# Explainable, Normative, and Justified Agency

**Pat Langley**

Institute for the Study of
Learning and Expertise

Center for Design Research
Stanford University

*http://www.isle.org/~langley/*

This research was supported by ONR Grant N00014-15-1-2424 and by AFOSR Grant FA9550-20-1-0130. Thanks to M. Barley, D. Choi, B. Meadows, M. Sridharan, and P. Stone for discussions.

# Prevalence of Autonomous Agents

Autonomous artifacts are becoming ever more widely deployed in the form of:

- Self-driving cars
- Delivery drones
- Military robots

Before such systems can gain widespread acceptance, they must be able to:

- Explain their behavior in understandable terms;
- Follow the laws, customs, and morals of society.

They must move beyond ***explainable agency*** to ***justified agency***.

# Driving to the Hospital

Suppose that Dan drives a friend, Eve, with a ruptured appendix to the hospital. On the way, he:

- Exceeds the speed limit
- Weaves in and out of traffic
- Slows at red lights but runs them
- Detours briefly onto a sidewalk
- Yet retains control and avoids collisions

Dan later defends his actions because Eve's life was in danger, so reaching the hospital was more important than traffic laws.

In doing so, he has illustrated the notion of justified agency.

# Two Senses of Explanation

We can distinguish between two uses of the term ‘explanation’ in English:

- A ***structure*** – mental, written, or spoken – that elucidates some phenomena or behaviors.
  - *E.g., a scientific account of pulsar variation or the clarification of a home-buying decision.*
- The ***process*** or ***activity*** that generates an explanatory structure of this sort.
  - *E.g., a scientist or buyer engages in the explanation of pulsar behavior or a purchase decision.*

We will use both senses in this talk, with meaning being clear from the context.

# Two *More* Senses of Explanation

We can further differentiate between two additional connotations of the term:

- ***Interpretive explanation*:** Construction of accounts for observed situations or events.
  - *E.g., a geologist posits processes that have created a landform, a mechanic hypotheses why a car does not start.*
- ***Communicative explanation:*** Conveying existing accounts to another agent.
  - *E.g., the geologist presents a talk about his model, the mechanic includes diagnostic notes on her estimate.*

The second applies not only to sharing accounts for external events, but to *communicating about plans and actions*.

# Representing Explanations

Whether an agent constructs or communicates an explanation, it must first represent its content.

Explanations always refer to *existing knowledge* and adopt a common form that includes:

- A set of *given / observed facts*, with optional goals / queries;
- A set of *instantiated knowledge elements* linking these facts;
- *Annotations* that modulate the entire structure or its elements.

Knowledge elements may be logical formulae, causal chains, numeric equations, or teleological relations.

These generic elements may be handcrafted or learned, but the explanations themselves are generated automatically.

# Explainable Agency

When we make a decision, we can often explain the choices we considered and why we selected one over others.

Definition:

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

Examples of explainable agency:

- Why did you prefer driving route A to work over route B?
  - Route A had fewer traffic signals and it was still pretty short.
- Why did you swerve suddenly into the next lane?
  - It was the only way to avoid hitting a fallen tree limb.

Explainable agency relies on a form of *communicative explanation*.

# Prior Research on Explainable Agency

There have been important studies of explainable agency in a number of settings:

- Diagnostic systems (Clancey, 1983; Swartout & Moore, 1993)
- Reactive execution (Johnson, 1994; van Lent et al., 2004)
- Personalized services (Rogers et al., 1998; Gervasio et al., 1999)
- Plan generation (Smith, 2012; Fox et al., 2017; Zhang et al., 2017)
- Robot planning / control (Colaco & Sridharan, 2015)

However, each effort emphasized one type of explanation while ignoring others.

# Three Types of Explanatory Content

Agents can provide different types of explanation about their problem-solving decisions:

- ***Structural accounts*** – How the component steps lead to goals
  - *E.g., a route must traverse each segment to reach a destination*
  - *This corresponds to Newell's (1980) definition of rationality*
- ***Preference accounts*** – Why some solutions are more desirable
  - *E.g., one route is shorter than another and has fewer turns*
- ***Process accounts*** – Search carried out when finding solutions
  - *E.g., considers a partial route but backtracks because of a toll*

Other forms of communicative explanation include why a given candidate is nonviable (e.g., violates constraints).

# Two Modes for Explanatory Agency

There are two primary ways an agent can communicate about its decision making:

- ***Reporting summaries*** of structures, preferences, or search
  - *E.g., presenting the turns and segments of a driving plan*
  - *E.g., summarizing actions in execution of a military mission*
- ***Answering questions*** about structures, preferences, or search
  - *E.g., which routes it considered, which one it selected, and why*
  - *E.g., what plan it executed, how it responded to surprises*

Dealing with questions is more difficult, as they may focus on particular elements and may address any previous problem.

# Component Abilities for Question Answering

To answer questions about its decision making, an explainable agent must:

- ***Store*** and ***index*** its solution traces in episodic memory
  - Linking their constituents separately for access later
- ***Extract*** cues from queries and ***retrieve*** content from memory
  - Accessing only solution elements relevant to questions
- ***Translate*** retrieved elements into an understandable form
  - Sharing the transformed result with the questioner

Answers may use language, graphics, or formal notations and should offer no more detail than necessary.

# Rationales vs. Rationalizations

The importance of retrieval raises the issue of what counts as legitimate communicative explanation.

- People can produce verbal protocols *during problem solving*.
  - They can access many aspects of the process while it happens.
- They are unreliable at reproducing their reasoning *after the fact*.
  - Imperfect storage or indexing means that traces are incomplete.

Retrospective reports rely on *reconstructive memory*, which has much in common with *interpretative explanation*.

Such rationalization is relevant to modeling human behavior, but it is less defensible for artificial agents.

# Normative Agency

Humans are driven by goals, but they must also operate within their society's norms.

Definition:

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

Examples of normative agency:

- Paying for food rather than stealing it
- Saluting to a superior officer
- Waiting in line rather than cutting ahead
- Recycling to help the environment

These all canalize people's behavior in certain directions.

# Varieties of Normative Agency

We want our intelligent agents to follow different types of social norms, including:

- ***Formal laws*** (e.g., obey traffic signals)
- ***Military orders*** (e.g., get up at reveille)
- ***Informal customs*** (e.g., Pittsburgh left turn)
- ***Moral tenets*** (e.g., favor life over property)

Society states few of its norms explicitly, but that makes them no less important.

Neither does it mean we cannot encode them as knowledge over which agents can reason.

# Prior Research on Normative Agency

The field has seen substantial research on normative reasoning in the context of:

- Legal knowledge (e.g., Gardner, 1987; Branting, 2000; Rissland et al., 2003; Ashley, 2017)
- Moral tenets (e.g., McLaren, 2005; Mikhail, 2007; Deghani et al., 2008; Iba & Langley, 2011; Malle et al., 2015)

However, this work has focused on judgement of other agents' behaviors, not plan generation / execution with social norms.

Naturally, laws and conventions are embedded in self-driving vehicles, but treatments to date have been shallow.

# Representing Social Norms

Before they can use social norms, agents must represent their content in terms of:

- ***Deontological*** or ***consequentialist*** approaches
  - What actions to take vs. what states are desirable
- ***Qualitative*** relations or ***quantitative*** criteria
  - Symbolic rules vs. numeric functions
- ***Prescriptions*** or ***proscriptions***
  - What is required vs. what is forbidden

Fortunately, none of these dichotomies are mutually exclusive; a unified framework can include them all.

# Challenges for Normative Agency

Social norms raise a number of issues related to representation and problem solving:

- Dealing with ***conflicting norms***
  - *E.g., legality vs. safety, authority vs. morality*
- ***Mitigating factors*** that modulate norms
  - *E.g., aggravated assault, self defense*
- Reasoning about ***others' mental states***
  - *E.g., telling white lies, the golden rule*

Once addressed, we can incorporate social norms into available methods for plan generation / execution.

# Justified Agency

When we make a decision, we can often state the choices we considered and how norms influenced our selection.

Definition:

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

Examples of justified agency:

- Stealing food to help a starving child (and explaining why)
- Disobeying an order that you consider illegal (and . . .)
- Cutting in line to avoid missing a flight (and . . .)
- Breaking traffic laws for a medical emergency

Note that all of these instances concern ***conflicting norms***.

# Enabling Justified Agency

To implement intelligent systems that exhibit justified agency as defined earlier, we must:

- ***Encode social norms*** as goals, values, and constraints
- Use this knowledge to ***generate / execute plans***
- ***Store / index*** the products in an episodic memory
- ***Retrieve*** relevant content in response to questions
- ***Communicate*** the retrieved answers to questioner

Social norms will figure centrally in the resulting explanations and thus produce justified agency.

# Two Hypotheses for Justified Agency

Analysis of these abilities suggests two plausible conjectures:

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

Treating social norms in terms of goals, values, and constraints should give this ability with no extra effort.

- ***Preference explanations*** *will play the most important role in demonstrating justified agency.*

This is because the most interesting cases involve conflicting norms and tradeoffs among them will be key to justification.

Either hypothesis may be false: both agency or norms may be more complex than we envision.

# Future Work on Justified Agency

We have offered a theoretical framework for justified agency, but we must still:

- Develop architectures for explainable agents (e.g., planners)
- Encode knowledge about social norms (e.g., for urban driving)
- Combine these components to produce justified agents
- Demonstrate / evaluate their behavior (e.g., in driving simulators)

The AI research community should pursue this agenda using:

- Different *forms of knowledge* (both handcrafted and learned)
- On distinct *problem types* (e.g., plan generation, plan execution)

These will clarify the nature of this important cognitive ability.

# The Main Points

In this talk, we defined and then examined three related ideas:

- *Explainable agents* – convey reasons behind their actions
- *Normative agents* – attempt to follow societal norms
- *Justified agents* – explain activities using such norms

We discussed challenges for each ability and how to combine them, producing two conjectures:

- *Merging explanation and social norms gives justified agency; which in turn relies centrally on preference accounts.*

But both hypotheses need testing with controlled experiments in mission-oriented testbeds.

# Concluding Remarks

Designing and constructing justifiable agents is an important step toward replicating the full range of human intelligence.

The ultimate demonstrations of such autonomous artifacts would be:

- Self-driving cars that sway judges in traffic court
- Police drones that defend themselves in civil suits
- Military robots that win court martials for actions in combat

We encourage other AI researchers to pursue this audacious vision of explainable, normative, and justified agency.

# References


- Meadows, B., Langley, P., & Emery, M. (2014). An abductive approach to understanding social interactions. *Advances in Cognitive Systems*, *3*, 87–106.
- 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.
- Langley, P. (2018). Planning systems and human problem solving. *Advances in Cognitive Systems*, *7*, 13–22.
- Langley, P. (2019). Explainable, normative, and justified agency. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 9775–9779). Honolulu, HI: AAAI Press.
- Langley, P. (2020). Explanation in cognitive systems. *Advances in Cognitive Systems*, *9*, 3–12.

*I, Robot (1964)*
*The Outer Limits*

End of Presentation