Explainable, Normative, and Justified Agency

Pat Langley
Institute for the Study of Learning and Expertise,
2164 Staunton Court, Palo Alto, CA 94306 USA
Department of Computer Science, University of Auckland,
Private Bag 92019, Auckland 1142 NZ

Abstract

In this paper, we pose a new challenge for AI researchers – to develop intelligent systems that support *justified agency*. We illustrate this ability with examples and relate it to two more basic topics that are receiving increased attention – agents that explain their decisions and ones that follow societal norms. In each case, we describe the target abilities, consider design alternatives, note some open questions, and review prior research. After this, we return to justified agency, offering a hypothesis about its relation to explanatory and normative behavior. We conclude by proposing testbeds and experiments to evaluate this empirical claim and encouraging other researchers to contribute to this crucial area.

1 Background and Motivation

Autonomous artifacts, from self-driving cars to drones to household robots, are becoming more widely adopted, and this trend seems likely to accelerate in the future. The increasing reliance on these devices has raised concerns about our ability to understand their behavior and our capacity to ensure their safety. Before intelligent agents can gain widespred acceptance in society, they must be able to communicate their decision making to humans in ways that convince us they actually share our aims.

This challenge involves two distinct but complementary issues. The first is the need for agents to explain the reasons they carried out particular courses of action in terms that we can understand. Langley et al. (2017) have referred to this as *explainable agency*. The second is the need for assurance that, when agents pursue explicit goals, they will also follow the many implicit rules of society. We will refer to this ability as *normative agency*. Both of these functions are necessary underpinnings of trustable autonomous systems, and two capabilities are closely intertwined.

Consider a scenario in which a person drives a friend with a ruptured appendix to the hospital. The driver exceeds the speed limit, weaves in and out of traffic, runs through red lights, and even drives on a sidewalk, although he is still careful to avoid hitting other cars or losing control on turns. Afterwards, the driver explains that he took such drastic actions because he thought the passenger's life was in danger, so reaching the hospital in short order had higher priority than being polite to others along the way or obeying traffic laws. Humans can defend their behavior in this manner even when they violate serious societal norms, and we want intelligent agents of the future to exhibit the same ability to justify their decisions and actions on demand.

We discuss this challenge in the sections that follow, arguing that such *justified agency* is an important topic that deserves substantial attention from AI researchers. We first examine the two related topics of explainable agency and normative agency. In each case, we describe the desired abilities, offer illustrative examples, and touch on relevant research. We also consider the space of agent designs, including some open issues that require additional work. After this, we turn to justified agency, arguing that it combines the first two abilities and proposing a hypothesis about what else is required to support it. We close by considering some testbeds and experiments that would let the research community evaluate progress in this critical arena.

2 Explainable Agency

People can usually explain their reasons for making decisions and taking actions. When someone purchases a microwave oven, rearranges furniture in a room, or plans a vacation, he can state the choices considered, why he selected one alternative over the others, and even how his decision might have differed in other circumstances. Retrospective reports are seldom perfect, but they often offer important windows into the decision process.

As intelligent agents become both more autonomous and more prevalent, it is essential that they support similar abilities. We will say that:

- *An intelligent system exhibits* ***explainable agency*** *if it can provide, on request, the reasons for its activities.*

Consider some examples from the realm of autonomous vehicles. If we ask a self-driving car why it followed a given route, it should be able to state that the path had few lights and stop signs while still being reasonably short. More importantly, when we ask the vehicle why it swerved into another car, it should explain that it was the only way to avoid hitting an unexpected jaywalker. Like humans, such agents should have reasons for their actions and they should be able to communicate them to others when asked.

Tackling this problem requires that we make design decisions about representations and processes needed to support explainable agency, some of which we will borrow from Langley et al.'s (2017) analysis. One issue concerns what will count as legitimate explanations. Should plausible post hoc rationalizations be acceptable if an agent's decision-making procedures are not interpretable or should we require genuine insights into why the agent took its actions? We argue that only the latter should be viewed as reasonable accounts, which implies that the agents should make decisions in ways that are transparent and easily communicated. There are many well-established methods in the AI arsenal that meet this criterion, but opaque policies induced from large training sets will not suffice.

Another design choice involves whether to explain activities in terms of individual actions or higher-level structures. Research in machine learning has emphasized reactive control (e.g., Erwig et al., 2018), which supports the first alternative, whereas AI planning systems make choices about entire sequences of actions. We predict that humans will find plan-oriented explanations more familiar and easier to understand, and thus offer a natural approach to adopt. Even when a plan goes awry during execution, an agent can still give the reasons it decided to change course and how its new plan responded to the surprise. Some frameworks, like hierarchical task networks, also specify plans at multiple levels of abstraction, which would let a system offer more or less detailed explanations, down to individual actions if desired.

Any intelligent agent must use information to make decisions about which activities to pursue, and a third issue concerns how to encode this content. One response, widely adopted in AI planning research, relies on symbolic goals, often stated as logical expressions. Another option, popular in game-playing systems, instead uses numeric evaluation functions that comprise sets of weighted features. Each approach has advantages for explainable agency: goals provide explicit end points for chains of actions, while functions show how such plans handle trade offs. Although these typically appear in isolation, they can also be combined. For instance, Langley et al. (2016) describe an agent architecture that associates functions with goals, using their weighted sum to guide planning. Such hybrid frameworks offer one promising approach to building explainable agents.

Within this design space, we still need research on a number of open issues about explainable agency: These include extending intelligent systems to:

- *Generate explanatory content*. When deciding on courses of action, an agent must consider different alternatives, evaluate them, and select one of the options to pursue. This should take place during generation of plans and during their execution, producing traces that can be used in later explanations of the agent's activities.
- *Store generated content*. As it makes these decisions, an agent must cache information about the choices that it considered and the reasons that it favored one over the others, in an episodic memory or similar repository. This requires not just retaining the content about decisions, but also indexing it in ways that support later access.
- *Retrieve stored content*. After it has completed an activity, an agent must be able to retrieve decision traces that are relevant to different types of questions. This requires transforming the queries into cues and using them to access content in episodic memory about alternatives considered, their evaluations, and the final choices made.
- *Communicate retrieved content*. Once it has retrieved episodic traces in response to a question, an agent must identify those aspects most relevant to the query, translate them into an understandable form, and share the answer. This should include no more detail than needed to convey the reasons for making the decision under examination.

Research on analogical planning (e.g., Jones and Langley, 2005; Veloso et al., 1995) has addressed issues of storage, indexing, and retrieval, but not for the purpose of self report. Leake (1992) presented a computational theory of explanation, but it focused on accounts of other agents' behaviors.

The AI literature also includes other research relevant to this topic. Early explanation systems recorded inference chains and recounted them when asked to justify their conclusions (Swartout, Paris, and Moore, 1991), with some systems supporting higher-level accounts with meta-level rules (Clancey, 1983), but these did not deal with physical activities. More relevant work comes from Johnson (1994) and van Lent et al. (2004), who developed agents for military mssions that recorded their decisions, offered reasons on request, and anwered counterfactual queries. However, they dealt with knowledge-guided execution rather than agent-generated plans and, despite linking actions to objectives, did not state why some activities were preferable to others.

In more recent work, Briggs and Scheutz (2015) have reported an interactive robot that gives reasons why it cannot carry out some task, drawing on five explanation types, including lack of knowledge and physical ability. The literature on computational models of argument (e.g., Bench-Capon and Dunne, 2007) is also relevant, as it examines the structures and processes that enable an intelligent system to support the conclusions it draws from evidence, although it does not address goal-directed agency. These previous efforts offer some key elements needed for explainable agents, but we need additional research on their combination into integrated cognitive systems.

3 Normative Agency

Humans are driven by goals, but they must operate within their societies. When a hungry person seeks food, he buys it rather than stealing it. When a passenger wants to board a bus, she waits in a queue rather than cutting in front of others. When a soldier desires sleep, he nevertheless gets up when he hears reveille. In other words, people generally follow the *norms* of their society. These may involve formal laws, military orders, informal customs, or moral tenets, but they all influence and canalize behavior in certain directions.

Now that intelligent agents are becoming prominent, we would like them to behave in similar ways. We will say that:

- *An intelligent system exhibits* ***normative agency*** *if, to the extent possible, it follows the norms of its society.*

Let us return to the domain of autonomous vehicles. Clearly, we want self-driving cars to obey established laws, such as staying within the posted speed limit, driving on the correct side of the road, and stopping at red lights. However, we also want them to follow informal customs, such as not cutting in front of another vehicle and moving over to let faster ones pass. At the same time, we want them to realize that sometimes norms come into conflict and they must violate one at the expense of another. Driving is a far more complex task that simply staying on the road and avoiding collisions.

Before intelligent agents can use norms to guide behavior, we must first decide what content they will encode. One option is to specify what actions the agent should or should not carry out in certain classes of situations. This view is closely related to *deontological* accounts of ethics, championed by Kant, which emphasize fulfilling one's duties or obligations. Another choice is to associate different values with distinct states and to favor actions that produce better outcomes. This idea is linked to *consequentialist* approaches to ethics, due originally to Hume, Bentham, and Mill, with utilitarianism an important special case. At first glance, these frameworks appear to be competitors, but Spranca, Minsk, and Baron (1991) report studies that suggest people use a mixture of deontic and consequentialist methods.

A related issue concerns how an intelligent agent represents such normative content. One approach, adopted by Mikhail (2007), specifies moral tenets using logical rules, much as one can do with many formal laws. Another alternative is to state norms in terms of numeric value functions, like those used in many game-playing systems. Rules are often linked to deontic frameworks and value functions to consequentialist ones, but one can also apply rules to states and functions to actions. These approaches seem mutually exclusive, but Iba and Langley (2011) have shown how they map onto an agent architecture that associates numeric values with rule-generated structures. Norms can also specify both prescribed and proscribed actions or states (Malle et al., 2015), akin to positive and negated trajectory goals.

Within the space of agent designs, there are a number of open issues that deserve attention from the research community. These include extending intelligent agents to handle:

- *Conditional values*. We can easily associate numeric values with normative rules, but some norms may only come into play in certain contexts, and their importance may vary with situational factors. Thus, we must develop representations for laws, morals, and other norms that specify conditional values or utilities.
- *Trade offs among norms*. In some cases, norms may be incompatible, forcing the agent to decide which to obey and which to ignore. We must develop agent architectures that examine the values of relevant norms, evaluate the trade offs among different choices, and select plans or actions that give better overall scores.
- *Mitigating factors*. The importance of norms can be altered by other factors that make their violation no less serious but more acceptable. We must develop representations of such mitigating factors and methods for combining them when making choices about actions.
- *Domain-independent norms*. Many norms are domain-specific, but others are quite general, like being sensitive to another's concerns or not causing unnecessary emotional harm. These require formalisms for norms that describe beliefs about others' mental states and methods for combining such abstract constraints with concrete ones.

The AI literature reports some work on such normative reasoning, with the earliest focused on legal inference (e.g., Branting, 2000). Equally relevant has been research on machine ethics and moral reasoning (e.g., Anderson et al., 2006; Bringsjord et al., 2006; Dehghani et al., 2008; Guarini, 2005; McLaren, 2005). Some researchers have developed new representations and mechanisms to support the ability to make normative judgements and decisions, but others (Iba and Langley, 2011; Liu et al., 2013) have instead treated moral reasoning as a form of everyday cognition. Authors have demonstrated their systems on a variety of scenarios, showing that AI can address many aspects of legal, moral, and other normative reasoning, but we need more work on integrated systems that combine these abilities to operate in complex and sometimes ambiguous settings.

4 Justified Agency

Although people can explain their goal-oriented actvities, many of their accounts incorporate societal norms. When a pedestrian clarifies why he followed an indirect path, he may say that he did it to avoid walking across a neighbor's lawn. When a homeless person is asked why he begs for a handout rather than mugging someone, he might state that he avoided the latter because it was against the law. And when a shopper explains why she let another customer with only a few items check out ahead of her, she might say that, if their positions were reversed, she would have appreciated the same treatment. Our explanations often include a mixture of personal goals and more generic social constraints.

We maintain, as we have before, that intelligent agents should demonstrate similar abilities. We will say that:

- *An intelligent system exhibits* ***justified agency*** *if it follows society's norms and explains its activities in those terms.*

Let us return to our original example of taking someone who has peritonitis to the emergency room, driving aggressively and breaking traffic laws along the way. This scenario is interesting because the explanation revolves almost entirely around social norms – not only the laws and customs the driver chose to ignore, but the idea that saving another's life should take precedence over nearly all other factors. The only personal agent goals that come into play, such as avoiding collisions and not turning over, arise because they support this top-level normative aim. Again, this example shows that urban driving is a far richer activity than has been typically assumed by the research community.

If we want to develop justified agents of this sort, we must again make some design decisions. The most basic of these concern the representation of justifications. One response is to adopt the same structure as the explanations discussed earlier, which relate the agent's goals and their associated utilities to decisions in favor of certain plans and actions, but to replace some goals with normative constraints. Many societal norms specify actions or states that the agent should avoid while achieving its aims, but we might encode these in much the same way as trajectory goals used in AI planning systems. Thus, initial analysis suggests there is no serious obstacle to adapting goal-oriented accounts to include norms in support of justified activities.

The introduction of norms into explanations has similar implications for the research challenges raised earlier. When generating and evaluating plans, a justified agent must consider not only its personal goals but also social concerns, and it must include the latter in episodic traces that it stores in memory. When asked a question about its activities, the agent must be able to retrieve the ways in which its choices relate to norms and then communicate them in accessible terms. The simplest forms of justifications will treat norms as hard constraints that forced the agent to carry out some actions and avoid others. However, more nuanced accounts will include reasoning about tradeoffs that arose when norms conflicted and thus explain why some plans were preferable to alternative courses of action.

This preliminary analysis suggests that developing agents with the ability to justify their behavior may not be as difficult as it appears. We can state this as a hypothesis:

- *Any intelligent system that supports explainable agency and normative agency will also exhibit justified agency.*

In other words, once we have developed the representations and mechanisms to support the first two abilities, we will require no additional structures or processes to let agents justify their activities in normative terms. If we simply augment our agent's goals and evaluation functions with similar encodings of social mores, then it seems plausible that we will obtain justified agency with no extra effort.

Some readers will think that this conclusion follows logically from our definitions, but it is actually a scientific hypothesis that merits empirical tests. The definition of justified agency requires that it incorporate both the ability to explain decisions and to reason about norms, but it does not imply these alone are sufficient. For example, agency may be more complex than we have posited (Bello and Bridewell, 2017) and fuller analysis may reveal that norms demand richer forms of explanation. Similarly, taking such factors into account during plan generation may depend on reasoning beyond that needed with goals and utilities, or answering normative questions may require new forms of response. Such extensions may not be necessary, but we need further research to determine whether the hypothesis is accurate.

This endeavor will benefit from testbeds that support compelling demonstrations of justified agency. One candidate is urban driving, which involves a combination of goals, formal laws, informal customs, and moral tenets. Choi et al. (2007) report a simulated environment for in-city driving, implemented in the Torque game engine, that supports many vehicles, multi-lane roads, traffic signals, buildings, and pedestrians. This environment would support the types of scenarios we have discussed, although different driving simulators, if sufficiently rich, could serve equally well. Most computer games, which emphasize achieving the goals of individual agents, are less obvious options, but other simulated environments hold promise. For instance, the RoboCup Rescue environment (http://rescuesim.robocup.org/) supports multi-agent simulation of urban scenarios that involve extracting people from disaster areas. Clearly, these could include mixtures of task-oriented goals and social norms that, after a mission ends, an agent uses to justify its behavior.

Experiments that test the hypothesis would naturally take the form of lesion studies (Langley and Messina, 2004) that compare system behavior with some elements removed. Here we might take an existing agent architecture and extend it in two ways, one that supports retrospective explanation of activities and another that relies on social norms to guide behavior. Once each ability has been demonstrated on target scenarios, we can merge the extensions and examine whether the combined system provides acceptable normative explanations. If not, then analysis should reveal sources of the problem and suggest ways to extend the framework. Much of the effort will involve devising scenarios that require normative reasoning, designing questions that elicit explanations, and combining them to demonstrate justified agency. The plausibility of our hypothesis depends largely on the intuition that we can treat norms as a variety of declarative goals. This would let architectural extensions for explainable agency apply directly to traces of normative decision making. However, a careful study of social norms may reveal they differ from standard goals in ways that require separate extensions to the architecture, which could lead to unexpected interactions that violate the hypothesis.

5 Closing Remarks

In this paper, we reviewed the notion of explainable agents, which communicate the reasoning that led to their actions, and normative agents, which take into account social norms when deciding which plans and actions to pursue. We examined design choices that arise in each case, cited relevant research, and noted some open issues that require additional work. We also defined justified agency as the ability to explain one's activities in terms of societal norms. Analysis led to the claim that any intelligent system which uses such norms in making decisions and which can explain its activities will also support justified agency. We formulated this as an empirical hypothesis that requires testing, and we suggested simulated testbeds and controlled experiments that could determine its adequacy.

Research on justified agency follows the grand tradition of early AI, which identified some cognitive ability not yet reproduced in computers and developed digital artifacts that exhibited it. Designing, constructing, and demonstrating this new functionality would be an important step toward covering the full range of human intelligence. Initial forays into this arena should use simulated environments and controlled experiments to study justified agency, but we can imagine a time when autonomous agents incorporate such structures and processes in the field. Their ultimate test would be self-driving cars that defend themselves in civil suits and military robots that win court martials about their actions in combat. We encourage other researchers to pursue this audacious vision of explainable, normative, and justified agency.

Acknowledgements

This research was supported in part by Grants N00014-15-1-2517 and N00014-17-1-2434 from the Office of Naval Research, which are not responsible for its contents.

References

Anderson, M.; Anderson, S. L.; and Armen, C. 2006. An approach to computing ethics. *IEEE Intelligent Systems* 21: 56–63.

Bello, P.; and Bridewell, W. 2017. There is no agency without attention. *AI Magazine* 38: 27–33.

Bench-Capon, T.; and Dunne, P. 2007. Argumentation in artificial intelligence. *Artificial Intelligence*, *171*, 619–641.

Branting, L. K. 2000. *Reasoning with rules and precedents: A computational model of legal analysis*. Dordrecht: Kluwer.

Briggs, G.; and Scheutz, M. 2015. "Sorry, I can't do that:" Developing mechanisms to appropriately reject directives in human-robot interactions. *Proceedings of the AAAI Fall Symposium on AI and HRI*. Arlington, VA: AAAI Press.

Bringsjord, S.; Arkoudas, K., Bello, P. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21: 38–44.

Choi, D.; Morgan, M.; Park, C.; and Langley, P. 2007. A testbed for evaluation of architectures for physical agents. *Proceedings of the AAAI-2007 Workshop on Evaluating Architectures for Intelligence*. Vancouver: AAAI Press.

Clancey, W. J. 1983. The epistemology of a rule-based expert system: A framework for explanation. *Artificial intelligence* 20: 215–251.

Colaco, Z.; and Sridharan, M. 2015. What happened and why? A mixed architecture for planning and explanation generation in robotics. *Australasian Conference on Robotics and Automation*. Canberra, Australia.

Dehghani, M.; Tomai, E.; Forbus, K.; and Klenk, M. 2008. An integrated reasoning approach to moral decision making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Erwig, M.; Fern, A.; Murali, M.; and Koul, A. 2018. Explaining deep adaptive programs via reward decomposition. *Proceedings of the 2018 IJCAI Workshop on Explainable Artificial Intelligence*, 40–44. Stockholm.

Guarini, M. 2005. Particularism and generalism: How AI can help us to better understand moral cognition. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.

Iba, W. F.; and Langley, P. 2011. Exploring moral reasoning in a cognitive architecture. *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society*. Boston, MA.

Johnson, W. 1994. Agents that learn to explain themselves. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1257–1263. Seattle, WA: AAAI Press.

Jones, R. M.; and Langley, P. 2005. A constrained architecture for learning and problem solving. *Computational Intelligence* 21: 480–502.

Langley, P.; Barley, M.; Meadows, B.; Choi, D.; and Katz, E. P. 2016. Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.

Langley, P.; Meadows, B.; Sridharan, M.; and Choi, D. 2017. Explainable agency for intelligent autonomous systems. *Proceedings of the Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence*, 4762–4763. San Francisco: AAAI Press.

Langley, P.; and Messina, E. 2004. Experimental studies of integrated cognitive systems. *Proceedings of the Performance Metrics for Intelligent Systems Workshop*. Gaithersburg, MD.

Leake, D. B. 1992. *Evaluating explanations: A content theory*. Hillsdale, NJ: Lawrence Erlbaum.

Liu, L.; Langley, P.; and Meadows, B. 2013. A computational account of complex moral judgement. *Proceedings of the Annual Meeting of the International Association for Computing and Philosophy*. College Park, MD: IACAP.

Malle, B. F.; Scheutz, M.; and Austerweil, J. L. 2015. Networks of social and moral norms in human and robot agents. *Proceedings of the International Conference on Robot Ethics*, 3–17. Lisbon, Portugal.

McLaren, B. M. 2005. Lessons in machine ethics from the perspective of two computational models of ethical reasoning. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*. Menlo Park, CA: AAAI Press.

Mikhail, J. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Science* 11: 143–152.

Swartout, W. R.; and Moore, J. D. 1993. Explanation in second generation expert systems. In J.-M. David, J.-P. Krivine, and R. Simmons, eds., *Second generation expert systems*. Berlin: Springer-Verlag.

Van Lent, M.; Fisher, W.; and Mancuso, M. 2004. An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 900–907. San Jose, CA: AAAI Press.

Veloso, M.; Carbonell, J.; Perez, A.; Borrajo, D.; Fink, E.; and Blythe, J. 1995. Integrating planning and learning: The PRODIGY architecture. *Journal of Experimental and Theoretical Artificial Intelligence* 7: 81–120.