Occam's razor is insufficient to infer the preferences of irrational agents

Stuart Armstrong and Sören Mindermann (equal contributions)

Future of Humanity Institute, Oxford University

Summary

Algorithms that infer human preferences from the human policy usually assume that the policy $\dot{\pi}$ is (noisily) rational. Since humans are *systematically* irrational, this can lead to catastrophically wrong inferences. Can we ever hope to learn human preferences on tasks that aren't trivial for humans? The only available answer is to learn the human rationality model (planner) $\dot{p}$ and jointly infer their reward function $\dot{R}$.

First, a No-Free-Lunch (NFL) result shows that this is impossible without assumptions about human rationality or reward. NFL results are usually solved by Occam's razor priors. Second, this strategy unexpectedly fails because some nonsensical interpretation of human rationality and reward are very simple.

We will need to investigate which additional assumptions humans, who are also subject to this problem, are making.

Notation

Let Π be the set of (deterministic) policies.
Define $\mathcal{R} = \{R : \mathcal{S} \times \mathcal{A} \to [-1, 1]\}$, the reward functions, $\mathcal{P} = \{p : \mathcal{R} \to \Pi\}$ the planners.

Definition

The pair (p, R) is *compatible* with $\dot{\pi}$ iff $p(R) = \dot{\pi}$.

If $p(R) = \dot{\pi}$ and the human follows $\dot{\pi}$, (p, R) predicts the human's actions perfectly. One can't update away from (p, R) by observation, so we need priors.

Theorem

For all R, and all p with $\dot{\pi} \in Im(p)$, there exists p' and R' with $p'(R) = p(R') = \dot{\pi}$. Regret from choosing the wrong (p, R) can be very large [1].

Degenerate pairs and rewards

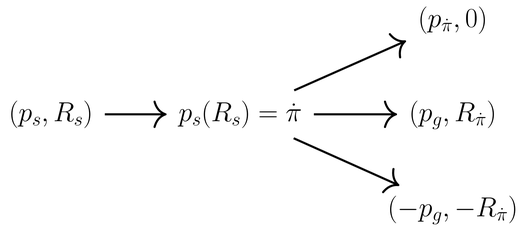
We construct 3 pairs (p, R) that are clearly undesirable, but valid decompositions of the human policy, i.e. $p(R) = \dot{\pi}$. Then we show these are programs with almost minimal description length.
Define $-p$: $-p(R) = p(-R)$. Note that $(p, R) \to (-p, -R)$ is a simple map and $-p(-R) = p(R)$. **Thus if $(\dot{p}, \dot{R})$ is compatible, so is $(-\dot{p}, -\dot{R})$.**

For $\dot{\pi}$, define $p_{\dot{\pi}}$, p_g (the greedy planner), and $R_{\dot{\pi}}$:

$$\begin{aligned} \forall R, s : p_{\dot{\pi}}(R)(s) &= \dot{\pi}(s) \\ p_g(R)(s) &= \operatorname{argmax}_a R(s, a) \\ \forall a, s : \; R_{\dot{\pi}}(s, a) &= \mathbf{1}(\dot{\pi}(s) = a) \end{aligned}$$

Note that $(p_{\dot{\pi}}, 0)$, $(p_g, R_{\dot{\pi}})$, and $(-p_g, -R_{\dot{\pi}})$ are compatible with $\dot{\pi}$. Let (p_s, R_s) be the simplest pair compatible with $\dot{\pi}$.

Then the following maps are simple programs in any reasonable language L:



Complexity of human rationality

Assume $(\dot{p}, \dot{R})$ fits with human judgement about human rationality and reward. The following imply it is of higher complexity than $\dot{\pi}$:

1. Any reasonable $(\dot{p}, \dot{R})$ is of high complexity, higher than it may intuitively seem to us [2, 3].
2. Even given $\dot{\pi}$, any reasonable $(\dot{p}, \dot{R})$ involves a high number of contingent choices:
   - i Distinct cultures have distinct rationality judgements [4].
   - ii Different individuals have distinct rationality judgements [5].
   - iii The same individual has different judgements at different times [5].
   - iv The judgement of an individual can be changed or manipulated [6].
3. Hence any given $(\dot{p}, \dot{R})$ has extra information (and thus extra complexity), even given $\dot{\pi}$. Points iii and iv prevent 'simply' using the human's own judgement of their rationality.
4. Past and present failures to find a simple $(\dot{p}, \dot{R})$ derived from $\dot{\pi}$ are evidence that this is tricky.

Simplicity of degenerate planner-rewards and negative rewards

For reasonable languages L, there are degenerate rationality-reward pairs (p, R) of minimal Kolmogorov complexity K_L among those compatible with the human policy $\dot{\pi}$.

$$K_L \begin{pmatrix} (p_{\dot{\pi}}, 0) \\ (p_g, R_{\dot{\pi}}) \\ (-p_g, -R_{\dot{\pi}}) \end{pmatrix} \quad \approx \quad \min_{(p,R):\; p(R)=\dot{\pi}} K_L(p, R) \quad \approx K_L(\dot{\pi}).$$

Any reasonable human planner-reward pair $(\dot{p}, \dot{R})$ is of non-negligible higher complexity than these.

$$K_L(\dot{p}, \dot{R}) \quad >> \quad K_L \begin{pmatrix} (p_{\dot{\pi}}, 0) \\ (p_g, R_{\dot{\pi}}) \\ (-p_g, -R_{\dot{\pi}}) \end{pmatrix}.$$

Also, there is a planner $p' = -\dot{p}$ such that the complexity of the opposite reward $-\dot{R}$ is comparable.

$$K_L(\dot{p}, \dot{R}) \quad \approx \quad K_L(-\dot{p}, -\dot{R}).$$

Conclusion

The simplest high-likelihood hypotheses about reward and planner are degenerate. How do humans themselves deal with this problem? They necessarily use some shared priors, that have never been informed by observation. Too strong assumptions can lead to catastrophic false inferences so we need to find minimal sufficient ones.
One potential approach is to look at the 'variables' and models within the human thought process, see Appendix C of the paper.

Contact Information

- Web: www.fhi.ox.ac.uk/
- Email: soeren.mindermann@gmail.com, stuart.armstrong@philosophy.ox.ac.uk

References

[1] T. Everitt, V. Krakovna, L. Orseau, M. Hutter, and S. Legg, "Reinforcement learning with a corrupted reward channel," *arXiv preprint arXiv:1705.08417*, 2017.

[2] P. W. Glimcher, C. F. Camerer, E. Fehr, and R. A. Poldrack, "Neuroeconomics: Decision making and the brain," 2009.

[3] L. Muehlhauser and L. Helm, "The singularity and machine ethics," in *Singularity Hypotheses*, pp. 101–126, Springer, 2012.

[4] J. G. Miller, "Culture and the development of everyday social explanation," *Journal of personality and social psychology*, vol. 46, no. 5, p. 961, 1984.

[5] P. Slovic and A. Tversky, "Who accepts savage's axiom?," *Behavioral science*, vol. 19, no. 6, pp. 368–373, 1974.

[6] H. Schuman and J. Ludwig, "The norm of even-handedness in surveys as in life," *American Sociological Review*, pp. 112–120, 1983.