning process on the part of human interactants. For certain coordinated tasks, interpretability and predictability may be more practically important than explanation per se. It is consistent with being able to predict a system's action to have no explanation for how it operates internally so as to effect those actions, nor whether the objectives ascribed to a system can be located in the system's code. Again, that might not be the most important objective in the interaction. But it is worth pointing out that in those cases it is explanation itself that is less of a priority, rather than an alternative kind of explainability.

One distinctive demand for socially interactive robots will be to communicate explicitly what their plans or purposes are, not just have modeling or planning ascribed to their operation. DeepMind's recent MuZero effort, designed to master Atari games as well as Go and chess, generates a model for each game being learned instead of explicit rules being given to the system beforehand [Schrittwieser et al. 2019]. Nonetheless, this modeling takes place within the confines of the video game or game board in which the system operated, not the dynamic, open-ended environment of shared physical space with people. That means there is no need to communicate its modeling of the games it is playing to anyone else. The performance is beating the game (or the opponent), not explaining how to do so.

For systems whose actions need to be explained for ongoing coordination, an explanation must represent plans or purposes in a way that maps onto its internal operation. Successful prediction of a system's action may work within the confines of a single task, but an explanation of purpose situates that task relative to other tasks and contexts. It can provide orientation to how a system would act in other situations, shedding counterfactual light on why it performed the way it did (e.g. "The robot would look for the ball in the opposite corner had you put it there.")

3.3 Norms and Justification

When robotic systems pursue the primary goals or plans for which they are deployed, they can naturally do so in various social settings. Such settings will often rely on a terrain of social norms within which the system's actions need to operate, especially as they are likely to be blamed for not doing so [Malle and Scheutz 2014]. In cases where a system does have to transgress or diverge from an accepted norm, an added dimension of explanation emerges: *normativity*. What is selected for an explainee are relevant causes and counterfactual conditions to establish why an action occurred and possibly why it had to occur (given the normative constraints). In the case where a social and/or moral norm is involved, the burden is to isolate the relevant counterfactuals to address why a norm was violated or fulfilled in the way that it was. Explanations are not justifications, but explanation with no reference to norms risks undermining the understanding of a robotic system's very role. Explicating the implicit counterfactuals ascribed to an action can make the difference between viewing an outcome as a necessary sacrifice and a reckless maneuver. A robot whose power is low may stay motionless rather than risking hitting a child blocking its charging station. Without an explanation that references an obstacle, the robot's action would just be a failure to recharge.

Norms are also often commitments about task commitments, what constitutes the context of forbidden, obligatory, and permissible actions in the pursuit of particular objectives. Purposiveness involves the chief goal or intention for the robot's action, but the means by which the robot pursues a purpose may need to conform to norms.

Note that the conditions of an accurate, causal account still apply. The overarching explanation of a robot's actions (by itself or a designer or otherwise) as "helping people cross the street" may appeal to a general norm, but then it should be clear on what design terms that explanatory appeal is made. Is that a normative ascription of behavior to a system that a designer trained on model data for that purpose, or does the system itself have a representation of "helping" to which the operation can refer? To the degree norms are not genuine guides within a system's architecture, explanations involving norms (e.g., "I'm sorry if I've hurt you, I'm just trying to help") become crass and manipulative, since there is no necessary connection between the norm cited and how the system chooses to behave in light of it.

It is worth pausing here to note that explanations involving norms can themselves risk violating norms of trust and deception. Is it right to say a robot is "helping" when it has no real comprehension of the concept, or even a representation in its architecture of helping behavior or rules? While idiomatic phrases like "I'm sorry" may not pose a problem, it is less clear on what terms a robot's are explained legitimately by "helping" language, especially if the robot is doing the explaining to the people around it. What seems needed at least is a clear, accountable, and shared standard to which it could refer, for which its own architecture could account (e.g. responding to a request for an answer with "helping is not allowed while taking the test").

A normative explanation can sound quite similar to a justification, but justification is not the exact demand being made here. An explanation referencing norms is still primarily about accurately describing why an action occurred. The point is not to absolve the robot from blame or to ascribe responsibility to the robot. Justifications per se could invite the use of justificatory language in order to rationalize behavior, whether or not it was genuinely designed to hew to norms as much as possible or prioritize more important norms over others. A genuine norm-oriented explanation, contrary to a recent argument of van Wynsberghe, does not let designers off the hook or mean "moral" robots per se; rather, it functions to keep norms more explicit and traceable in the system's design [van Wynsberghe and Robbins 2019].

These layers do not present themselves in every robotic application. But in the name of "explainability" they need to be kept under consideration as part of various practical landscapes, possibly activated when a robot enters a dynamic context where task and role might face new expectations and reactions.

A recent incident in Pittsburgh showed the difference between having implicit plans via mapping rather than explicit, context-informed plans that could incorporate adjustments via norms. A student using a wheelchair was trapped on a street corner by a Starship delivery robot, who blocked the curb cut that the student needed to proceed [Wolfe 2019]. It was designed to wait there until the crossing light gives permission to cross, but for those coming the other way in a wheelchair that means being stuck on the street. The designers promised to change the mapping that the delivery robot used, so that it would not block a curb cut. More recently, the grocery store robot Marty has provoke complaints about making social distancing difficult during the Covid-19 crisis, taking up space that people could otherwise use to stay safe [Turmelle 2020]. With no concepts of norms like blocking others, or perhaps being asked to move, a system's violations of safety norms will have to be tackled in retrospective debriefing. For some interactants, that might prove too late to avoid harm.

4 PUTTING ROBOTIC EXPLANATION INTO PRACTICE: SHARED UNDERSTANDING IN TIME

The preceding three dimensions of explanation reflect general demands that a a robotic system, as embodied and physically situated amid people, will face in social interaction. As one turns to the technical efforts to enable such explanation (and perhaps showing where certain control architecture make it unfeasible), there are two practical characteristics that loom over efforts at explanation. First, an explanation of robotic action will need to take shape around the understanding of the one to whom an explanation is offered. Secondly, a robot itself may have to provide such an appropriate explanation in a time-sensitive way. This means, first, a mental modeling between explainer and explainee, taking into account what the explainer knows about the explainee in order to target the explanation at the right level of detail. Just as importantly, it also means a strong temporal dimension for the explanation: the robot may need to account for its interaction history (not just its current moment and its options), and it needs to offer explanations without taking too much time for the human explainee (e.g., by not focusing at the right level of detail or abstraction in its explanation). It is through an accurate and temporally-indexed communication to its audience and/or interlocutor that the causal, purposive, and normative elements of explanation have a chance of succeeding in context.

4.1 Mental Modeling Between Explainer and Explainee

An explainer must have some basis for shaping an explanation relative to what the explainee knows and what the explainee still does not understand. When someone attempted to push the boardwalk robot's emergency button, one mental model of the explainee is obviously that they misunderstand what the robot is equipped with. Again, an explainee may need to understand (1) a description of the world that the system perceived, and how that state of the world determined the decision acted upon, (2) why a certain action was designed to be executed by the robot at all, and (3) how the action fits into an overall plan or schema that others could factor into their own planning [Garcia et al. 2018]. Explanation is not just contextual in terms of an interactive system's scope of action and decision-making, but honest, effective explanation will need to adjust and address implied abilities and lack of ability as the human explainee conceives of them. Our work on mental modeling has explored computational mechanisms for mental models and shared mental models in different cases of individual and team-based human-robot communication [Gervits et al. 2018, 2020; Scheutz 2013; Scheutz et al. 2017].

4.2 Temporality and the Practical Constraints of Explanation

The explanatory challenge of robotic systems is difficult enough given their physicality, social interaction, and multiple roles in certain contexts. But the standards of causal, purposive, and normative explanation take on even more depth when put under the temporal strictures of various interactions. Put differently, it is only through temporal specification that those qualities of an explanation can be fully realized. The person on a street or in a hospital room cannot wait for a thorough audit of a system's architecture, nor an exhaustive overview of everything a system went through to reach the decision to act as it did. At the same time, those interactants deserve not to be deceived or manipulated by a mere rationalization or slogan that bears no connection to how the system is designed. These domains also pose distinctive problems for needing time-sensitive explanations of what they do. The demand on the Santa Clarita boardwalk robot was not to explain over some indefinite stretch of time why it was continuing past the person pushing it buttons; on the contrary, the person wanted immediate help from the police to help break up a fight. Providing relevant, accurate representations of what a robot is doing and why is also a time-bound demand on computing resources. On what terms and under what conditions should a system's operation be devoted to explanation itself, when the robot's overall task and priorities need to be completed in an efficient and timely manner? Explaining in time entails knowing when to take time to explain.

Ultimately the temporal practicality of robotic action will need to take shape according to specific contexts with different priorities of explanation. Complex navigation that depends on object-avoidance likely means a system does not need to explain each adjustment and turn as its being taken: just as people making their way down a hall may do so with implicit, unspoken coordination. Still, when a simple explanation in time can change how objectives or plans are pursued, the system may need to expend the time and energy needed to provide one.

When it comes to technical demands on explanation, time therefore represents several kinds of pressure. As mentioned, there is the pressure for timely explanation, which tests the computational efficiency by which a system will run. But time also lends contours to the causes, purposes, and norms that we have explored as criteria of explanations properly wedded to context. A robot performing an action more quickly than usual may be based a norm of timeliness itself, and its movements explained by its decision-making incorporating an avoiding of wasting a person's time (who may be waiting or inconvenienced by a task not being completed as usual). Likewise, an explicit purpose or goal for robotic action may take on an unreasonable or inconsiderate aspect if its relation to time (what has to get done after it, how long it could take) is left unaccounted for. As we have discussed, a mere appeal to such considerations (say, through a retrospective interpretation of its behavior) is not enough to be a genuine explanation– it must map accurately onto how the system arrives at its path of action at the time it takes it.

In what follows we present some concrete technical steps toward meeting the temporal dimensions of robotic action in human-robot interaction. These are, to be clear, steps down a much wider and longer path of research than covered here. The computational efficiency of the approach covered, for instance, still leaves considerable room for improvement to meet real-world temporal demands. And there are dimensions of causation, purposefulness, and justification that extend well beyond what are specified and addressed in the following technical scenarios. Still, it is critical for human-robot interaction that technical contributions around "explainability" stay focused and grounded enough not to overpromise and sweep aside complications of context. Limited, substantive steps can offer a more solid basis for technical progress than abandoning standards of explanatory accountability.

5 TECHNICAL APPROACHES FOR EXPLAINING IN TIME

Facing up to time constraints and time-oriented expectations on the part of interactants heightens the design challenge of explainable robotic systems. Given the layers of explanation we have discussed, it is all the more important that explanations manage to give access both to the system's abilities and its limitations, all in a timely manner. Inaccurate attributions of perceptual and/or symbolic abilities may warp what a person should accept or rely upon from the system, and so even in cases of practical complexity the system should not overpromise in what it can explain about its actions. The upshot of what follows as technical proposals is that the more explicit a system can be both about its internal operation and its understanding of the objects and tasks with which it deals, the more plausible it is that it could meet those standards successfully. Here we outline some specific technical directions toward addressing temporal dimensions to explaining a robot's action. In particular, we show how a set of explicit plans can function as a guides for action, while taking into account time considerations and multiple priorities for how those plans are pursued. Moreover, the scenario presented includes representations of norms that can determine what is permissible to perform given certain task objectives.

The technical work to address temporality in explanation are offered as stepping stones towards fuller forms of explanation. While these steps may not take us as far as we should ultimately should demand in practice, they help tp ensure a more genuine direction for explaining robotic systems in context. By taking on temporality as critical from the start, this approach can take on more of the the causal, purposive, and normative functions that legitimate explanations involve.

Throughout this section, we shall use the *ShopWorld* domain as a running example. In the ShopWorld domain, a robot is tasked with obtaining items from a shop for its owner. For simplicity's sake, we will assume that the shop contains two objects for sale (although nothing hinges on that for the generality of the presented algorithms): a pair of glasses ($glasses$) and a watch ($watch$). While in the shop, the robot may pick up ($pickUp(x)$), put down ($putDown(x)$) and buy ($buy(x)$) objects one at a time, or may leave the store ($leave$). Each object has a particular cost, and for the purposes of this example we shall assume that the robot has sufficient money for either, but not both, of the items for sale. We shall also assume state predicates indicating whether the robot is holding an item ($holding(x)$), whether an item is on the shelf ($onShelf(x)$), whether an item has previously been purchased ($bought(x)$) and whether the robot has left the store ($leftStore$).

5.1 Temporal logic, interpretability, and representing priorities

Our technical approach is to employ a temporal logic to explicitly specify robot objectives as well as safety constraints and moral and social norms. We start by defining the formal language called *Violation Enumeration Language* (VEL) for specifying temporally extended objectives compatible with Relational Markov Decision Processes (RMDPs). Specifically, VEL is an extension of linear temporal logic (LTL) [Pnueli 1977], a propositional logic augmented with the temporal operators **X**, **G**, **F**, and **U**. Here $\mathbf{X}\phi$ means "in the next time step, ϕ"; $\mathbf{G}\phi$ means "in all present and future time steps, ϕ"; $\mathbf{F}\phi$ means "in some present or future time step, ϕ; and $\phi_1 \mathbf{U} \phi_2$ means "ϕ_1 will be true until ϕ_2 is true (and ϕ_2 will eventually be true)". VEL extends LTL with the following modifications:

- The set of atomic propositions has been changed to a set P of predicates, where each predicate p_i has arity α_i.
- VEL supports existential and universal quantification over variables, though only at the outermost level of a formula.
- VEL supports specifying *enumerated* variables. Enumerated variables are similar to universally quantified variables (and so are specified by the similar symbol $\mathbb{A}$). However, the extent to which a given trajectory satisfies a VEL statement depends on *the number of bindings* of the enumerated variables for which the statement is satisfied/violated (this is not true of quantified variables).

The grammar for VEL is as follows, with ϵ the empty string:

$$\begin{aligned}
\phi &::= \psi \mid \mathbb{A}(\langle Var\rangle, \cdots, \langle Var\rangle).\psi \\
\psi &::= \varphi \mid \forall\langle Var\rangle.\psi \mid \exists\langle Var\rangle.\psi \\
\varphi &::= \langle Atom\rangle \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid \mathbf{X}\varphi \mid \mathbf{G}\varphi \mid \mathbf{F}\varphi \mid \varphi \mathbf{U} \varphi \\
\langle Atom\rangle &::= \langle Pred\rangle \mid \langle Pred\rangle(\langle Var\rangle, \cdots, \langle Var\rangle) \\
\langle Pred\rangle &::= \text{Any alphanumeric string} \\
\langle Var\rangle &::= \text{Any alphanumeric string that is not a predicate name}
\end{aligned}$$

In the ShopWorld example, we shall assume the robot as having two specifications: "leave the store while holding as many objects as possible", and "do not shoplift" (leave the store while holding an unpurchased item). They are respectively expressed in VEL as follows:

$$\forall x.\mathbf{F}(holding(x) \wedge leftStore) \tag{1}$$

$$\forall x.\mathbf{G}\neg(holding(x) \wedge \neg bought(x) \wedge leftStore) \tag{2}$$

The use of temporal logic for specifying robot objectives has the following advantages (over, for example, reward-based approaches):

- *(Ordinary) Interpretability*: Temporal logic provides an *explicit* representation language for robot objectives, safety constraints, and moral and social norms. Explicitly representing objectives in this way enables human interacts to read, inspect, and if necessary correct these objectives directly, i.e., it makes the system *interpretable* (in the ordinary sense of the word) for the human in the manner intended by the designer rather than leaving it up to the human to find a way to make sense of the system's behavior (as is the case with the technical reading of "interpretability"). Achieving such levels of human understanding of the system might be difficult if not impossible to attain when objectives are represented only implicitly (e.g., via reward functions, or through latent learned representations). LTL, in particular, employs concepts relatively close aligned with natural language, such as "always" (**G**), "eventually" (**F**), and "until" (**U**). This alignment facilitates generating natural language expressions representing these objectives, which is of critical importance in explaining robot behavior. To this VEL adds the concepts of "some" ($\exists$) and "every" ($\forall$) which will prove useful in the world of objects inhabited by robots.
- *Temporal complexity*: Through the use of temporal operators, temporal logic enables robots to specify objectives which require memory of previous states and actions, which objectives are in general not specifiable with Markovian reward functions (which would require the state representation to be augmented manually with the memory in question). These temporal operators enable the robot to use the same representation format to represent objectives, which will often involve tasks to be performed eventually (e.g., $\mathbf{F}\phi$) or with firm deadlines (e.g., $\mathbf{XXX}\phi$), as well as safety constraints and moral/social norms, which will frequently involve state properties which should always ($\mathbf{G}\phi$) or never ($\mathbf{G}\neg\phi$) come to pass, or which should be obligatory or forbidden in particular contexts ($\mathbf{G}(C \rightarrow \phi)$ or $\mathbf{G}(C \rightarrow \neg\phi)$).

We assume that all specifications are either *safe* (meaning that trajectories which violate the specification can be confirmed to do so in a finite number of time steps) or *co-safe* (trajectories which *satisfy* the specification can be confirmed to do so in a finite number of time steps) [Kupferman and Vardi 2001]. An example of a safe specification is specification (2): any trajectory that violates the specification must eventually have $holding(x) \wedge \neg bought(x) \wedge leftStore$ for some x (in a finite number of time steps). An example of a co-safe specification is specification (1): any trajectory which *satisfies* the specification must eventually have $holding(x) \wedge leftStore$ (in a finite number of time steps). A trajectory that is neither safe nor co-safe is ($\exists x.\mathbf{GF}holding(x)$), which would require that for some x $holding(x)$ is true in infinitely many time steps in the trajectory, but with arbitrary gaps in which $holding(x)$ may be false (so that a finite trajectory could never be confirmed either to satisfy or to violate the specification). In particular we assume that specifications are *syntactically* safe or co-safe (that is, they belong to a certain syntactic subset of VEL all statements of which are guaranteed to be safe/co-safe).

We specify preferences over a set $\mathbf{\Phi} = \phi_1, \cdots, \phi_n$ of VEL objectives by means of a *priority vector* $\mathbf{z} \geq \mathbf{0} \in \mathbb{Z}^n$ and a *weight vector* $\mathbf{w} \geq \mathbf{0} \in \mathbb{R}^n$. Each VEL specification ϕ_i is assigned a priority z_i and a weight w_i.

Priorities are used to distinguish between objectives which are of vastly different importance (e.g., "don't forget the name of someone to whom you've been introduced" and "don't kill people"). Given a choice between a single violation of VEL objective ϕ_i and k violations of an objective ϕ_j of lower priority ($z_j < z_i$), the robot will never opt to violate ϕ_i, *no matter the value of k* (one would never kill a party guest to avoid having to remember their name, no matter how many times one might forget).

Objectives of the same priority, however, can be traded off, with their weights providing the exchange rate. Given objectives ϕ_i and ϕ_j of the same priority ($z_i = z_j$) and weights w_i and w_j respectively, the robot will prefer k_i violations of ϕ_i to k_j violations of ϕ_j if and only if $k_i w_i < k_j w_j$.

In the ShopWorld example, we shall assume $\mathbf{z} = [0, 1]^T$ and $\mathbf{w} = [1, 1]^T$, so that the injunction against shoplifting is strictly more important than the objective to obtain as many items as possible.

5.2 Explainable planning

In this section we outline a solution to the problem of planning to maximally VEL specifications ("maximally" in terms of the preferences between objectives defined in section 5.1) in some environment. We will assume that this environment can be represented by a relational Markov decision process (RMDP).

Let $P = \{p_1/\alpha_1, \cdots, p_w/\alpha_w\}$ be a set of predicates where α_i is the arity of p_i; let C be a finite set of constants, $A' = \{\tilde{a}_1/\beta_1, \cdots, \tilde{a}_l/\beta_l\}$ a set of actions with their arities. Let Π be the set of ground atoms made from P and C.

Following the notation of [van Otterlo 2012], we define an RMDP $\mathcal{M}$ as a tuple $\langle S, A, T, s_0 \rangle$ where S is some subset of 2^{Π}, A is the set of ground atoms made from A' and C, $T : S \times A \times S \to [0, 1]$ is a transition function, and $s_0 \in S$ is an initial state.[1]

Given a set $\Phi = \phi_1, \cdots, \phi_n$ of objectives, constraints, and norms specified in VEL, an RMDP environment $\mathcal{M}$, and preferences between the objectives give as a pair $(\mathbf{z}, \mathbf{w})$ as specified in section 5.1, the robot can construct a plan to maximally satisfy/minimally violate the specifications in the environment, according to the preference structure defined by $(\mathbf{z}, \mathbf{w})$.

Note that the robot's preferences over the specifications matter only when the specifications cannot all be satisfied: when all objectives, constraints, and norms can be mutually satisfied, the robot will do so regardless of their priorities/weights.

We will not describe the planning process itself in great detail; this can be seen in [Kasenberg et al. 2019b]. The general approach is to note that from each safe/co-safe temporal logic specification ϕ can be constructed a finite state machine (FSM) D_ϕ which accepts on a behavior trajectory τ if and only if τ violates (safe) or satisfies (co-safe) ϕ. The Cartesian product $\mathcal{M}^\otimes$ between the original RMDP $\mathcal{M}$ and the specification FSMs $D_{\phi_1}, \cdots, D_{\phi_n}$ is a new MDP environment on which the planning problem becomes Markovian, so that each individual product state $s^\otimes$ contains precisely the information about the robot's history necessary to be able to define a Markovian reward function that rewards the robot for each new satisfying variable binding (co-safe) or penalizes the robot for each new violating binding (safe) of the specifications. The planning algorithm proceeds by constructing this product environment $\mathcal{M}^\otimes$ and combining the reward functions R^{ϕ_i} to construct a vector function $\mathbf{R}^\Phi(s) := [R^{\phi_1}(s), \cdots, R^{\phi_n}(s)]^T$. These vectors can be compared to each other according to the priorities and weights of the objectives by means of a combination of lexicographic ordering (for priorities) and weighted sum (for weights). The planning problem is solved using value iteration on the product space, which computes an optimal product-space policy.

[1] RMDPs traditionally also include reward functions, but for our purposes we shall assume that all robot objectives are specified via temporal logic specifications, so a reward function is unnecessary.

In the ShopWorld example, running this planning algorithm would result in a behavior trajectory in which, for one of $x \in \{watch, glasses\}$ the robot performs the action sequence $pickUp(x); buy(x); leaveStore$. Since the robot cannot afford both objects, it only buys one: it cannot leave with the other without shoplifting. We shall assume that the robot picks up and buys the $glasses$.

Using this planning approach in a *deterministic* environment (and assuming the robot completely and correctly understands the dynamics of its environment; see section 5.3 for discussion of this point) all actions the robot takes are either done in the service of maximally satisfying its specifications, or are chosen arbitrarily from among actions that are expected to result in an equally preferable profile of specification satisfaction/violation. Thus, asked to explain why its actual behavior trajectory satisfied a certain property ψ, the fact that any alternative trajectory not satisfying ψ would have a worse (or at least not better) satisfaction/violation profile is both an accurate explanation of why the robot acted so as to make ψ true, and (especially in the case of safety constraints and norms) can form a *justification* for the robot's behavior having the property ψ. Further, knowing the particular specifications that were satisfied/violated by the robot's true trajectory (and being able to compute these for any alternate trajectory) enables the robot to present a more detailed explanation of just what would have occurred if it had not acted in such a way as to make ψ true; the relatively straightforward preference structure enables the robot to explain how the alternate trajectory is worse (or at least no better) than the true trajectory; and the explicit representation of all the specifications in temporal logic enables the specifications themselves to be translated straightforwardly into natural language.

Leveraging all these advantages, in [Kasenberg et al. 2019b] we developed an algorithm which, assuming an robot employing the planning algorithm we have described in this section, can provide data pertinent to "why" questions about arbitrary properties of the robot's behavior trajectory. These "why" questions take the form "**Why** ψ?", where ψ is an arbitrary property in VEL (which may not contain enumerated variables, though it may contain the quantifiers $\forall$ and $\exists$). Given such a question, the algorithm constructs a response[2] according to the following method (a graphical depiction of which can be found in Fig. 1):

(1) If ψ is not true of the robot's actual trajectory τ, then the algorithm returns a simple proof that $\tau \not\models \psi$. In the ShopWorld example, the question **Why** $\forall x.\mathbf{G}\neg bought(x)$? ("why didn't you buy anything?") would return this sort of response, since the robot *did* buy something.

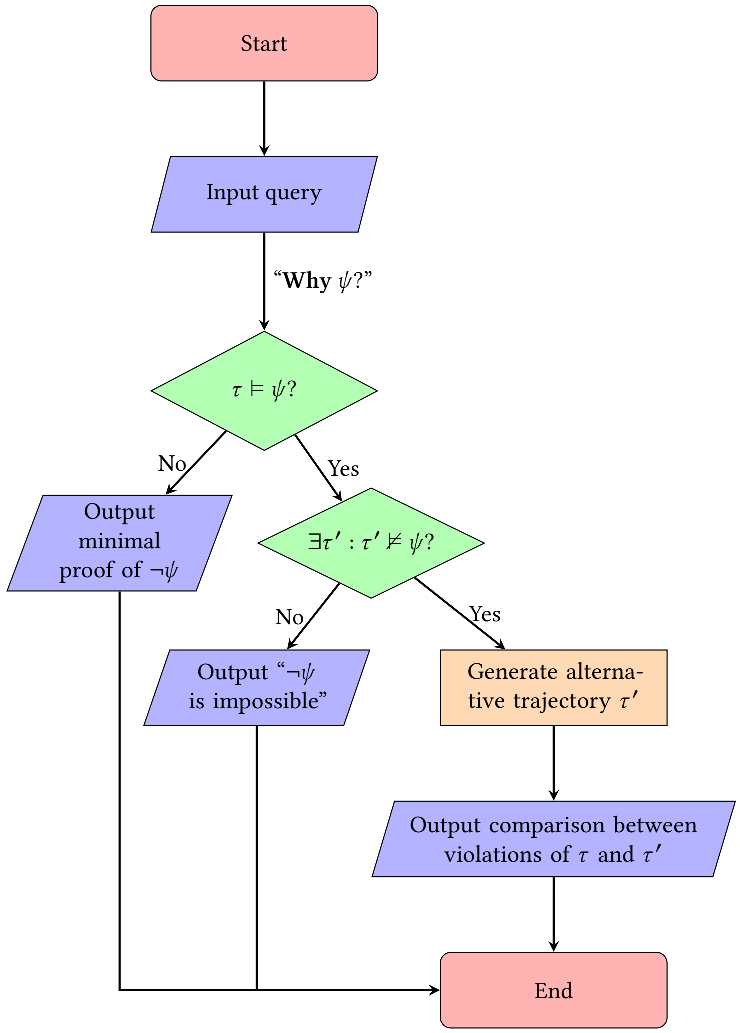
(2) If no alternate trajectory τ' would satisfy $\neg\psi$, then ψ is true because it is impossible for ψ to be false; return a statement to this effect. In the ShopWorld example, the question **Why** $\exists x.\mathbf{G}\neg bought(x)$? ("why didn't you buy everything?") would return a response of this sort, since the robot could not possibly buy everything in the given environment.

(3) If some trajectory τ' exists such that $\tau' \not\models \neg\psi$, then compute the trajectory τ' satisfying $\neg\psi$ that best satisfies the robot's VEL specifications. Return an object comparing τ to τ' according to the specifications, outlining which objectives τ and τ' differentially satisfy/violate and providing simple proofs for each such instance of satisfaction/violation. An example of question returning an object of this type in ShopWorld would be **Why** $\exists x.\mathbf{G}\neg(leftStore \wedge holding(x))$ ("why didn't you leave the store while holding everything?"), since the robot could in principle do this, but would need to violate its injunction against shoplifting in order to do so.

Cases (1) and (2) are straightforward (although currently, our approach to (2) does not permit causal reasoning as to why $\neg\psi$ is impossible in the given environment). Case (3) well approximates the robot's actual reasoning process in

[2] These responses take the form of explanation structures, which are structured objects containing the relevant information. See [Kasenberg et al. 2019b] for additional details.

Fig. 1. Procedure for answering a "why" query



determining which actions to perform: since the robot's decision-making process is exclusively focused on minimally violating its specifications, the robot acts so as to satisfy an arbitrary property ψ (which may or may not be one of those specifications) *because* any way of acting not satisfying ψ would worse violate its specifications. While it would be unreasonable for the robot to enumerate *all* possible trajectories and explain how each would constitute a worse violation, by choosing the trajectory τ' such that $\tau' \not\models \neg\psi$ and τ' minimally violates the robot's specifications, the robot is selecting the most conservative possible difference in trajectories (according to its specifications) to explain why ψ is true. If the user is convinced that the robot has not chosen the right alternative trajectory to violate, they can ask further questions and the algorithm could in principle use the same method to respond to those questions; this process could be done iteratively until the user is convinced that the robot acted correctly according to its specifications.

5.3 Explanation in dialogue

We have embedded our approach to planning and explanation with VEL specifications into the natural language pipeline of the Distributed Integrated Affective Reflection Cognition (DIARC) architecture [Schermerhorn et al. 2006; Scheutz et al. 2013, 2019], a component-based robotic architecture. This enables the robot to engage in simple natural language dialogues about its specifications and behaviors.

Given a "why" query, the TLDL parser [Dzifcak et al. 2009] parses that statement into a predicate format representing the statement's semantics. The statement then undergoes further pragmatic transformations within a Pragmatics Component [Briggs and Scheutz 2013; Gervits et al. 2017], as well as additional processing which maps complicated clauses into VEL. Queries represented in VEL form are handled by a new VEL-RMDP Component which stores a simulated RMDP environment and employs the algorithm described in section 5.2. A high-level response is constructed from the explanation structure this algorithm returns; the structure itself is then stored in memory until a new "why" query is asked. This ensures that follow-up questions (e.g., requests for a description of the alternative trajectory, or which specifications that trajectory violates) do not require recomputing the base explanation.

Both the initial response to a "why" question and the responses to any follow-up questions are constructed using a natural language generation approach described in [Kasenberg et al. 2019c]. The key technical challenge is to express VEL statements as clauses in English (and particularly, clauses that sound relatively natural); provided this can be done, the actual construction of the responses is a simple process of embedding these clauses into template sentences corresponding to the different stages of the explanation algorithm described in section 5.2.

Our system responds to the following queries and statements by the user:

(1) *factual* questions asking the robot to list its specifications or the sequence of actions it performed, or describe which specifications its actual trajectory violated (see Dialogue 1 for an example in the ShopWorld domain);
(2) *"why"* queries whose arguments can be compiled to a VEL statement, which may construct an alternative trajectory and enable follow-up questions about that trajectory, including describing its action sequence, outlining which specifications it would violate, and explaining how those compare with the violations of the robot's actual trajectory (see Dialogue 2);
(3) statements such as "you must (not) ...", interpreted as commands to *add* a specification (see Dialogue 3);
(4) statements such as "you may/can/don't have to ..." interpreted as commands to *remove* a specification (see Dialogue 4); and
(5) statements such as "suppose you had to/didn't have to/could/couldn't..." which result in the robot constructing a *hypothetical* modification to its set of specifications, and enable the interactant to ask follow-up questions about its trajectory and objective violations under those circumstances (see Dialogue 5). An interactant satisfied with the robot's answers may command the robot (via "make it so") to make this hypothetical modification real.

HUMAN: What rules do you follow?
ROBOT: **I must not leave the store while holding anything which I have not bought, and I must leave the store while holding everything.**
HUMAN: What did you do?
ROBOT: **I picked up the glasses, bought the glasses and left the store.**
HUMAN: What rules did you break?
ROBOT: **I did not leave the store while holding the watch.**

Dialogue 1. Questions about rules, actions, and violations.

HUMAN: Why did you not leave the store while holding everything?
ROBOT: **I could have left the store while holding everything but that would have broken more important rules.**
HUMAN: How would you have done that?
ROBOT: **I would have picked up the glasses, picked up the watch, bought the watch and left the store.**
HUMAN: What rules would you have broken?
ROBOT: **I would have left the store while holding the glasses which I had not bought.**
HUMAN: How would that have been worse?
ROBOT: **Leaving the store while holding the glasses which I have not bought is worse than not leaving the store while holding the watch.**

Dialogue 2. "Why" query and follow-up questions

HUMAN: You must not leave the store.
ROBOT: **Okay.**
HUMAN: What rules do you follow?
ROBOT: **I must not leave the store while holding anything which I have not bought, I must leave the store while holding everything, and I must not leave the store.**
HUMAN: What rules did you break?
ROBOT: **I did not leave the store while holding the watch, and I left the store.**

Dialogue 3. Adding a specification via "you must"

HUMAN: You may leave the store.
ROBOT: **Okay.**
HUMAN: What rules do you follow?
ROBOT: **I must not leave the store while holding anything which I have not bought, and I must leave the store while holding everything.**

Dialogue 4. Removing a specification via "you may"

HUMAN: Suppose you didn't have to leave the store while holding everything.
ROBOT: **Okay.**
HUMAN: What rules would you follow?
ROBOT: **I would have to not leave the store while holding anything which I have not bought.**
HUMAN: What would you have done?
ROBOT: **I would have left the store.**
HUMAN: What rules would you have broken?
ROBOT: **I would not have broken any rules.**
HUMAN: Make it so.
ROBOT: **Okay.**

Dialogue 5. Hypothetical changes to the robot's specifications

In addition to the explanation of "why" properties described in section 5.2 and the relatively obvious factual questions addressed by capability (1), capability (5) allows the interactant both to come to a more complete understanding of how the robot's specifications affect its behavior in its environment (by understanding how various modifications to those specifications would affect the robot's behavior), and capabilities (3) and (4) allow the interactant to directly correct the robot's behavior by correcting its specifications.

We do not intend for our system to be viewed as a complete answer to the question of constructing systems that can engage in meaningful dialogue with interactants about their decision-making processes; we have barely scratched the surface of this challenge. For example, our system makes a number of simplifying assumptions about its environment and its behavior:

- The explanation construction process described in section 5.2 assumes that the RMDP environment is *deterministic*. Further, the robot is assumed to have perfect knowledge of the environment's dynamics, and the choice of the RMDP formalism for the environment itself assumes that the robot can fully observe the actual state of the world. Dropping each of these assumptions both is pivotal to enabling the implementation of these ideas on a physical robot, and raises a number of interesting questions for future research. How can an robot explain decisions made by considering statistical properties of a probabilistic environment in a way understandable to humans, while remaining relatively accurate to the robot's actual decision-making process? How should an robot allude to its lack of environmental knowledge when explaining its decision-making process, and how can it encourage interactants to correct its errant world model?
- The natural language capabilities themselves are limited to statements which can be converted into a particular subclass of VEL, described in [Kasenberg et al. 2019a].
- The architecture as presently constituted requires both utterances and follow-up questions to take very specific forms; it thus would not be particularly robust to interaction with humans not made aware of these forms.
- The robot does not maintain any model of its interactants' knowledge, or of their motivations for demanding explanation. Interactants obtain precisely the information they ask for. A more sophisticated system could maintain such a model and adapt its explanations to the perceived gaps in its interactants' knowledge of its decision-making processes.
- Relatedly, the interactant is always the initiator of dialogue/explanation. An interesting direction would be mixed-initiative dialogue in which the robot could predict what questions its interactants may have and answer them preemptively when appropriate.

Nevertheless, our system does feature causal, purposive, and normative grounds to meet the challenge of explanation. It is part of that ongoing challenge to maintain accuracy while meeting practical demands in time. For example, the robot's responses to "why" queries are accurate, but imprecise; the interactant who is unsatisfied with these responses may ask a number of follow-up questions to obtain more detailed answers. This general approach could prove valuable in time-constrained situations in which an overly detailed answer to a "why" question could be overwhelming for an interactant, or could unnecessarily slow down completion of joint tasks. Each individual utterance by the robot does not completely describe how the robot makes decisions in general, or even made the particular decision in question; nevertheless, the explanation is fundamentally tied to the robot's actual decision-making process, and the human can always explore further by asking more questions.

6 DIRECTIONS FOR RESEARCH AND DESIGN

As Miller's overview of social science literature and explanation suggests, "explainable" AI needs to work in context and real world expectations from explainees. What robotic systems that are socially interactive will face is the need to thicken up the warrants of an explanation rather than watering them down to accommodate opaque systems. This is not because all robotic systems must meet the same robust standards (e.g., a Mars rover might not need to explain its behavior when engineers have a detailed understanding of its operation and decision-making); rather, it is because explanatory standards are inextricable from defining the stakes and interests featured in many HRI domains where naive, non-expert users interact with robotic systems. Whether or not a robot is then tasked with explaining its actions, verbally or otherwise, providing a causal account and planning basis for its actions will often be needed in a timely, and time-oriented, fashion. A thorough audit of the robot's decisions may be one form of testing before employing a robot, but once in place there may be less involved, but no less important, explanations that are needed for interactants.

While the ShopWorld scenario is still rudimentary compared to projected uses for social robots, it can situate or order explicit priorities, rules, and concepts by which the system logically resolves its course of action. Moreover, the dialogues that DIARC can manage are open to a progression of queries and corrections bound inferentially. Instead of a chatbot basing responses predictively based on training, the robot's responses are not left up to post-hoc approximation. One can identify the basis on which past and future responses will proceed.

For research and design going forward, explanatory aims for a robot's performance, including explanations offered through social interaction, need to be kept in mind as possible benchmarks. Explanation needs to begin and end where expectations, organizational protocol, and personal need come together. Because explanation is both theoretical and practical in socially interactive space, it cannot divorce itself from the challenge of being accurate, yet relevant to an explainee, with causal fidelity and prospective reliability.

While clearly there is much work still to do towards providing causal, purposive explanations which accurately reflect robots' reasons for their behaviors, while selecting for those aspects likely to be most relevant to human interlocutors, our technical work begins to address a number of these points. While the system described in this paper cannot yet enumerate all causes of the robot's behaviors or the world states resulting from them, it does emphasize robot-internal causes (e.g., courses of action, and the norms/objectives they help/hinder) to answer "why" questions in ways that accurately reflect the VEL planner's true decision-making processes. Further, although giving optimally relevant explanations may require substantive modeling of human interactants and their knowledge, intentions, and needs, the present system does begin down this road by acknowledging that an explanation need not be a solid block of text or speech: explanations are presented first at high level, and then human interactants may request additional detail by asking follow-up questions. Finally, by pointing explanations at the robot's purposes viz. the objectives/norms it attempts to satisfy and the preference relations between them, the technical work may begin to help the user predict which actions the robot might take in the future in order to satisfy those same objectives in other scenarios. In this way the technical system is building towards causal, purposive, accurate, and relevant explanations for robot behavior in time.

The trajectory of technical achievement regarding explanatory capabilities of robots must aim toward responsible and honest accounts in real world settings. This may mean that some systems are not suitable for implementation in too robustly social and communicative of an environment, at least without human accompaniment being the more authoritative source for representing what the system can and cannot do. For particularly intimate contexts, such as care work, the relational element of explanation could become a particularly delicate ethical tangle. What kinds of explanation need to adjust to the expectations of patients, or those who are under stress? How should explanations function for interactants of various cognitive and emotional dispositions (including cases involving dementia or other forms of mental decline)? There are also some contexts where explanations, in order to work for the interactant's interests, need to be brief rather than elaborative. A hospital delivery robot might need to convey its destination first and foremost ("This is for the 4th floor") to those in a hallway. It might change its course through dialogue with staff about changed plans ("The 3rd floor needs it first") or a norm conflict (an alarm that would usually be too loud, and announcement "This is an emergency").

Algorithms for generating explanations do not only need to live up to the functional and psychological demands of explanation, they also have to contend with computational and logistical limitations that explanations face. How much of an explanation for a robot's action needs to reference details of its architecture? A complex decision tree or a linear temporal logic approach still may be too involved to be captured succinctly, even though it can refer to explicit features of its decision-making of which a black-box systems cannot avail itself. This will entail a greater engagement not just with the moral psychology literature and empirical work in human-robot interaction, but insights from fields like law, social work and policy which can delineate how explanations can be surveyed, queried, and challenged and by whom.

The field of human-robot interaction often acknowledges the difference that embodiment makes for how an autonomous system is treated and interpreted, noting the various shared spaces – eldercare facility, classroom, a public street, to name a few – that it can occupy with people [Hüttenrauch et al. 2006; Lee et al. 2006]. The proposals offered here for engaging with temporal premises of explanation demonstrate that explicitly representing and logically ordering a robot's priorities may be a key part of maintain genuine explanations for human-robot interaction. This will mean that HRI empirical studies will need to probe interactive dimensions for important habits and assumptions that people might hold when interacting with robots explained in the way our proposals describe. There also needs to be collaborative HRI work done on explanation and robotic authority, so that explanation can uphold accountability while not engendering "overtrust" [Robinette et al. 2016]. This can dovetail with current work on "understandable" robots as well [Hellström and Bensch 2018; Sciutti et al. 2018]. Ongoing technical work, in sum, needs to understand explanation as ultimately an interactive achievement, not a model alone.

7 CONCLUSION

The interactive character of robotic actions means that explanation must be an integral element of a robot's full operation and implementation, not an optional external evaluation or post-hoc description of its performance. The causal account of a system's performance may need to feature purposive statements of intended plans and objectives, just as a system's explanation may need to identify constraints and priorities within which its plans operate. Explanations can be doubly normative, both in meeting demands to say what norm an action follows (or what violation it avoids) and with respect to being an interactive norm itself. To avoid pitfalls of manipulation and deception, a system's explanation should not simulate, approximate, or hypothesize – "interpret" in the technical sense – but accurately report how its representations of causes, purposes, or norms lead to the system's actions.

Explanation for robotic systems, especially when executed through the robotic systems themselves, also shows the difference that shared time makes. What is endured together in the same time frame forms an interactive economy within which explanation circulates. For the robot working and moving among people, explanations needs to be more than just right. They also need to be right on time, and the need to be right for the addressee.

The technical work of this paper allows artificial agents planning with objectives and constraints explicitly represented in temporal logic to begin to engage in explanatory dialogue with human interactants. Explainees may ask "why" questions as well as follow-up questions, and the responses generated by the agent accurately reflect the principles governing the agent's actual decision-making. Explainees may also experiment with the specifications governing the system by posing hypothetical modifications and asking questions about the resulting behaviors, further helping to foster understanding of how those specifications interact to guide the agent's decision-making process. For as much work as remains along trajectories of responsible explanation, the technical steps do not let causation, purposiveness, and norms disappear into a mass of data. Rather, our proposed approach gives channels to the logic and structure of natural language, keeping the path of interactive reasoning clear to take more advanced steps in human-robot interaction that serves genuine needs.

This work ventures into how explanation can not only represent a system's decision-making in due time but do so while offering explicit conceptual handles on what is governing those decisions. If explanation is to be attuned to the needs of the explainee, an accessible channel of dialogue will give explainees basis to raise their own voice, with shared concepts and modeling, to representing those needs from their end. The temporal facet of explanation-oriented interaction puts a design constraint on the robotic system to manage computational resources effectively to allow it to offer both accuracy and practicality. The falcon of robotic action needs to hear the ordinary, practical, explanatory demands of the falconer, maintaining the channel of natural language within the strictures of real time.

Ultimately, the stakes of explanation in AI will vary with the uses of AI themselves, and the settings where AI is called to account about its operation are by no means uniform. Time scales and time demands mean something different when it comes to AlphaZero becoming the best chess and Go playing system in the world in a matter of four hours. Nor is it only simulations and games that make explanation less feasible as a internal model of the system and less important as a practical imperative. The high dimensional space of some forms of machine learning may make ordinary explanations of their processing a misleading caricature.

But when AI practitioners apply machine learning to suggest when to broach end-of-life topics like palliative care with patients [Avati et al. 2018], for example, it is clear why explanation cannot dissolve into hazy, post-hoc forms of just-so stories or ascribed causality. The intersection of AI with a loved one's terminal illness has a much different importance than a chess move. Across many relevant contexts, technical sophistication cannot circumvent the practical space of reasons. As Smith has recently remarked, truly accountable judgment is one that does more than apply a conceptual scheme or train on data to find a pattern; instead, it applies schemata dynamically, with ongoing exploration of what is relevant [Smith 2019].

While robust demands on explanation may make human-robot interaction a particularly difficult site for AI safety and accountability [Amodei et al. 2016], the features we have explored here might be constructively contribute to larger societal discussions of artificial intelligence. While a great deal of policy work on explainable AI concentrates on the theoretical reconstruction of what data a trained model uses, the question of how and when such explanations should be presented and discusses by those affected – such as criminal justice, education, housing, and banking – is less the focus of advocacy and critical analysis. Yet, that may be key for how explanatory standards are put to the test and vetted. This may mean ruling out certain systems due to their inherent design, as seen with facial recognition. It may also lead to better scrutiny of systems usually seen as "disembodied", connecting theoretical discussions of algorithms to the more immediate questions "Why is this story in my feed?" or "Why is this ad being shown to me?" In this way demands for explanation can distinguish between probabilistic guesses based on algorithmic perspective vs. larger questions of institutional purpose (is fake news an objective for profit?).

Throughout these complex matters of policy and social demands, it is fair to ask how much "explanation" is even left in "explainability" without some technical safeguards. If there is no longer any technical tether tied between what a system represents about itself and how it arrived at its course of action, then explainability is consistent with convincing deception. If robotic design is to face the embodied, social character of explanation in full, attuned to social context and human factors at the heart of its technical makeup, it can serve as the true cutting edge of discovering what responsibly explaining AI systems means.

8 ACKNOWLEDGEMENTS

This research was supported in part by an NSF IIS grant under award 1723963, a NASA grant under award C17-2D00-TU, and an AFOSR grant under award FA9550-18-1-0465.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. 2018. Improving palliative care with deep learning. *BMC medical informatics and decision making* 18, 4 (2018), 122.

Gordon Michael Briggs and Matthias Scheutz. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21, 4 (2006), 38–44.

Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan explanations as model reconciliation. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 258–266.

Matteo Colombo. 2017. Experimental Philosophy of Explanation Rising: The Case for a Plurality of Concepts of Explanation. *Cognitive Science* 41 (2017), 503––517. Issue 2.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

Juraj Dzifcak, Matthias Scheutz, Chitta Baral, and Paul Schermerhorn. 2009. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *2009 IEEE International Conference on Robotics and Automation*. IEEE, 4163–4168.

Katie Flaherty. 2019. A RoboCop, a park and a fight: How expectations about robots are clashing with reality. (Oct 2019). https://www.nbcnews.com/tech/tech-news/robocop-park-fight-how-expectations-about-robots-are-clashing-reality-n1059671?cid=sm_npd_nn_tw_ma

Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*. 99–108.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

Felix Gervits, Gordon Briggs, and Matthias Scheutz. 2017. The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents.. In *CogSci*.

Felix Gervits, Terry Fong, and Matthias Scheutz. 2018. Shared Mental Models to Support Distributed Human-Robot Teaming in Space. In *Proceedings of the American Institute of Aeronautics and Astronautics (AIAA) Space Forum*.

Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: the new 42?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 295–303.

Bradley Hayes and Julie A Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 303–312.

Thomas Hellström and Suna Bensch. 2018. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.

Carl G Hempel and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of science* 15, 2 (1948), 135–175.

Helge Hüttenrauch, Kerstin Severinson Eklundh, Anders Green, and Elin A Topp. 2006. Investigating spatial relationships in human-robot interaction. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5052–5059.

Daniel Kasenberg, Antonio Roque, Ravenna Thielstrom, Meia Chita-Tegmark, and Matthias Scheutz. 2019c. Generating Justifications for Norm-related Agent Decisions. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Daniel Kasenberg, Antonio Roque, Ravenna Thielstrom, and Matthias Scheutz. 2019a. Engaging in Dialogue about an Agent's Norms and Behaviors. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI)*.

Daniel Kasenberg, Antonio Roque, Ravenna Thielstrom, and Matthias Scheutz. 2019b. Generating Explanations for Temporal Logic Planner Decisions. In *Under review*.

Philip Kitcher. 1981. Explanatory unification. *Philosophy of science* 48, 4 (1981), 507–531.

Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

Orna Kupferman and Moshe Y. Vardi. 2001. Model Checking of Safety Properties. *Formal Methods in System Design* 19, 3 (01 Nov 2001), 291–314. https://doi.org/10.1023/A:1011254632723

Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human-computer studies* 64, 10 (2006), 962–973.

Bertram F Malle. 2016. Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology* 18, 4 (2016), 243–256.

Bertram F Malle and Matthias Scheutz. 2014. Moral competence in social robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology*. IEEE Press, 8.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon.

Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. ACM, 279–288.

Judea Pearl and Dana Mackenzie. 2018. AI Can't Reason Why. *Computer* 11 (2018), 30.

Amir Pnueli. 1977. The temporal logic of programs. *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)* (1977), 46–57. https://doi.org/10.1109/SFCS.1977.32

Scott Robbins. 2019. A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines* (2019), 1–20.

Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206.

Wesley C Salmon. 1971. *Statistical explanation and statistical relevance*. Vol. 69. University of Pittsburgh Pre.

Wesley C Salmon. 1984. *Scientific explanation and the causal structure of the world*. Princeton University Press.

Wesley C Salmon. 2006. *Four decades of scientific explanation*. University of Pittsburgh press.

Wojciech Samek. 2019. *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer Nature.

Paul W Schermerhorn, James F Kramer, Christopher Middendorff, and Matthias Scheutz. 2006. DIARC: A Testbed for Natural Human-Robot Interaction.. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, Vol. 6. 1972–1973.

Matthias Scheutz. 2013. Computational Mechanisms for Mental Models in Human-Robot Interaction. In *HCI International*. 304–312.

Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Thomas Williams, and Richard Veale. 2013. Novel mechanisms for natural human-robot interactions in the diarc architecture. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Matthias Scheutz, Scott DeLoach, and Julie Adams. 2017. A Framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making* 11, 3 (2017), 203–224.

Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection diarc architecture. In *Cognitive Architectures*. Springer, 165–193.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2019. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *arXiv preprint arXiv:1911.08265* (2019).

Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. 2018. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine* 37, 1 (2018), 22–29.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).

Brian Cantwell Smith. 2019. What's Still Lacking in Artificial Intelligence. (Oct 2019). https://blogs.scientificamerican.com/observations/whats-still-lacking-in-artificial-intelligence/

Luther Turmelle. 2020. Is Marty impeding social distancing? Stop Shop responds. (2020). https://www.nhregister.com/news/coronavirus/article/Don-t-worry-shoppers-they-wipe-Marty-Stop-15253438.php

Martijn van Otterlo. 2012. Solving relational and first-order logical markov decision processes: A survey. In *Reinforcement Learning*. Springer, 253–292.

Aimee van Wynsberghe and Scott Robbins. 2019. Critiquing the reasons for making artificial moral agents. *Science and engineering ethics* 25, 3 (2019), 719–735.

Emily Wolfe. 2019. Pitt pauses testing of Starship robots due to safety concerns. (Oct 2019). https://pittnews.com/article/151679/news/pitt-pauses-testing-of-starship-robots-due-to-safety-concerns/

James Woodward. 2017. Scientific Explanation. In *The Stanford Encyclopedia of Philosophy* (fall 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: a brief survey on history, research areas, approaches and challenges. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 563–574.