# From Explainable to Justified Agency

*Pat Langley*
Center for Design Research, Stanford University
Stanford, California 94305 USA
Institute for the Study of Learning and Expertise
Palo Alto, California 94306 USA
patrick.w.langley@gmail.com

***Abstract***
This chapter introduces explainable agents, which communicate the reasons behind their activities, and identifies three types of self explanations – structural, preference, and process – that store different forms of content about agent decisions. In addition, it considers three component abilities – indexing, retrieval, and transmission – that are required to communicate this stored content. Finally, it examines normative agents, which attempt to follow their society's maxims, and justified agents, which explain their actions in terms of such norms. The chapter also presents hypotheses about when different forms of self explanation will be most useful and about relations among explainable, normative, and justified agency.

## *Introduction*

Intelligent systems are becoming more widely adopted for critical tasks like driving cars and controlling military robots. Naturally, increased reliance on such devices has led to concerns about the interpretability of their complex behavior. Before people will fully trust such autonomous agents, they must be able to explain their decisions so that we can gain insight into their operation. There is now a substantial literature on explanation in systems that learn from experience, but it has focused on tasks like object recognition and reactive control, typically using opaque encodings of expertise that lend themselves only to shallow elucidation, as in 'heat maps' that display activation levels.

However, we also need research on explanation for more complex tasks that involve multi-step decision making, such as the generation and execution of plans. Approaches to these problems rely on high-level representations that are themselves easily interpreted, but challenges arise in communicating solutions that combine these elements and the reasons they were chosen. In this chapter, I focus on such settings. Some work on explanation, especially with opaque models, has dealt with post hoc rationalizations of behavior, rather than the actual reasons for it. In the pages that follow, I concentrate on the latter. Moreover, I will focus on *self explanations*, that is, the reasons the explaining agent carried out certain activities. Elsewhere (Langley, 2019), I have referred to this ability as *explainable agency*.[1]

We can specify the task of explainable agency in generic terms. Given domain knowledge for generating task solutions and criteria for evaluating candidates, the agent attempts to find one or more solutions. After generating, and possibly executing, these solutions, a human asks the agent to clarify its decisions, at which point it must share its reasoning in comprehensible terms. One example involves an intelligent robot that plans and executes a reconnaissance mission, after which it takes part in an 'after-action review' where it

[1]This problem is arguably less challenging than postulating the reasons that another agent behaved as it did, sometimes called *plan recognition*, as the system can store and access traces of its own decision making.

answers questions from a human supervisor. There has been some research on such *explainable planning* (Fox et al., 2017; Smith, 2012; Zhang et al., 2017), but we need more effort devoted to this important topic.

In the sections that follow, I discuss different senses of the term 'explanation' and consider some factors that arise when representing such structures. Next, I discuss three types of self explanation, along with approaches to indexing, retrieving, and transmitting them. After this, I introduce the notion of *normative agency*, which takes social maxims into account during decision making, and *justified agency*, which explains choices in terms of social norms. Along the way, I also propose some hypotheses about self explanation that merit further study.

## *Aspects of Explanation*

Two aspects of human explanations place constraints on AI approaches to replicating their generation. First, they invariably involve some form of *cognitive structure* that relates items of interest. For instance, a diagnosis links observed symptoms to hypothesized problems, often through multiple steps. Second, these structures typically comprise elements of *knowledge* that have been instantiated for the task at hand. Thus, the steps in a diagnosis might be instances of generic rules that relate symptoms to causes. Explanatory structures vary along a number of dimensions. They may be entirely qualitative, as in a geometry proof, or they may include quantitative annotations, as in the solution to a physics word problem. Accounts also differ in their complexity (e.g., the number of knowledge elements) and their depth (e.g., the length of reasoning chains). Nevertheless, they share many features that one can discuss in general terms.

We should distinguish between two uses of 'explanation' that commonly appear in English. The word sometimes refers to a mental, written, or spoken *structure* that serves to elucidate some phenomena or behaviors. Thus, we refer to a scientific explanation of pulsar cycles, a mechanical explanation of how a toilet flushes, or an introspective explanation for one's home-buying decision. In other cases, the term denotes the *process* or *activity* of generating such an explanatory structure. We say that an astrophysicist engages in explanation of pulsar behavior, a plumber focuses on explanation of a leak, or a home buyer carries out explanation of his residential choice. This chapter will use both senses of the term, but its meaning should be clear from the context in which it appears.

We can further differentiate between two specializations of explanatory processes. The first refers to the *construction* of accounts for observed situations or events. A geologist posits a set of processes for the origin of a landform, a reader infers the goals of a novel's character, and a home buyer records the reasons for his decisions. The result is a cognitive structure in the explainer's own mind. The second meaning instead deals with the *communication* of such mental structures once they exist. The geologist presents a talk about his account of a landform's evolution, the reader shares with a friend his guesses about the character's motivations, and the home buyer tells his partner why he favors one house over others. This second sense applies not only to sharing accounts of external events, but also to communicating why one made a given decision or generated a particular plan. Thus, it includes the process of *self explanation*, the important specialization on which I will concentrate here.

## *Representing Explanations*

We have seen that explanations are cognitive structures an intelligent system can construct or communicate, so both their form and content merit discussion. Such accounts link a set of observations or decisions to

each other through a set of relations that serve as connective tissue. Explanations invariably draw on background knowledge, typically at the domain level (e.g., how refrigerators operate, regulations about driving) but they sometimes involve the meta level (e.g., conventions of dialogue). However, they do not incorporate generalized knowledge elements themselves, but rather refer to *instances* of such knowledge elements that connect facts or queries to each other.

In rule-based frameworks, explanations are organized as one or more proof trees with shared subproofs, where each rule instance links observed or inferred beliefs (e.g., Ng & Mooney, 1990). For instance, an account for why an automobile does not start might connect observed behaviors through instantiated rules that describe a generic car's operation (e.g., Reiter, 1987). In script and frame paradigms, the knowledge elements are large enough that some accounts involve a single instantiated structure, although they can combine more than one (e.g., Shrager, 1987). An explanation can also involve an analogy, where knowledge corresponds to stored cases (linked facts), one of which maps onto elements of the new situation. Any formalism (e.g., rules, scripts, frames, or cases) that encodes knowledge structures can serve in this capacity.

In addition, explanations can differ in the ontological character of the knowledge elements on which they draw. These may denote logical relations, like those in geometry proofs, but they may also incorporate numeric calculations, as arise in solutions to textbook physics problems (e.g., VanLehn & Jones, 1993). Moreover, the knowledge elements can include likelihood information, as in the rules of a probabilistic context-free grammar. In such frameworks, explanations can have the same organization as in logical ones (e.g., proof trees), but they attach probabilities to constituents. Knowledge structures may also have a causal interpretation, which can be either deterministic (e.g., a broken wire leads a starter to fail) or stochastic (e.g., a loose wire sometimes causes failure). Accounts that focus on an agent's behavior may be teleological in that they refer to the goals that guide its decisions and actions (e.g., Meadows, Langley, & Emery, 2014). Other explanations involve predictable patterns that lack further justification; many social norms and conventions (e.g., expected behavior in churches or restaurants) take this form.

Finally, facts can play two distinct roles in explanatory structures, as Langley and Meadows (2019) have noted. In *derivational* explanations, observations serve as root nodes in a set of connected proof trees, while rule instances or other instantiated knowledge structures show how they follow from other facts and assumptions. Many scientific explanations adopt this scheme, as do causal diagnoses and teleological plans. In *associative* explanations, observed beliefs appear only as terminal nodes, which let one deduce new beliefs that follow from these facts. Such accounts use instantiated knowledge structures to connect observations to each other, but not to derive them. Parse trees for sentences are classic instances of this paradigm, but script-based interpretations of stories also illustrate the idea. This distinction is less relevant to self explanations, our focus here, as agents have access to their reasoning chains, but some (e.g., plans) have a hierarchical or derivational structure, whereas others (e.g., schedules) are relational but nonhierarchical.

## *Varieties of Self Explanation*

With these points in mind, we can now examine three forms of self explanation[2] and how they differ. Efforts to develop new AI functionality often start with a cognitive task analysis that identifies component abilities. Elsewhere (Langley, 2019) I have proposed four such abilities that underlie explainable agency:

- *Generating decision-making content*. When carrying out problem solving, the agent must consider different candidate solutions, evaluate them, and select which ones to pursue.
- *Indexing generated content*. When making decisions, the agent must store and index details about its choices in an episodic memory or similar repository.
- *Retrieving stored content*. After it has solved a problem, the agent must transform questions into cues that let it retrieve relevant information from this memory.
- *Transmitting retrieved content*. Once it has retrieved this information, the agent must translate the results into an understandable form and convey it to others.

All approaches to explainable agency must draw on their generated content, which in turn influences their downstream processing. Thus, it makes sense to discuss in some detail not the mechanisms involved in the first stage of processing, but instead the results they produce.

### *Structural Explanations*

One form of self explanation – *structural* – clarifies how a collection of steps is *rational* in Newell's (1982) sense that an agent believes they could help achieve its goals. For instance, a plan incorporates a sequence of actions that, if carried out, should produce an end state that satisfies some goal description while not violating any known constraints. Thus, a route for driving must include contiguous segments from the starting point to the target destination. The explanatory structure shows how the steps link the goals or query to the initial situation through knowledge: it focuses on the *means* of achieving objectives. We can specify the generic task of explaining the qualitative structure of a problem solution in terms of inputs and outputs:

- *Given*: A solution to a problem that specifies steps linking the initial state to the goal description;
- *Given*: Domain knowledge that defines the problem space in which the agent sought solutions;
- *Given*: A query about whether or why the candidate is acceptable or about the role played by given steps;
- *Produce*: An explanation for why the candidate is or is not acceptable or how given steps aid the solution.

Structural explanations need not focus on successful solutions; they can also clarify why a candidate does not resolve the problem. Note that this formulation does not mention how the agent generated its reasoning chain and concerns only its logical or causal structure.

The details of a structural explanation depend on the problem-solving strategy that generates it. For example, many planners find a sequence of actions that transform the initial state into one that satisfies the goal description, with each step moving closer to the objective. Other systems create partial-order plans that specify which actions must occur before others and which do not, giving a finer-grained analysis of

[2]Another important variety addresses how the agent revised a plan during execution because unexpected events occurred.

causal dependencies. Deductive proofs specify how a conclusion follows logically from a set of given facts through chains of inference steps. Each explanation type describes structural dependences among their elements and each has a recursive character in which subgraphs are themselves explanations. Storage happens during construction, with the causal or logical links serving as building blocks.

The character of structural explanations has implications for later stages of processing. This lets the agent answer questions like *Why did you take action A?*, *How did you achieve goal G*?, and *Why did you do A before B?*, but requires appropriate indexing, retrieval, and transmission.[3] For instance, given a partial order plan, one might index actions by the goals or subgoals they achieve and by their matched conditions. When asked a question about the role an action plays in a given plan, the agent translates the query into a retrieval cue, maps it to an appropriate index, and returns the retrieved structure. Finally, the transmission process converts this content into natural language, a diagram, or other format to provide an answer. This may invoke templates associated with different question types and instantiate them as needed, producing a response like *I turned left from Main onto Campus so I would be heading north on Campus*.

The AI literature includes some relevant research on these topics. For instance, work on analogical planning (e.g., Jones & Langley, 2005; Veloso et al., 1995) has addressed generation, storage, and retrieval, but not their use for self explanation. Some expert systems recorded their reasoning and played them back on request (Clancey, 1983; Swartout et al., 1991), while Johnson (1994) and van Lent et al. (2004) developed agents that recorded their decisions during execution of military missions and later answered questions about their reasoning, including what they would have done in counterfactual scenarios. In other work, Bench-Capon and Dunne (2007) adapted computational models of argument to explain how alternative conclusions are supported or contradicted by available evidence, whereas Briggs and Scheutz (2015) reported an interactive robot that gives five types of reasons why it cannot carry out a task.

### *Preference Explanations*

A second form of self explanation focuses on the *desirability* of solutions that an agent's finds, without concern for their internal structures. This is especially relevant for tasks like route finding and job scheduling that have many possible solutions, some of which are more desirable than others. We can state the task of explaining such preferences more precisely in terms of inputs required and the outputs it produces:

- *Given*: A set of solutions that the agent has generated for some decision-making task;
- *Given*: Domain knowledge that defines a problem space of candidate solutions and their quality;
- *Given*: A query about why the agent ranks a given solution above other candidates;
- *Produce*: An explanation for why the agent prefers that solution over alternatives.

This activity is quite different from explaining how the component steps of a plan or derivation achieve some goal. Rather, it more closely resembles the task addressed by recommender systems, which often produce a ranked list of candidates for users to consider.

[3]In this chapter, I focus on indexing and retrieval of elements for a specified task, rather than dealing with cases in which the agent must access structures from a memory that stores results for many distinct problems.

The distinction between structural and preference explanations is not a matter of granularity, but whether one cares about *means* of reaching results or about their overall *quality*. To clarify this point, consider a travel planner that finds multiple routes for reaching some target location. A structural account would store, for each route, the road segments and turns that lead from the start to end point, including how each step enables the next one. In contrast, a preference explanation would describe each candidate route in terms of driving distance, number of traffic lights, or other global characteristics. When multiple criteria come into play, preference accounts clarify their relative importance and how decisions resolve tradeoffs. They may also specify why a candidate's score did not exceed an acceptability threshold.

The details of this self-explanation ability will depend on how the agent's scoring and ranking process operates. One common method uses a linear utility function that computes each candidate's score on $k$ features, multiplies each score by a weight, and calculates a weighted sum, then orders candidates by this total. A second scheme uses a lexicographic function, which orders attributes by importance. Candidates are partitioned based on scores for the initial attribute, then ranked within these sets based on the second attribute, and so forth, much as words in a dictionary. A third alternative relies on preference rules that rank some candidates as better than others, without assigning numeric scores, to give a partial ordering over them.

Preference explanations support different types of questions than structural accounts. These include queries like *Why did you prefer solution X to solution Y?*, *How did X compare to Y on criterion C?*, and *Why did X not appear in the solution set?*. In this case, indexing and retrieval are simple processes, as the agent can store values for individual attributes with each solution and retrieve them as needed. As before, the final transmission stage can draw on templates that specify forms of answers for alternative types of queries, although these will differ from those for structural explanations. They will also depend on whether orderings are based on a numeric evaluation function, a lexicographic scheme, or preference rules. For instance, to clarify why it favored one solution over another, the agent might unpack calculations for the two candidates, note that they tied on the first attribute but that one did better on the second, or report the rule responsible for the decision.

This emphasis on preferences does not imply that explanation must deal only with complete solution structures. For example, if a planner uses a hierarchical task network to guide its search, then a user should be able to question why it selected one subplan for a given subtask rather than an alternative. The same idea applies to a system that finds proofs using monotonic inference rules, where a user may ask why it favored one subproof over a different candidate that leads to the same intermediate conclusion. The ability to focus attention on elements of hierarchical solutions does not necessarily mean that explanations must touch on their logical structure or how they were found. Moreover, the same mechanisms for indexing, retrieving, and transmitting results can apply to any level of hierarchical explanations.

As noted above, recommender systems often rely on a learned user profile to rank candidate items like books or movies, but one can also use such profiles as heuristics to guide search on complex reasoning tasks and to rank the solutions. Rogers et al. (1999) applied this idea to route planning, drawing on a user profile, represented as weights on complete route features, to find personalized directions in a digital road map. Gervasio et al. (1999) adopted a similar approach to personalized scheduling, invoking a user profile, encoded as weights on global schedule features, to evaluate candidates and rank solutions. These two efforts

are interesting because one used best-first search through a space of partial routes, whereas the other used repair-space search through a space of complete schedules. This shows that radically different search methods can produce the same type of preference accounts.

### *Process Explanations*

The final form of self explanation focuses on the *processes* by which an agent generates its plans or other mental structures. This view revolves around the widespread assumption, which had its origins in the earliest days of artificial intelligence, that complex cognition requires heuristic search through a problem space (Newell & Simon, 1976). This posits that the recipients of explanations are interested in details about how the system carried out that search, including which alternatives it considered, why it decided to pursue some in favor of others, and even when it decided to change its mind (e.g., by deciding to backtrack).

We can specify the generic task of explaining the problem-solving processes that an agent used to make its decisions and generate it solutions as:

- *Given*: An annotated search tree that stores options considered and decisions made in problem solving;
- *Given*: Domain knowledge that defines a problem space in which the agent seeks solutions;
- *Given*: A query about why the agent considered an alternative or made a choice during problem solving;
- *Produce*: An explanation for why the agent considered that alternative or made that choice.

This task formulation is similar in spirit to the generation of think-aloud protocols (Newell & Simon, 1972), which gave early insights about human problem solving and which led directly to the creation of early AI systems. In this setting, a researcher presents a subject with some problem (e.g., a theorem to prove or a puzzle to solve), asking the subject to talk aloud as he works on it. The scientist records this verbal report, transcribes it, and analyzes it to understand the subject's thinking processes. One important difference is that our explanation task occurs after problem solving is complete.

As before, the details of process explanations differ considerably depending on the problem-solving strategy. For instance, a forward-chaining planner would store actions it considers at each state, including the successor states that would result and the order in which each was generated. The system would also retain its reasons for pursuing one option before others, as well as reasons for backtracking or declaring success. In contrast, a means-ends problem solver would record its reasons for selecting a goal on which to focus or an action on which to chain backward. Alternatively, a case-based planner would note why it favored one retrieved solution over competitors, why it took certain adaptation steps, and so forth. Even within the same framework and given the same goals, different heuristics can guide search down different paths. This means that different problem solvers can arrive at the same solutions by divergent trajectories, each of which constitutes a separate process account of the agent's decision making.

Process explanations combine elements of structural and preference accounts, the key difference being that they retain decisions about the search effort itself rather than only about solutions. As a result, they support questions like *Why did you select action A on step S?*, *How did you achieve goal G on step S?*, *Why did you prefer A over B on step S?*, and *Why did you backtrack after trying action A?*. Note that each of these refers to some point in the search process, as the agent may consider the same action or goal in different contexts. Thus, the agent must incorporate this information during indexing and retrieval in addition to the cues used for structural and preference accounts. There appears to have been little AI research on storing, retrieving,

and transmitting process explanations either during problem solving or during retrospective reports, although studies of verbal protocols (Ericsson & Simon, 1984) offer clues about the mechanisms that produce them.

The concern with traces of decision making raises the question of what counts as a legitimate process explanation. People are good at generating verbal protocols during problem solving, but they are notoriously unreliable at reproducing their reasoning later and instead often provide at least partial rationalizations. Such reconstructions are similar to accounts of external events, in that they explain incomplete memories in terms of plausible inferences over background knowledge. This form of explanation is relevant to modeling humans, but it is less defensible when developing synthetic agents, which need not suffer from the same memory limitations. For most applications, researchers can assume that process accounts are based on accurate traces based on the decision maker's actual reasoning and conclusions.

### *Hypotheses about Explanation Types*

Now that we have identified and characterized three forms of self explanations, we can ask which of them is most useful to humans who interact with intelligent agents. Some might argue that process explanations are the natural choice, as they provide more details and thus will offer greater insight into an agent's operation. Others might instead hold that structural or preference accounts are inherently superior, because people have no need to know how an intelligent system decided on its actions but will care only how it achieved the objectives how it ranked the alternative solutions.

I will not take either position, but instead claim that the most appropriate form of self explanation depends on its intended purpose. This argument assumes that there are different types of consumers, which leads to two hypotheses. We can state the first as:

- *Process explanations will be favored by researchers interested in the details of problem solving.*

This conjecture posits that some users care primarily about the process of finding solutions. This group includes cognitive psychologists who want to understand the ways in which an intelligent system mimics, or fails to mimic, a human problem solver. Yet it also includes many AI researchers who are concerned with the detailed operation of their AI systems, both for debugging purposes and for improving the effectiveness of their search mechanisms.

However, not all people who interact with intelligent systems will care about detailed traces of their problem-solving behavior. This suggests a second conjecture, which we can state as:

- *Structural and process explanations will be favored by users interested in outcomes of problem solving.*

This group includes end users of autonomous agents who had no role in their development. These are analogous to people who use recommender systems but have little idea how they operate, but who still want to know why one option was ranked as better than another. But it will also include AI researchers, and even psychologists, who are concerned more with the correctness of solutions and the criteria used to evaluate them than with the mechanisms used to find them. Preference accounts are likely to be more useful on tasks that involve many solutions of differing quality.

### *Normative Agency*

Explainable agency is linked to the pursuit of goals, but not all goals are egocentric, which requires us to take a slight detour, as humans 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, and we would like intelligent agents to behave in similar ways. We will say that:

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

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 other vehicles and moving over to let faster ones pass. At the same time, we want them to realize that norms may come into conflict and they may need to favor some at the expense of others.

Consider a scenario in which a driver takes a friend with a ruptured appendix to the hospital. He exceeds the speed limit, weaves in and out of traffic, slows for red lights but then runs them, and even drives briefly on a sidewalk, although he is still careful to avoid hitting other cars or losing control on turns. The driver takes these drastic actions because he thinks the passenger's life is in danger, so reaching medical treatment rapidly is more important than being polite to others along the way or obeying routine traffic laws. This example of normative agency illustrates that societal norms can conflict with each other and thus requires reasoning about tradeoffs. The scenario also reminds us that driving is a far more complex task than simply staying on the road and avoiding collisions.

Before intelligent agents can use norms to guide behavior in such a human-like manner, 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. A second 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.

To develop human-like normative agents, the research community must address a number of open issues that deserve attention. These include extending intelligent systems 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 are 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 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 forgivable. 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 a friend's concerns or avoiding unnecessary emotional harm. These require formalisms for beliefs about others' mental states and ways to combine such constraints with domain-level concerns.

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 normative judgements and decisions, but others (Iba & Langley, 2011; Liu et al., 2013) have 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 this remains a relatively unexplored arena.

### *Justified Agency*

Although people can explain their goal-oriented activities, 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 knew the latter 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 that intelligent agents should demonstrate similar abilities and we will say that:

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

Let us return to the example of taking someone with 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 someone's life should take precedence over other factors. Personal goals come into play, such as avoiding collisions and not turning over, but they also support this top-level normative aim.

If we want to develop justified agents of this sort, we must decide on how their justifications map onto the three forms of explanations discussed earlier. Recall that structural accounts specify how a sequence of steps leads to the agent's goals, so the natural response is to replace some egocentric goals with societal ones. Many societal norms specify actions or states that the agent should avoid while achieving its aims, but we can encode these in much the same way as trajectory constraints in AI planning systems. Preference explanations specify the overall qualities of problem solutions, values of their constituents, and how these are combined. They are relevant to scenarios that involve tradeoffs among norms, where the agent must balance societal aims against each other or against its own. Process accounts that describe the course of the agent's decision making, including structural relations and preferences, can also incorporate social norms.

Thus, initial analysis suggests there are no serious obstacles to adapting the three types of self explanation to include norms in support of justified activities. When generating, evaluating, and storing plans, a justified agent must encode, consider, and record not only its personal goals but also social concerns. Some justifications will treat norms as hard constraints that forced the agent to carry out some actions and avoid others, but others will include reasoning about tradeoffs that arose when norms came into conflict. 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. This leads to another hypothesis:

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

In other words, once we have developed the representations and mechanisms to support the first two abilities, we will need no additional structures or processes to let agents justify their activities in normative terms. If we simply augment our goals and preferences with similar encodings of social mores, then we will obtain justified agency with no extra effort. This means that developing agents with the ability to justify their behavior will not be as difficult as it first appears.

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 & 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.

One can also ask which form of self explanation is more relevant to settings that require justified agency. We have already seen that social norms can appear, in different guises, in structural, preference, and process accounts. However, the most challenging instances of justified agency in humans involve conflicts and tradeoffs among norms. These are the mainstay of moral dilemmas studied by philosophers, but they also occur in legal cases and everyday life. The centrality of tradeoffs suggests that preference explanations will play the most important role in justified agency, but we must develop intelligent systems that communicate their reasoning about social norms to test this conjecture.

## *Concluding Remarks*

In this chapter, I defined the notion of explainable agents, which convey the reasons behind their decisions and actions. I also distinguished among three varieties of self explanation – structural, preference, and process – that store different types of content and I hypothesized when each of them is likely to be most useful. In each case, I examined how these accounts might be encoded, along with their implications for indexing, retrieval, and transmission. After this, I introduced the idea of normative agents, which attempt to follow societal maxims, and justified agents, which explain their decisions and activities in terms of those norms, along with a conjecture that joining explainable and normative agency will enable justified agency with no additional effort.

The theoretical analysis that I offered for explainable, normative, and justified agency is far from complete, but it suggests clear avenues for how to elaborate it. Researchers interested in the topic should develop architectures for agents that support all three types of self explanation, develop normative agents that guide their decisions by knowledge about social norms, and combine these elements to produce justified agents. They should demonstrate and evaluate these agents' ability to plan and act in complex domains (e.g., in urban driving simulations), to take into account laws, customs, and moral tenets when making decisions in these settings, and answer questions about the reasons for these decisions. Undoubtedly, these efforts will encounter unexpected obstacles that reveal new challenges, but they will take us closer to understanding the structures and processes needed to replicate explainable agency in humans.

## *Acknowledgements*

This chapter incorporates and elaborates on content from previous publications, including Langley, Meadows, Sridharan, and Choi (2017) and Langley (2019a, 2019b, 2020). The analysis was supported by AFOSR Grant FA9550-20-1-0130 and by Grant N00014-20-1-2643 from the Office of Naval Research, neither of which are responsible for its contents. I owe thanks to many colleagues – especially David Aha, Dongkyu Choi, Ben Meadows, and Mohan Sridharan – for discussions that led to these ideas about explainable agency.

## *References*

Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, *21*, 56–63.

Bello, P., & Bridewell, W. (2017). There is no agency without attention. *AI Magazine, 38*, 27–33.

Bench-Capon, T., & 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., & Scheutz, M. (2015). "Sorry, I can't do that:" Developing mechanisms to appropriately reject directives in human-robot interactions. In *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.

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

Dehghani, M., Tomai, E., Forbus, K., & Klenk, M. (2008). An integrated reasoning approach to moral decision making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (pp. 1280-1286). Menlo Park, CA: AAAI Press.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Fox, M., Long, D., & Magazzeni, D. (2017). Explainable planning. *Proceedings of the IJCAI-17 Workshop on Explainable AI* (pp. 24–30). Melbourne.

Gervasio, M. T., Iba, W., & Langley, P. (1999). Learning user evaluation functions for adaptive scheduling assistance. *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 152–161). Bled, Slovenia: Morgan Kaufmann.

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

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

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

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

Langley, P. (2019a). Explainable, normative, and justified agency. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 9775-9779). Honolulu, HI: AAAI Press.

Langley, P. (2019b). Varieties of explainable agency. In *Proceedings of the Second ICAPS Workshop on Explainable Planning*. Berkeley, CA.

Langley, P. (2020). Explanation in cognitive systems. *Advances in Cognitive Systems, 9*, 3–12.

Langley, P., & Meadows, B. (2019). Heuristic construction of explanations through associative abduction. *Advances in Cognitive Systems, 8*, 93–112.

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

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

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

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

Meadows, B., Langley, P., & Emery, M. (2014). An abductive approach to understanding social interactions. *Advances in Cognitive Systems, 3*, 87–106.

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

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87–127.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ng, H. T. & Mooney, R. J. (1990). On the role of coherence in abductive explanation. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 337–342). Cambridge, MA: AAAI Press.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence, 32*, 57–95.

Rogers, S., Fiechter, C., & Langley, P. (1999). An adaptive interactive agent for route advice. *Proceedings of the Third International Conference on Autonomous Agents* (pp. 198–205). Seattle: ACM Press.

Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning, 2,* 247–276.

Smith, D. E. (2012). Planning as an iterative process. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 2180–2185). Toronto: AAAI Press.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76–105.

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

VanLehn, K., & Jones, R. M. (1993). Integration of analogical search control and explanation-based learning of correctness. In S. Minton (Ed.), *Machine learning methods for planning*. San Mateo, CA: Morgan Kaufman.

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

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

Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., & Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. *Proceedings of the 2017 International Conference on Robotics and Automation* (pp. 1313–1320). Singapore.