The Case for Longtermism & Avoiding X-Risk

Ryan Doody

February 8, 2022

Axiological & Deontological Longtermism

Greaves & MacAskill defend *Axiological Strong Longtermism (ASL)*: roughly, that the best actions are best because of their effects on the very far future.

Their argument involves making three (controversial) assumptions, and demonstrating that ASL is true if those assumptions hold. They then argue that those assumptions aren't essential: their conclusion—ASL—is fairly plausible even without them.

Let's suppose ASL is true. What follows about what we ought to do? *Deontic Strong Longtermism (DSL)*: roughly, that one ought to choose the option that's best for the very far future.

They offer the following argument:

Axiological strong longtermism (ASL):

(i) Every option that is near-best overall is near-best for the far future.

(ii) Every option that is near-best overall delivers much larger benefits in the far future than in the near future.

Assumptions:

(1) The value of action is its *expected value*.

(2) The value of a complete world-history is the *total welfare* it contains.

(3) Time-separability: the value of one period of time is independent of the value of any other.

Stakes-Sensitivity Argument

P1 If the stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor, one ought to choose a near-best option.

P2 In the most important decisions facing agents today, the stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor.

C In the most important decisions facing agents today, one ought to choose a near-best option.

When evaluating an argument, there are two pertinent questions:

1. Is the argument *valid*?
2. Is the argument *sound?*

There's what's good, and there's what one should do.

Consequentialism: one ought to do what's best.

Deontology: in some cases, we aren't required to do what's best (we have the *prerogative* not to); and, in some cases, we shouldn't do what's best (e.g., because it violates a "side-constraint").

"what's best" = what has the best consequences

Greaves & MacAskill: Given the overwhelming importance of the very far future, decisions about *where to direct* one's altruistic spending have high stakes, there are no side-constraints, and minor personal prerogatives.

Question to Consider: Suppose you have a rich friend who left their wallet unattended. You could easily swipe a few hundred dollars—they're so rich they probably won't even notice—and donate it to your favorite Longtermist cause. Should you?

(Offhand, it's wrong to steal from a friend. But, on the other hand, the stakes are very very high!)

Beckstead's Argument for Longtermism

Argument adapted from Beckstead's "A Brief Argument for the Overwhelming Importance of Shaping the Far Future" in *Effective Altruism: Philosophical Issues*, ed., Hilary Greaves and Theron Pummer.

A Brief Argument for the Overwhelming Importance of Shaping the Far Future

P1 There's a chance humanity will survive for billions of years.

P2 If there's a chance humanity will survive for billions of years, then the expected value of the future is extremely great.

P3 There are actions we can take now that would shape the expected trajectory of humanity's future in important ways.

P4 If the expected value of the future is extremely great and and there are actions we can take now that would shape the expected trajectory of humanity's future in important ways, then what matters most (in expectation) is that we do what is best (in expectation) for the trajectory of humanity's future over the coming billions years.

C What matters most (in expectation) is that we do what is best (in expectation) for the trajectory of humanity's future over the coming billions years.

Beckstead's argument appeals to a number of assumptions (which provide support for **P2**).

Period Independence: How well history goes as a whole is a function of how well things go during each period of history.

Additionality: If there are people who have good lives (etc.) during a period of history, that makes that period go better than it would have if nothing good had happened.

Temporal Neutrality: The value of a period is independent of when it occurs.

Risk Neutrality: The value of a prospect equals its expected value.

What is best for the trajectory of humanity's future? He thinks *existential risk reduction* is more important than achieving proximate benefits and speeding up development

. . . but that, in general, "What matters most for shaping the far future is producing positive trajectory changes and avoiding negative ones," (p. 95).

How Valuable Is Existential Risk Reduction?

Suppose existential risk reduction is valuable. *How* valuable is it? Ord provides a simple model to estimate the expected value of the future:

$$EV(Future) = \sum_{i=0}^{\infty}(1-r)^i \cdot v = \frac{v}{r} \tag{1}$$

Assumptions: (i) In each century, there is a (constant) risk r of extinction; (ii) We have the ability to reduce r in our century; (iii) Each century (prior to catastrophe) has the same intrinsic value v.

Surprising Conclusions: The value of eliminating all risk this century is the same no matter the size of r; the value of reducing r in all future centuries is *higher* the *lower* r is.

For example, the value of *halving* all future risk is:

$$\frac{v}{r/2} - \frac{v}{r} = \frac{v}{r}$$

Which is higher the closer r is to 0.