Toward a formal pragmatics of explanation

Jacqueline Harding, Tobias Gerstenberg, Thomas Icard

Motivation

- Wanted: a formal model of explanation that meets classical desiderata and makes testable predictions about explanatory judgments
- Key idea: most interesting features of explanation arise from a combination of conversational pragmatics and ordinary causal cognition

Existing formal accounts

- Hempel & Oppenheim 1948: derivations of observations from "laws" rendering them retrodictable ('deductive-nomological' model)
- Explanation in terms of *statistical relevance* (Salmon 1970) or tracing *causal processes* along space-time (Salmon 1984; Dowe 2000)
- Gärdenfors 1980, 1988, 1990: Explaining why FACT is a matter of specifying probabilistic or factual information that would have—in the agent's epistemic state $\mathcal{K}$ prior to learning FACT—rendered it less surprising.
- Halpern & Pearl 2005: Causal version of Gärfenfors account. $\mathbf{X} = \mathbf{x}$ is an explanation of FACT relative to knowledge state $\mathcal{K}$ iff:

  (EX1) FACT is true at all models $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$.

  (EX2) $\mathbf{X} = \mathbf{x}$ is an actual cause of FACT for all $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$ in which $\mathbf{X} = \mathbf{x}$ is true.

  (EX3) No proper subset of $\mathbf{X}$ satisfies **EX2**.

  (EX4) There exists $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$ in which $\mathbf{X} \neq \mathbf{x}$.

EX2 just says that explanations should cite actual causes; empirically, **EX1**, **EX3**, and **EX4** all admit of counterexamples.

A Communication-First Alternative

- Make the communicative context—including not just the 'listener', but also the speaker S who produces the explanation—explicit.
- Couched in framework of Rational Speech Acts (Frank & Goodman 2012; Sumers et al. 2023; etc), beginning with a 'literal' listener:

$$P_{L0}(\mathcal{M}, \mathbf{u} \mid \text{"FACT because } \mathbf{X} = \mathbf{x}\text{"}) \quad \propto \quad \text{Prior}(\mathcal{M}, \mathbf{u}) \cdot \mathbf{1}_{\mathcal{M},\mathbf{u} \models \text{"FACT because } \mathbf{X}=\mathbf{x}\text{"}}.$$

  $L0$ assigns probabilities $\text{Prior}(\mathcal{M}, \mathbf{u})$ to causal situations; "FACT because $\mathbf{X} = \mathbf{x}$" literally means $\mathbf{X} = \mathbf{x}$ is an actual cause of FACT.

- Listener faces decision problem $R : \mathcal{A} \times \mathcal{K} \to \mathbb{R}$ ($\mathcal{A}$ some set of actions); upon receiving message m follows policy π, with reward U:

$$\pi(a \mid m) \quad \propto \quad \exp\left(\beta_L \cdot \left[\sum_{(\mathcal{M},\mathbf{u}) \in \mathcal{K}} P_{L_0}(\mathcal{M}, \mathbf{u} \mid m) \cdot R(a, \mathcal{M}, \mathbf{u})\right]\right) \qquad U(m, \mathcal{M}, \mathbf{u}) \quad = \quad \sum_{a \in \mathcal{A}} \pi(a \mid m) \cdot R(a, \mathcal{M}, \mathbf{u}).$$

- Pragmatic speaker S chooses explanation m to help with L's decision problem, respecting processing costs (for both L and S herself):

$$P_S(m \mid \mathcal{M}, \mathbf{u}) \quad \propto \quad \exp\Big(\beta_S \cdot \big[U(m, \mathcal{M}, \mathbf{u}) - \text{Cost}(m)\big]\Big).$$

- The pragmatic listener responds appropriately, with overall goodness a function of the listener's (expected) increase in expected utility:

$$P_L(\mathcal{M}, \mathbf{u} \mid m) \;\propto\; \text{Prior}(\mathcal{M}, \mathbf{u}) \cdot P_S(m \mid \mathcal{M}, \mathbf{u}) \qquad \text{Goodness}(m, \mathcal{M}, \mathbf{u}) \;=\; \sum_a \pi(a \mid m) \cdot R(a, \mathcal{M}, \mathbf{u}) - \sum_a \pi_{\text{Prior}}(a \mid m) \cdot R(a, \mathcal{M}, \mathbf{u}).$$

Manipulation Games

Generic decision problem $(\mathcal{A}, \mathcal{R})$ capturing 'manipulation & control':

- $\mathcal{A}$ the set of endogenous variables outside FACT.
- For endogenous $X \in \mathcal{A}$ and possibility $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$,

$$\mathcal{R}(X, \mathcal{M}, \mathbf{u}) \;=\; \sum_{\mathbf{u}'} P(\mathbf{u}') \cdot \text{Manipulates}(X, \text{FACT} \mid \mathcal{M}, \mathbf{u}'),$$

  $\text{Manipulates}(X, \text{FACT} \mid \mathcal{M}, \mathbf{u}') = 1$ iff some x flips value of FACT.

Connection to **causal strength measures**: with binary variables this gives ΔP measure (Jenkins & Ward 1965; Cheng & Novick 1992). Replacing $P(\mathbf{u}')$ with sampling procedure in (Lucas & Kemp 2015) gives the *counterfactual effect size model* (Quillien & Lucas 2023).

Explaining Explanation?

The account derives the following 'explanatory virtues':

- Sensitivity to the **downstream interests** of the listener.
- Appropriateness to the listener's **background knowledge**.
- Identification of explanatory relationships which are **invariant** across background conditions.
- Account for **influence of norms** on explanatory judgments.
- Preference for **simpler** theories or more **'minimal'** explanantia.
- Situation at the 'right' level of **abstraction**.
- 'Soft' versions of Halpern & Pearl's $\mathbf{EX1}$, $\mathbf{EX3}$, and $\mathbf{EX4}$.