Longtermism (cont'd) & Avoiding Existential Risk

PHIL 1561 Ethics, Economics, and the Future
Ryan Doody

Contents:

Axiological Longtermism
Deontological Longtermism
Ord's "Simple Model"
The Time of Perils

What Is (Axiological Strong) Longtermism?

Axiological Strong Longtermism:
In the most important decision situations facing agents today,

(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**.

The Far Future?
Everything after some time *t* (where *t* is, e.g., 100 years after the point of decision).

The Near Future?
Everything before *t* and after the point of decision.

Why Think It's True?

There is (in expectation) a *vast* number of lives in the future of human civilization.

V(Near-future)

V(Far-future) = sum of each person's well-being

V(Overall) = **V**(Near-future) + **V**(Far-future)

Objections

Objections to (Axiological) Longtermism

1. The Washing-out Hypothesis

2. The argument rests on many controversial assumptions

3. Epistemic worries

Objections to (Axiological) Longtermism

1. **The Washing-out Hypothesis**
2. The argument rests on many controversial assumptions
3. Epistemic worries

"Might it be that the expected instantaneous value differences between available actions **decay with time** from the point of action, and **decay sufficiently fast** that in fact the near-future effects tend to be the most important contributor to expected value?"

Response:
There are things we can do now that we can be fairly confident will affect the far-future in positive ways.

Example: Existential Risk Reduction

Objections to (Axiological) Longtermism

1. The Washing-out Hypothesis
2. **The argument rests on many controversial assumptions**
3. Epistemic worries

For example:
Ex Ante Value of an option is its *expected* value;

Value is *total* welfare;

Time-separability for benefits.

Objections to (Axiological) Longtermism

1. The Washing-out Hypothesis

2. The argument rests on many controversial assumptions

3. **Epistemic worries**

"[W]e are clueless both about what the far future will be like, and about the differences that we might be able to make to that future."

We will discuss these more later on.

Deontic Strong Longtermism:

One ought to choose the option that's best for the very far future.

The 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.

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**").

The 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.

Discussion Question:
Suppose you have a rich friend who has 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?

The 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.

Discussion Question:

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

Should you?

How Valuable is Existential Risk Reduction?

Ord's "Simple Model" of Existential Risk Reduction

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*.

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

Ord's "Simple Model" of Existential Risk Reduction

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*.

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

Interesting Results:

1. The value of eliminating **all risk this century** is the same no matter the size of r.
2. The value of reducing r in **all future centuries** is higher the lower r is.

High Risk, Low Reward?

Thorstad's 'High Risk, Low Reward'

Thorstad argues that there is a tension between the following two claims:

the astronomical value thesis: he best available options for reducing existential risk today have astronomical value.

existential risk pessimism: existential risk this century is very high.

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

Thorstad's 'High Risk, Low Reward'

Thorstad argues that there is a tension between the following two claims:

the astronomical value thesis: he best available options for reducing existential risk today have astronomical value.

existential risk pessimism: existential risk this century is very high.

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

Although the future itself may be astronomically valuable, the expected value of reducing existential risk in this century is capped at the value v of an additional century of human existence. [377]

Thorstad's 'High Risk, Low Reward'

Thorstad argues that there is a tension between the following two claims:

the astronomical value thesis: he best available options for reducing existential risk today have astronomical value.

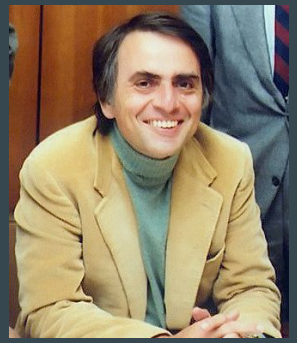
existential risk pessimism: existential risk this century is very high.

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

although the value of existential risk reduction is in principle unbounded, in practice this value may be modest if we are pessimistic about existential risk. By way of illustration, setting r to a pessimistic 20% values a 10% relative reduction in existential risk across all centuries at once at a modest five-ninths of the value of the present century. Even a 90% reduction in risk across all centuries would carry just 45 times the value of the present century. [381]

Time of Perils

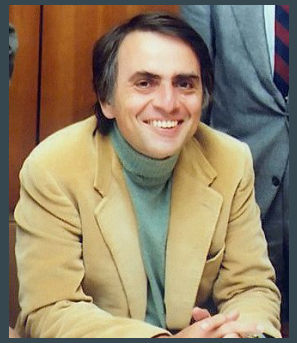
Time of Perils



“It might be a familiar progression, transpiring on many worlds ... life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges ... and then technology is invented. It dawns on them that there are such things as laws of Nature ::: and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others [who] are not so lucky or so prudent, perish.”

Time of Perils

“It might be a familiar progression, transpiring on many worlds ... life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges ... and then technology is invented. It dawns on them that there are such things as laws of Nature ::: and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others [who] are not so lucky or so prudent, perish.”



But how realistic is this, really?