AI Alignment and Human Reward

Patrick Butlin

The Alignment Problem: Suppose that we will build powerful, autonomous AI agents. How can we determine their values so as to ensure that their actions will benefit us?

An approach to the problem: AI agents should learn what individual humans value (Russell 2019). Their objectives will be derived from these values.
For example, an AI built to serve the public good might learn what many people value, then promote an aggregate of these values.

A question for this approach: Humans value things in many different ways. AIs could learn what we value in any one of these ways, or some combination. Which of the ways in which we value things should be the **target for alignment**?

Human values, and criteria for their assessment as targets ←

Assessment of human reward functions as a target for alignment →

An Assortment of Human Values

Ways in which we treat things as valuable	Examples
Choices or dispositions to choose	Preferences as dispositions to choose (Russell 2019)
Evaluative attitudes	Desires, evaluative beliefs, intentions, preferences as comparative judgments
Subpersonal evaluative representations	Action and outcome values posited by RL framework (Dolan & Dayan 2013, Sotala 2016)
Dispositions to treat stimuli as valuable in cognition	Dispositions to take pleasure, reward functions

Human evaluative cognition is complex; these are some of the ways in which we treat things as having value. If AIs are to learn what humans value we must be precise about which values are the target, and understand how these values are related to our interests.

Well-being and the Assessment of Targets

Basic criterion: To be a good target for alignment, a set of some person's values must be such that if their life scored highly on a metric derived from this set, it would be good for them.

What is needed for a good human life? Philosophers advocate hedonist, desire-satisfaction and objective list theories (Parfit 1984). The lists of goods in objective list theories provide a useful heuristic test for assessing alignment targets.

A list of objective goods summarised from Fletcher (2016):
- **Experiential goods:** pleasure, happiness, aesthetic experience
- **Social goods:** friendship, virtue
- **Perfectionist goods:** knowledge, achievement, development of abilities, rational activity, excellence in play, work and agency

A further heuristic: Many philosophers judge that life in a scenario like the Experience Machine (Nozick 1974) would not be good. We should be wary of targets which would give high scores to lives of simulated experience or direct brain stimulation.

References

Barto, A. 2013. Intrinsic motivation and reinforcement learning. In Baldassarre & Minolli, eds., *Intrinsically Motivated Learning in Natural and Artificial Systems*.
Daw, N. & J. P. O'Doherty. 2013. Multiple systems for value learning. In Fehr & Glimcher, eds., *Neuroeconomics: Decision-Making and the Brain.*
Deci, E. & R. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.*
Dolan, R. & P. Dayan. 2013. Goals and habits in the brain. *Neuron* 80: 312-325.
Fletcher, G. 2016. Objective list theories. In *The Routledge Handbook of Philosophy of Well-Being*.
Jeuchems, K. & C. Summerfield. 2019. Where does value come from? *Trends in Cognitive Sciences* 23: 836-850.
Nozick, R. 1974. *Anarchy, State and Utopia*.
Oudeyer, P.-Y., F. Kaplan & V. Hafner. 2007. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation* 11: 265-286.
Parfit, D. 1984. *Reasons and Persons*.
Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*.
Schmidhuber, J. 2010. Formal theory of creativity, fun and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development* 2: 230-247.
Sotala, K. 2016. Defining human values for value learners. In *Papers from the 2016 AAAI Workshop on AI, Ethics and Society*.

This research was supported by Survival and Flourishing.

Human Reward Functions

In standard RL theory, the concept of a reward function plays two roles:
- Optimal behaviour is defined as that which maximises reward
- The reward function describes evaluative feedback which the agent receives from the environment

We can use the latter role to understand human reward functions, although humans do not perceive reward itself.

My reward function describes my innate disposition to treat stimuli as having values for the purpose of value learning. 'Innate' is needed because learnt value representations also influence subsequent learning.

Stimuli with non-zero values in a normal human reward function	Stimuli which are not rewards, because we must learn to value them
Positive: food, sex, some social interactions… Negative: injury, illness	Ice-cream, books, sports, specific friendships…

My reward function describes my most fundamental values, in the sense that my other values are learnt based on this function, and are contingent on my circumstances.

Reasons for Optimism

Would a high-reward life be a good life?
Suppose that high levels of reward from social interaction require real friendships and family relationships. Then a high-reward life would involve plenty of food, good relationships, and little physical suffering. It would be good in important ways.

Which of the objective goods might be missing?
- Pleasure? – see below
- Happiness? – would presumably follow if life was good in other ways
- Perfectionist goods are the most likely missing elements

Intrinsic motivation and learning as a reward
However, psychologists argue that humans have 'intrinsic' motivation to learn, explore, play, and achieve goals (Deci & Ryan 1985). Schmidhuber (2010) and Oudeyer et al. (2007) give a partial explanation of this by claiming that progress in learning is rewarding.

This indicates that a high-reward life would involve acquiring knowledge and developing abilities. Other rewards may explain other aspects of intrinsic motivation, so that the high-reward life would also involve achievement and excellence.

Reasons for Pessimism

Empirical and conceptual problems in the application of RL theory to human psychology give us three reasons to be pessimistic about reward functions as a target.

1. The Boundary between Agent and Environment
Barto (2013) argues that RL agents are homunculi inside our minds. This is because in standard RL, reward signals are inputs to the agent. But organisms must infer reward levels from perceptible stimuli, and generate internal reward signals.
Barto's perspective implies that a high-reward life could be produced by brain stimulation.

2. Pleasure and Reward
The relationship between pleasure and reward is uncertain. If pleasure *is* reward, any highly pleasurable life, including an Experience Machine-like one, will be highly rewarding.
If pleasure is a reward *signal*, or is *a kind* of reward, this does not follow.

3. Do Humans Have Reward Functions?
It is widely accepted that humans use multiple systems for value learning (Daw & O'Doherty 2013). If these use different reward functions, we do not have unique reward functions.
It is also possible that the RL framework is not a good model for human value learning and choice (Jeuchems & Summerfield 2019). So we may not have reward functions at all.