Motivation Hierarchy

Oliver Ricahrdson

February 12, 2019

How do we acquire a moral compass? How do we acquire and change our motivations?

1 Mechanisms in Humans

There seem, at least on the surface, to be several mechanisms for this. I will go through some of them, with examples:

1.1 From Authority

Is it ok to take money? You ask your dad, who says no. From now on, you believe taking money is wrong. It's just how you were raised.

1.2 From Consistency

2 Motivations, Morals, and Preferences

I've been using the terms interchangably, but it may be worth looking more closely at where differences between these might lie.

2.1 Preferences.

A partial order on worlds $\mathcal{W}$, $(\mathcal{W}, \leq)$, which gets at which worlds an agent would "prefer" to exist. If you prefer tea to coffee, then you prefer the world where you get tea to the one where you get coffee, and $w_{\text{tea}} \geq w_{\text{coffee}}$.

This is all find if you're a cognitively unbounded agent and know everything, but it's impossible to have an explicit preference over all possible futures. It's impossible to actually store or compute what the entire future of a world will look like, must less evaluate it for preferences — and so when we say w_{tea}, we don't really know what that means.

What's worse, is that nobody actually can know which future universes are even possible. It might not even be possible for you to choose tea given the current state of the universe! So instead we have to take small slices of the universe and have preferences over them, which may or may not be consistent, or possible, or interact with the laws of physics in any meaningful way. The way we put them together is also totally unclear.

Preferences over internal models. We can imagine instead having preferences about our mental models about future imaginings, or preferences over what we ourselves will see and feel. So what's the difference between the two? Imaginings are models of the world in some "objective" sense, giving more weight to certain things you interact with, whereas models of your inputs directly could be easily hijacked. It is possible to have preferences which state things like: I would prefer to ACTUALLY be in paradise, than to think I'm in paradise while I'm actually in a drug den.

This is not at all an incoherent thing to say, and does indeed seem to be a preference as described in the first kind. As a result, we have to consider preferences over internal representations of some

2.2 Utilities.

An embedding of a totally ordered preference in real space, which explicitly exposes the sum + for combining preferences.

$$U : \mathcal{W} \to \mathbb{R}$$

2.3 Motivations.

Generally, "rational" agents are thought to act in such a way as to maximize their utilities. The word "motivation" seems to be a more situated version of a preference. Perhaps the right way to formalize them is not by explicit preferences over futures, but rather the ones induced by motivation

Problems. Why is it possible to be unmotivated to do something you thinks is important and you derive pleasure from? Even worse, how can it be possible that one does nothing even though it makes a person feel bad? Even though even the person themself does not know

The revealed preferences model says: -

2.4 Morals.

3 Active Learning