Explainable, Normative, and Justified Agency

Pat Langley

Institute for the Study of
Learning and Expertise
Palo Alto, California

This work was supported by ONR Grants N00014-15-1-2517 and N00014-15-1-2424.
Thanks to M. Barley, D. Choi, B. Meadows, M. Sridharan, and P. Stone for discussions.

A Motivating Example

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.

We will say that, in this scenario, Dan exhibits ***justified agency***.

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 first be able to:

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

We claim that ***both*** abilities are required for justified agency.

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.

Facets of Explainable Agency

An explainable agent should be able to answer questions about:

- Alternatives that it considered / selected (e.g., routes, lanes)
- Reasons selected / criteria used (e.g., shorter, less crowded)
- What it would do in another situation (e.g., near ambulance)

The agent should be able to answer questions about both its *generation* and *execution* of plans.

Previous research:

- Explainable expert systems (Swartout, 1991)
- Explainable reactive execution (Johnson, 1994; van Lent, 2004)
- Explainable planning (Smith, 2012; Fox et al., 2017)

Design Decisions for Explainable Agency

Design issues for devising explainable agency include:

- Rationalizations vs. actual reasons
- Individual actions vs. entire plans
- Symbolic goals vs. numeric functions

System designs should specify how the agent will:

- Create content during planning / execution
- Store and index content in episodic memory
- Retrieve content in response to questions
- Communicate this content to answer queries

We need substantially more research on these components.

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.

Facets of Normative Agency

A normative agent's behavior should take into account:

- 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)

Different norms may conflict, so that the agent must handle tradeoffs among them.

Previous research:

- Legal reasoning (e.g., Branting, 2000)
- Moral reasoning (e.g., McLaren, 2005; Deghani, 2008; Mikhail, 2007; Iba & Langley, 2011; Malle et al., 2015)

Design Decisions for Normative Agency

Design issues for devising normative agency include:

- Actions (deontic) vs. states (consequentialist)
- Symbolic rules vs. numeric functions
- Prescriptions vs. proscriptions

System designs should specify how the agent handles:

- Tradeoffs among conflicting norms
- Mitigating factors that modulate acceptability
- Domain-independent norms (e.g., others' mental states)

We need substantially more research on these components.

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

Justified agency is most interesting when norms ***conflict***.

The Character of Justified Agency

Three major design issues arise in devising justified agency:

- Generating, storing, and using explanations
- Encoding, using, and combining norms
- How to integrate explanations with social norms

We have considered the first two issues, but what of the third?

Consider a plausible hypothesis:

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

If we include social norms in our agent's goals and values, then we get justified agency with no extra effort.

Testing the Hypothesis

The hypothesis does ***not*** follow logically from our definitions.

- Justified agency requires the ability to explain decisions and reason about norms, but they may not be sufficient.
  - *Agency may be more complex than assumed*
  - *Norms may demand richer forms of explanation*

To test it, we must construct explainable and normative agents, combine them, and measure their ability to justify.

Simulated testbeds for urban driving, robotic rescue, and other mission-oriented settings support such studies.

The Main Message

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 also examined design issues for each type of agent and noted component functions that require further work.

Analysis suggested any system that supports both explanation and normative behavior ***will also*** exhibit justified agency.

But this empirical hypothesis requires 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.

End of Presentation