Avoiding Existential Risk

•••

PHIL 1561 Ethics, Economics, and the Future
Ryan Doody

Contents:

Ord’s “Simple Model”

“High Risk, Low Reward”

The Time of Perils

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 bitcoin ~~dollars~~—they're so rich they probably won't even notice—and donate it to your favorite Longtermist cause.

Should you?

Discussion Question:
Suppose you have a rich friend who has left their crypto wallet unattended. You could easily swipe a few hundred bitcoin ~~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.

"If the stakes are very high…"

Are the stakes very high?

What *is* the (expected) value of reducing the amount of existential risk we face **this century**?
What about reducing it for **all future centuries**?

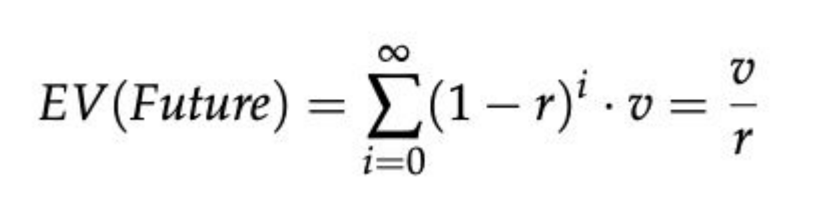
How Valuable is Existential Risk Reduction?

Thorstad’s ‘High Risk, Low Reward’

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

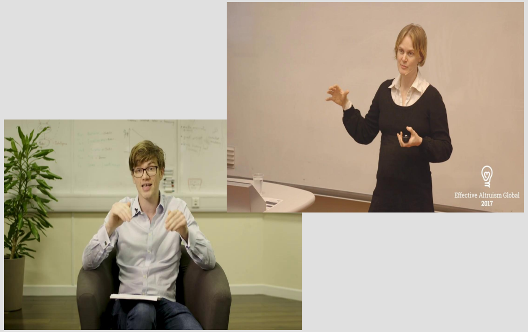
the astronomical value thesis: the best available options for reducing existential risk today have astronomical value.

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



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

the astronomical value thesis is what supports the claim that “the stakes are very high”

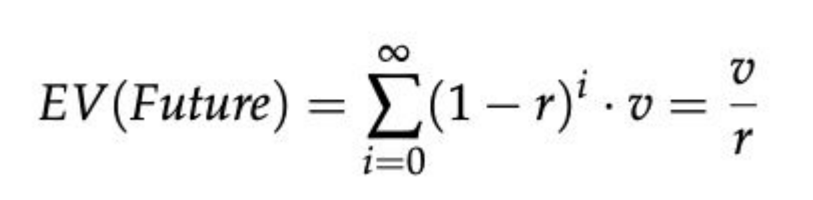


Thorstad’s ‘High Risk, Low Reward’

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

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

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



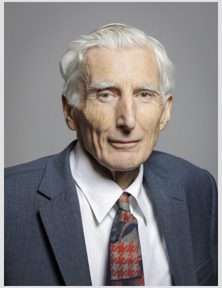
$$EV(Future) = \sum_{i=0}^{\infty}(1-r)^i \cdot v = \frac{v}{r}$$

existential risk pessimism is often taken to bolster the astronomical value thesis

existential risk pessimism is often taken to bolster the astronomical value thesis

Existential Risk this Century:

“One in six”



“50% chance of collapse”

“19% chance of human extinction”

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.

Ord's "Simple Model" of Existential Risk Reduction

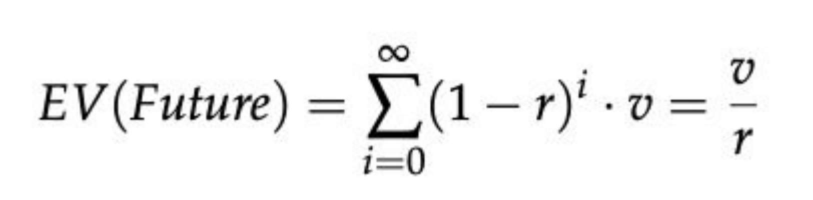
The value of eliminating all risk this century is the same no matter the size of r.

EV(*Future*) =
$(1 - r)v + (1 - r)^2 v + (1 - r)^3 v + (1 - r)^4 v + \ldots$

If we reduce existential risk to 0 in the first century, we get:

$v + (1 - r)v + (1 - r)^2 v + (1 - r)^3 v + \ldots$

The difference is: *v*.



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

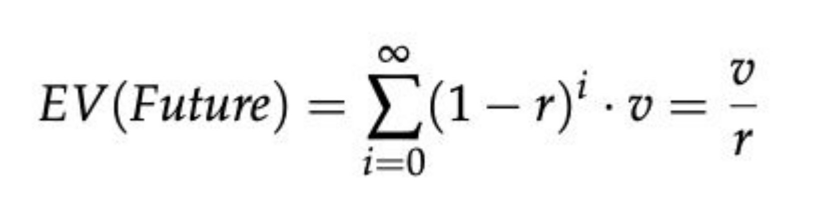
Interesting Results:

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.

Thorstad's 'High Risk, Low Reward'

The value of eliminating all risk this century is the same no matter the size of r—it's *v*.

The value of a century (*v*) might be large, but it's not astronomical!



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

"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’

The value of eliminating all risk this century is the same no matter the size of r—it’s *v*.

The value of a century (*v*) might be large, but it’s not astronomical!

And so:

“This means that interventions which present a small chance of preventing existential catastrophe in this century may not be obviously more valuable than other altruistic interventions, such as work done to mitigate extreme poverty.” [377]

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

“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]

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.

Ord's "Simple Model" of Existential Risk Reduction

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:

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

In general, the value of reducing per-century risk, from *r* to *(1 - f)r*, in all centuries is:

f/(1 - f) × v/r

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

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.

Thorstad’s ‘High Risk, Low Reward’

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:

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

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

In general, the value of reducing per-century risk, from *r* to *(1 - f)r*, in all centuries is:

$$f/(1 - f) \times v/r$$

“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.” [381]

Thorstad's 'High Risk, Low Reward'

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:

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

In general, the value of reducing per-century risk, from *r* to *(1 - f)r*, in all centuries is:

$$f/(1 - f) \times v/r$$

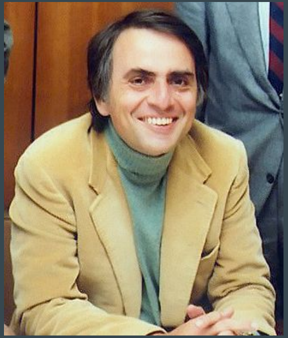
$$EV(Future) = \sum_{i=0}^{\infty} (1 - r)^i \cdot v = \frac{v}{r}$$

"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.”

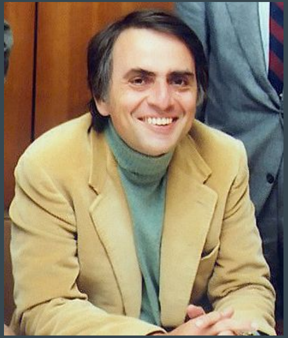


Rapid technological growth has given us the means to quickly **destroy** ourselves.

But if we learn to manage the risks, we will enter into a period of relative **safety**.

Time of Perils

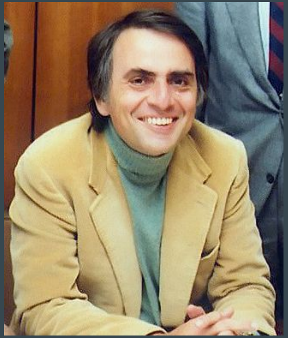
“It might be a familiar progression, transpiring on many worlds...life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges...and then technology is invented. It dawns on them that there are such things as laws of Nature...and that knowledge of these laws can be made both 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?

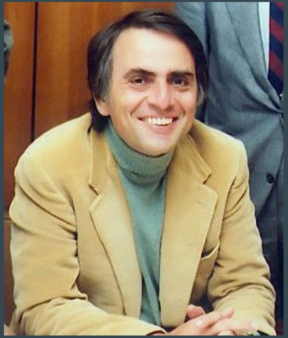
Time of Perils

The Time of Perils Hypothesis:
Existential risk will remain high for several centuries, but drop to a low level if humanity is able to survive this time of perils.



But how realistic is this, really?

Time of Perils



The Time of Perils Hypothesis: Existential risk will remain high for several centuries, but drop to a low level if humanity is able to survive this time of perils.

If the **Time of Perils Hypothesis** is correct, then reducing existential risk *can* be astronomically valuable.

It depends on two factors:

But how realistic is this, really?

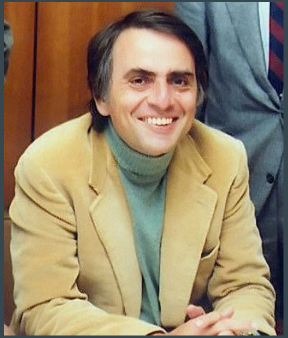
Time of Perils

The Time of Perils Hypothesis: Existential risk will remain high for **several centuries**, but drop to a low level if humanity is able to survive this time of perils.

If the **Time of Perils Hypothesis** is correct, then reducing existential risk *can* be astronomically valuable.

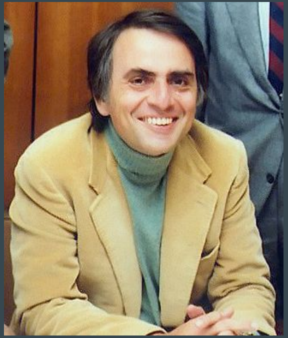
It depends on two factors:

N = the length of the perilous period



But how realistic is this, really?

Time of Perils



The Time of Perils Hypothesis: Existential risk will remain high for **several centuries**, but drop to **a low level** if humanity is able to survive this time of perils.

If the **Time of Perils Hypothesis** is correct, then reducing existential risk *can* be astronomically valuable.

It depends on two factors:

N = the length of the perilous period
$\mathbf{r_1}$ = the risk of extinction in the post-peril period

But how realistic is this, really?

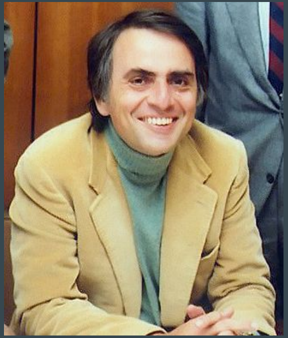
Time of Perils

The Time of Perils Hypothesis: Existential risk will remain high for **several centuries**, but drop to **a low level** if humanity is able to survive this time of perils.

If the **Time of Perils Hypothesis** is correct, then reducing existential risk *can* be astronomically valuable.

It depends on two factors:

N = the length of the perilous period
$\mathbf{r_1}$ = the risk of extinction in the post-peril period

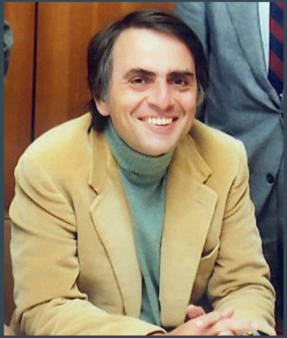


The perilous period **N** must be fairly short.

The post-peril risk $\mathbf{r_1}$ must be low.

Time of Perils

The Time of Perils Hypothesis: Existential risk will remain high for **several centuries**, but drop to **a low level** if humanity is able to survive this time of perils.



The perilous period **N** must be fairly short.

The post-peril risk **r_1** must be low.

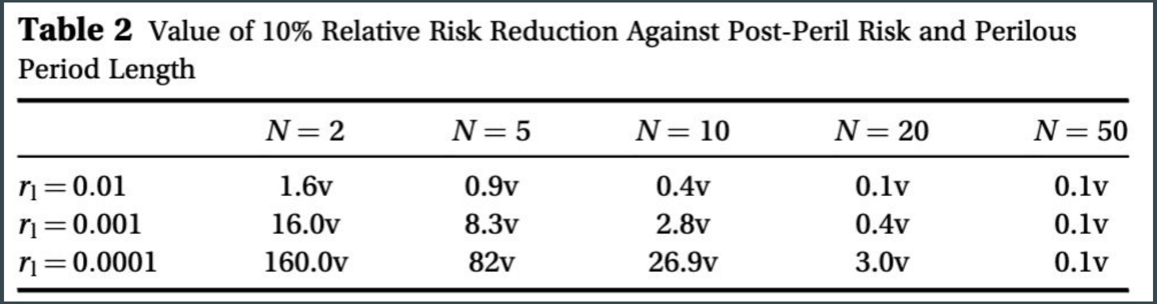
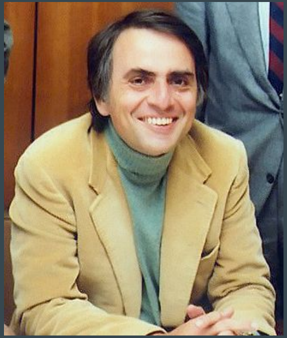


Table 2 Value of 10% Relative Risk Reduction Against Post-Peril Risk and Perilous Period Length

	$N=2$	$N=5$	$N=10$	$N=20$	$N=50$
$r_1=0.01$	1.6v	0.9v	0.4v	0.1v	0.1v
$r_1=0.001$	16.0v	8.3v	2.8v	0.4v	0.1v
$r_1=0.0001$	160.0v	82v	26.9v	3.0v	0.1v

Time of Perils



The Time of Perils Hypothesis:
Existential risk will remain high for **several centuries**, but drop to **a low level** if humanity is able to survive this time of perils.

But how realistic is this, really?

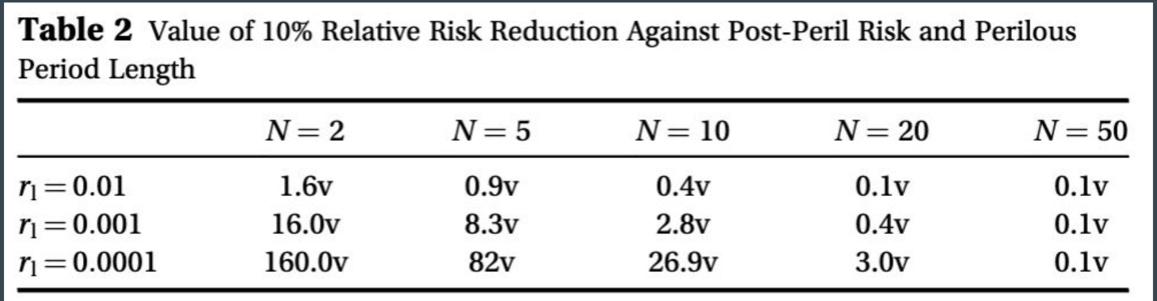
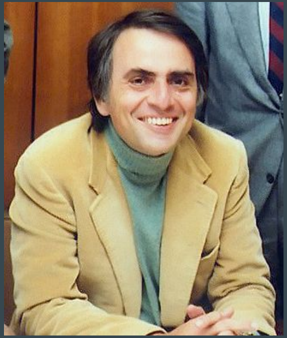


Table 2 Value of 10% Relative Risk Reduction Against Post-Peril Risk and Perilous Period Length

	$N=2$	$N=5$	$N=10$	$N=20$	$N=50$
$r_1=0.01$	1.6v	0.9v	0.4v	0.1v	0.1v
$r_1=0.001$	16.0v	8.3v	2.8v	0.4v	0.1v
$r_1=0.0001$	160.0v	82v	26.9v	3.0v	0.1v

Time of Perils

The Time of Perils Hypothesis:
Existential risk will remain high for **several centuries**, but drop to **a low level** if humanity is able to survive this time of perils.



Are we really that special?!?

But how realistic is this, really?

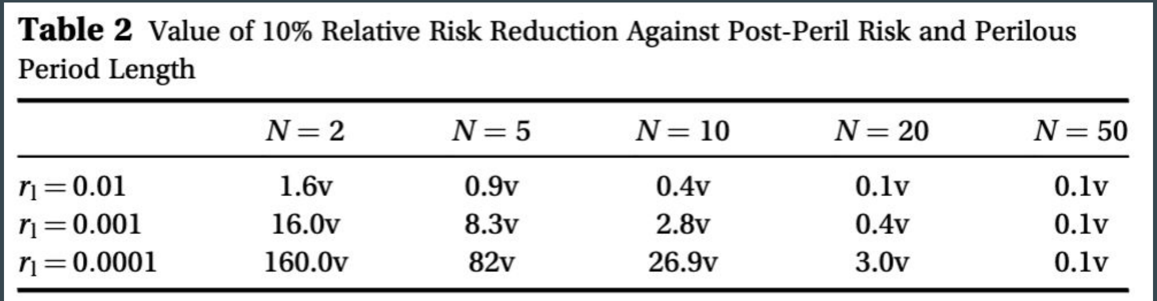


Table 2 Value of 10% Relative Risk Reduction Against Post-Peril Risk and Perilous Period Length

	$N=2$	$N=5$	$N=10$	$N=20$	$N=50$
$r_1=0.01$	1.6v	0.9v	0.4v	0.1v	0.1v
$r_1=0.001$	16.0v	8.3v	2.8v	0.4v	0.1v
$r_1=0.0001$	160.0v	82v	26.9v	3.0v	0.1v

Revisiting (the assumptions behind) Ord’s “Simple Model”

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*.**

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

How should we value the lives of future people?

How should we value the lives of future people?

What if our actions affect - not only *how well-off* future people will be - but *who* those future people will be?

The Non-Identity Problem