Explanations Meet Decision Theory

Jack Beasley, jbeasley@stanford.edu

October 22nd, 2020

Note: I changed topics late in the quarter after my proposal. My advisor (Thomas Icard) and I found a new research direction that I noticed overlapped well with this course, so I decided to get that project started with this class project instead of going with my old power optimization project.

Introduction

The goal of this paper is to start to bridge the gap between engineering decision theory and theoretical work on explanations in epistemology. Given the very different aims of these two communities, there is little to no cross-pollination between the two, however, there are ample opportunities for connections given that both center around inference and justified belief to varying extents.

In this paper, I plan to first introduce the philosophical issues involved in inference, critically including *abduction* or *Inference to the Best Explanation* (IBE). Once these philosophical concepts are out of the way, I'll introduce some of the simulation-based experiments philosophers have proposed as evidence for inference by IBE or for inference according to Bayes rule. I'll then frame one of these experiments in terms of a POMDP and show how many of these differences in performance between these belief update rules can be matched with optimal policies over bayesian update rules given different reward structures. Essentially, I use the POMDP framing to make the model more representative of real-world situations and to provide a way to quickly evaluate and explore how different belief updating rules, policies and reward functions affect the problem of explanationism. Thus, this paper is best viewed as an exploratory project that seeks to present POMDPs as a good lens through which to study.

Background and Literature Review

Inference in Philosophy

There are three distinct types of inference that often show up in the philosophy of science and epistemology literature: deduction, abduction and induction. To motivate these different forms of reasoning, I'll provide three election-themed inferences that we can draw in each different form.

First, define a simple popular vote election between two candidates where the candidates with more votes wins the election. Say we then know the following premiss about candidate A and candidate B:

1. The candidate with more votes wins the election.
2. Candidate A won more votes than candidate B.

From these two facts, we can conclude *with certainty* that Candidate A won the election. This is the reasoning lets us derive truths through proof in math, logic or any other axiomatizable system, however, it doesn't help us much when we don't know the axioms for certain, which is common in the real world. In philosophy, these sorts of inferences are lead to *a priori* truths, or facts that don't require observation to know.

Thus, we turn to a famously shaky type of inference: induction. While famously criticized by Hume, induction is the process of observing the world and drawing inferences from prior observations. To see this form, and the problems with it, consider a reliably Democratic precinct in an American presidential election. Say we've observed this precinct vote for Democratic candidates for the past twenty elections. By induction, we might conclude that that same district will vote for the Democratic candidate in the next election. However, this relies on the implicit assumption that the district and the world at large doesn't change much. Usually this works out and in practice this inference is often quite accurate. However, it is not inconceivable that this inference would turn out to be wrong if people changed their minds or the people in the district changed significantly. This is a key difference with deduction as a correct deductive inference will never be wrong.

Finally, we move to abduction, which is sort of like induction in reverse. Where induction starts from a statistic or collection of observations and makes an inference from that, abduction takes observations and infers a statistic that might explain those observations best. Say you are a poll worker for a single precinct in a city and are counting ballots. You might count 80 votes for candidate A and 5 for candidate B. From these observations, you might infer that candidate A is getting the most votes in that precinct's total vote because that would explain the skewed distribution of votes that you observed. The key difference from induction is that we appeal to our sense of what a good explanation is to make our inference for abduction.

Abduction, formally

Now that we've gone over the high-level ideas behind each mode of inference, we can discuss how we might go about formalizing these modes. First, we must define what it means to "explain" in mathematical terms. Here, I'll follow from David Glass' formalisms which generally track the consensus (Glass 2012). Here we have some evidence E and n different hypotheses $H_1, \dots H_n$ that could possibly explain that evidence. The problem of choosing an explanation is then a decision problem of choosing the hypothesis H_1 that is the best explanation of

the evidence E.

This formal framing crucially begs the contentious question of how we define what a "good explanation" actually is, which is the big, interesting question that those studying abduction seek to answer. Now that the problem is clear, I'll give a high level description of the various approaches to formalizing explanation.

Simply Bayes Initial approaches follow directly from Bayes rule and define the best explanation as the explanation which maximizes posterior probability (MAP). In terms of problem specification given above:

$$Pr(H_i|E) = \frac{Pr(E|H_i)P(H_i)}{Pr(E)}$$

However, this approach is called "trivial" by many explanationists as it reduces abduction to induction as we don't really appeal to explanations explicitly in this argument. Instead, we just use induction to find which explanation is most probable, which can be seen as merely a simple appeal to probability. In essence, using Bayes theorem to derive MAP alone equates the most explanatory explanation with the least surprising one.

Next, some approach explanation in terms of likelihoods, where H_1 is better than H_2 if and only if:

$$Pr(E|H_1) > Pr(E|H_2)$$

This approach also has problems as a crazy hypothesis might make our evidence very likely. For example, a crazy conspiracy theory H_c might make $Pr(E|H_c)$ because it can explain everything, however, the hypothesis itself is so far-fetched that $P(H_c) \approx 0$ so we wouldn't call it a good explanation.

So, if we take that there is something more to explanation than MAP, how do we balance likelihood and probability? That is the key question for the explanationist and there are many competing formalisms, however nearly all of them fit into the following generalized formulation proposed by critic IBE:

$$Pr(E|H_i) = \frac{Pr(E|H_i)P(H_i) + f(H_i, E)}{\sum_j^n Pr(E|H_j)Pr(H_j) + f(H_j, E)}$$

Where f is a function that assigns a "bonus" to the best explanation by some other rule. This f can be something based on various rules including Popper's rule, Good's rule and Schupbach and Sprenger's rule (Douven and Schupbach 2015), however, I plan not to dwell too much on these rules as there isn't consensus in the literature as to which is best and the bigger point of this work is to contextualize these rules within broader decision-based formation of explanation. Thus, we need only take these rules charitably as creating well-reasoned heuristics that might update our beliefs in a way that is pragmatically better by some

measures such as being faster to converge (as argued by (Schupbach 2018)) or more reflective of how people update their own beliefs (a claim that has some experimental backing (Douven and Schupbach 2015)).

Now that we understand what abduction formalisms look like, lets consider a key argument against these "explanationist" theories that stray from Bayes rule and MAP to try to better capture what makes for a good explanation. This argument is known as the "Dutch book" argument and was originally presented by David Lewis (Lewis 1999). Essentially, the argument states that any agent that is using a non-Bayesian update rule, like the explanationist ones described above, will always face a sure loss when using that rule to place bets against a certain legal, but carefully constructed set of odds presented to you by a house which knows you are using that rule. Lewis has a detail proof of this, however, the major response is that most scenarios in which we explain have different reward structures than betting at a casino. However, this has not resulted in a formal discussion about how rewards might actually impact the explanation rules' efficacies.

IBE within a POMDP

Given that the key argument against IBE approaches hinges on a specific reward structure, expanding the scope of the problem to include rewards and decisions seems like a natural direction to better evaluate these different modes of explanation. Thus, we can frame these experiments as POMDPs where we don't ever know exactly what is true and seek to find explanations which maximize a given reward function. Many of the explanationist arguments use computer simulations to demonstrate their update rule works well, thus I'll pick one such simulation to formulate as a POMDP to show that good or bad policies over normal Bayesisan-based belief can also lead to similar differences.

One key reason for IBE, according to abduction supporters, is that abduction can result in faster convergence, which might be better in the case of deciding when to announce a discovery when researchers want to avoid getting "scooped". Douven uses a simplified experiment to illustrate this by setting up an Urn with n balls that are all either violet or green. The agent's goal is observer k draws from the urn with replacement and decide which configuration of the Urn best explains this draw (Douven 2013). This follows from prior experiments on human subjects that showed humans are not exactly Bayesian at this task.

Now, we frame that as a POMDP.

State space

The states in this game are the configurations of the Urns, of which there are n, one for each possible number of violet balls (the number of violet balls sets the number of green balls as they must add to n).

Action space

Each agent can do one of two things:

1. Observe a ball
2. Select an explanation

Observation space

When an agent draws a ball, they observe a violet ball with probability:

$$p = \frac{\#violet}{n}$$

When the agent selects an action, they get confirmation of the true value before the problem is reset.

Transition

Whenever the agent is observing balls, the state remains the same and once the agent picks a hypothesis, the configuration is shuffled.

Rewards

There are three distinct rewards given that can vary depending on what problem setting we want to model. First, there is a reward or cost to observing a ball. In science, this might reflect the fact that experiments are not free to run. Second, there is a reward for correctly picking the correct hypothesis and third there may be a penalty for selecting the wrong hypothesis. In science, this might reflect that being right in public is often rewarded and being shown to be wrong in public may have negative outcomes. For my implementation, I chose a -1 penalty for observations, a $+5$ reward for a correct explanation and a -5 penalty for a wrong explanation.

Methods

I wrote a model of the POMDP described above and used both offline and online solvers to determine how good policies might behave. Beyond the above setup, I chose uniform priors for the belief and state distributions. Because this was mainly an exploratory analysis mainly aimed at showing that this POMDP formulation can automatically generate interesting leads and problems for existing theories, I mainly focused on the simplest case of seeing how policies from existing solvers differ from the explanationist update and decision rules presented by Douvan. Thus, I used the QMDP solver (Littman, Cassandra, and Kaelbling 1995) to solve for a policy and then evaluated that policy on 100 simulated games.

Results

Interestingly, the solved policies are much more aggressive than the decision policies specified by Douvan of waiting until beliefs are above 0.9 and picking the highest belief. These policies end up with a mean reward of 40, but with wildly varying results with a minimum reward of-408.0 and a maximum reward of 470.0. This contrasts sharply with what we'd expect a human do do as a human would likely be much more risk-averse, as in Douvan's rule, even if that is not necessarily optimal. This variation alone is worthy of a second look to determine under what reward structures an aggressive policy like this might be warranted.

Why is this policy aggressive, because it sets a much lower level of confidence as a point to decide at. This policy chose the highest option at a mean belief of 0.733 achieved using standard POMDP discrete belief representations, which follow from standard bayesian rules. However, this lower threshold shows that an optimal policy over Bayesian beliefs might be able to achieve the speed of convergence of an explanationist approach by merely lowering the threshold for choosing a belief in the decision rule. This points to a possible formalization where explanations might move to be a way of better deciding between bayesian-updated beliefs when we have bad priors or not enough time to reach a confident belief through traditional means. In this world, we might find reward functions that demand a more aggressive policy or decision rule and wee if explanationist approaches help in these contexts, rather than changing the Bayesian update mechanism that seems to work so well in many contexts.

This lower threshold of belief meant that the agent made an average of 79 selections averaging 46 correct selections. This gives an accuracy of 0.58, which, while not great, allows for collecting the larger rewards more frequently. It is a high-variance, high-expectation strategy that seeks to beat the opponent to making choices rather than get beat waiting for confidence to improve.

Conclusion

This work revealed a very interesting approach to explanation in an adversarial setting, but also opens the door to further work on abduction that centers the decision-making aspects of explanation, rather than merely the epistemic issues at hand. From this initial exploration, there seems to be a good case to be made for studying the various explanation rules as decision rules that are functions of a reward landscape, rather than as a universal update rule to compete with Bayes Theorem.

References

Douven, Igor. 2013. "Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation." *The Philosophical Quarterly* 63 (252): 428–44. https://doi.org/10.1111/1467-9213.12032.

Douven, Igor, and Jonah N. Schupbach. 2015. "Probabilistic Alternatives to Bayesianism: The Case of Explanationism." *Frontiers in Psychology* 6 (April). https://doi.org/10.3389/fpsyg.2015.00459.

Glass, David H. 2012. "Inference to the Best Explanation: Does It Track Truth?" *Synthese* 185 (3): 411–27. https://doi.org/10.1007/s11229-010-9829-9.

Lewis, David K. 1999. "Why Conditionalize." In *Philosophy of Probability: Contemporary Readings*, edited by Antony Eagle, 403–7. Routledge.

Littman, Michael L., Anthony R. Cassandra, and Leslie Pack Kaelbling. 1995. "Learning Policies for Partially Observable Environments: Scaling up." In *Proceedings of the Twelfth International Conference on Machine Learning*, 362–70. Morgan Kaufmann.

Schupbach, Jonah N. 2018. *Inference to the Best Explanation, Cleaned up and Made Respectable*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198746904.003.0004.