# Infinite Ignorance

Loren K. Fryxell

University of Oxford and Global Priorities Institute

*loren.fryxell@economics.ox.ac.uk*

December 9, 2024

15th Oxford Workshop on Global Priorities Research

# Introduction

# The Question

How should we make ethical decisions in a world that is potentially infinite when we can affect only a finite part of it?

# Philosophy

# Infinite Paralysis

It is well-known that most ethical theories break down when applied to this question

In fact, all aggregative consequentialist theories suffer from infinite paralysis (Bostrom, 2011)

> **Infinite Paralysis.** Suppose every feasible action causes a finite change in value. If there is any positive probability the universe contains infinite moral value, then we should be morally indifferent among all our feasible actions.

# Infinite Paralysis: Intuition

The intuition is simple. If we

1. assign moral values to universes and
2. evaluate probability distributions over universes by their expected moral value,

then the expected moral value of any probability distribution which places positive probability on universes with infinite moral value is infinite or undefined

So making a mere finite change to an infinite expected moral value changes nothing

# Economics

# Standard Model

$$V(x) = \sum_{t=0}^{\infty} \delta^t u(x_t)$$

# The Pure Rate of Time Preference

$$V(x) = \sum_{t=0}^{\infty} \delta^t u(x_t)$$

If we have a positive pure rate of time preference, then a temporally infinite universe causes no issues (though a spatially infinite universe would)

But many, including me, believe that weighing people's welfare less the more distant they are from us in time or space seems morally inadmissible

So we should not have a positive pure rate of time preference

# Standard Model Assumes a Finite Universe

$$V(x) = \sum_{t=0}^{\infty} \delta^t u(x_t)$$

With a zero rate of pure time preference, the discount factor must instead represent extinction risk

This means that, in this model, the world is finite (with probability one)

What happens if we have a universe that is possibly infinite?

# Extended Model with Potentially Infinite Universe

**A Very Simple Model with Potentially Infinite Universe**

$$V(x) = p \sum_{t=0}^{\infty} S^{\text{finite}}(t) u(x_t) + (1-p) \sum_{t=0}^{\infty} S^{\infty}(t) u(x_t)$$

**Example**

$$V(x) = p \sum_{t=0}^{\infty} \delta^t u(x_t) + (1-p) \sum_{t=0}^{\infty} u(x_t)$$

$V(x)$ is undefined for most $x$

In sum, economists generally do not consider models with a potentially infinite universe—and if they did, the standard approach wouldn't work

# Back to Foundations

# Back to the Foundations

In both philosophy and economics, why would we write it this way?

$$V(x) = p \times \text{well-defined sum} + (1-p) \times \text{undefined sum}$$

In particular, **how did we decide** that what we care about can be measured as an expectation over some numeric utility assigned to states of affairs?

And does this foundation say that we should take expectations over undefined numbers?

# Back to the Foundations

In my view, the most sensible and convincing foundation for expected utility is the von Neumann–Morgenstern (vNM) axioms

But these axioms say nothing about coherent preferences in a potentially infinite universe

The axioms themselves imply the world is finite (the continuity axiom)

Hence, in my view, the logic underpinning infinite paralysis is based on **theories which are being applied outside the domain in which they are grounded**

Expected utility theory was never supposed to apply to potentially infinite worlds

# Back to the Foundations

In this paper,

- I go back to the beginning,
- formulate that we have a potentially infinite world from the start, and
- build the theory from there,

just as the vNM axioms do in a finite world

# Result: Infinite Ignorance

The result is what I call *infinite ignorance*

**Infinite Ignorance.** Suppose every feasible action causes a finite change in value. If there is any positive probability that the universe contains finite moral value, then we should evaluate our feasible actions conditional on the universe containing finite moral value.

Saving a life is clearly good in a finite world, but it's not clearly good in a potentially infinite world. Infinite ignorance says that if it's good in a finite world, it's good in a potentially infinite world too.

# Infinite Paralysis vs Infinite Ignorance

**Infinite Paralysis.** Suppose every feasible action causes a finite change in value. If there is any positive probability the universe contains infinite moral value, then we should be morally indifferent among all our feasible actions.

**Infinite Ignorance.** Suppose every feasible action causes a finite change in value. If there is any positive probability that the universe contains finite moral value, then we should evaluate our feasible actions conditional on the universe containing finite moral value.

# Infinite Ignorance and Economics

Infinite ignorance implies we can evaluate $x$ by

$$V(x) = p \sum_{t=0}^{\infty} S^{\text{finite}}(t) u(x_t) + \cancel{(1-p) \sum_{t=0}^{\infty} S^{\infty}(t) u(x_t)}$$

# Quick Note Before We Begin

# Ranking Infinite Universes is Supremely Difficult

I think ranking surely infinite universes is *supremely* difficult, confusing, and paradoxical

Most approaches model universes by cutting them into pieces, assigning values to the pieces, and lining up the pieces into an infinite sequence

They then impose axioms on orderings $\succeq$ over the space of infinite sequences of real numbers

One common axiom is Pareto: If each element of sequence $A$ is strictly greater than it's corresponding element in sequence $B$, then $A \succ B$

# Pareto Axiom on Infinite Domains

However, in infinite domains this Pareto axiom changes how it ranks universes depending on how you cut the universe into pieces

Universe $A >_{\text{Pareto}}$ Universe $B$ under one way of segmenting the universe. And the very same Universe $A <_{\text{Pareto}}$ Universe $B$ under another way of segmenting the universe

This is true if one way is segmenting the universe by people's lives and the other way is segmenting the universe by blocks of spacetime

This is *also* true if both ways segment the universe by spacetime, but they are just different segmentations of spacetime

# The Worry and its Avoidance

I don't think how we construct segmentations of universes (including within spacetime) should affect how we morally rank them

Hence, in my view, the approach of modeling universes as infinite sequences and then imposing Pareto is worrying

In this analysis, I neither segment a universe into infinitely many pieces nor do I impose Pareto

# Model

# Model

Let $\Omega$ be a set of complete histories of the universe across all of spacetime (past and future)

Let $\Delta\Omega$ be the set of all probability distributions on $\Omega$

Each action we can take gives rise to a probability distribution over complete histories of the universe

Hence, we will call such distributions $p, q \in \Delta\Omega$ *actions*

# Natural and Divine Preferences

Let $\succeq$ be a transitive (and possibly incomplete) relation on $\Delta\Omega$ reflecting the moral preferences over $\Delta\Omega$ that a decision maker feels confident in

Call $\succeq$ the **natural preference**

Let $\succeq^*$ be a transitive (and possibly incomplete) relation on $\Delta\Omega$ reflecting the moral *idealized* preferences over $\Delta\Omega$ that the decision maker aspires to but may presently be ignorant of—the true relation known only to God

Call $\succeq^*$ the **divine preference**

# Axiom 1 (Compatibility)

Axiom 1 states that $\succsim$ is compatible with $\succsim^*$, i.e., any preference $\succsim$ that a decision maker feels confident in is aligned with her idealized preference $\succsim^*$

**Axiom 1 (Compatibility).** For any $p, q \in \Delta\Omega$,

$$p \succ q \implies p \succ^* q \quad \text{and} \quad p \sim q \implies p \sim^* q.$$

# Our Sphere of Influence

Decompose $\Omega$ into $\Omega = \mathbb{I} \times \mathbb{O}$, where

- $\mathbb{I}$ describes aspects of the universe that the actions under consideration can affect—they are *inside* our sphere of influence
- $\mathbb{O}$ describes aspects of the universe that the actions under consideration cannot affect—they are *outside* our sphere of influence

# Inside Our Sphere of Influence

Suppose I ask you how much you'd bet on the event

Donald Trump eats a Big Mac in the first week of 2025.

Would your bet change if you take action $p$ vs $q$?

You can see how it might, since your actions $p$ and $q$ might predictably influence the chance this happens

This is what it *means* for an aspect of the universe to be within your sphere of influence—that your bet may change upon taking different feasible actions

# Outside Our Sphere of Influence

Suppose I ask you how much you'd bet on the event

An alien creature on a planet unreachable by light shone from earth today, Dlanod Pmurt, eats a Gib Cam in the first week of 2025.

Would your bet change given you take action $p$ vs $q$?

You can see how it might not, since this planet is outside your future light cone

This is what it *means* for an aspect of the universe to be outside your sphere of influence—that your bet would never change upon taking different feasible actions

# Formal Definition

We say that two actions $p$ and $q$ have no ex ante effect on $\mathbb{O}$ relative to each other if, for any $O \subseteq \mathbb{O}$, the probability that $O$ occurs is the same

**Definition.** Two actions $p, q \in \Delta\Omega$ *have no (ex ante) effect on* $\mathbb{O}$ if, for any $O \subseteq \mathbb{O}$, $p(\mathbb{I} \times O) = q(\mathbb{I} \times O)$.

Note that $\mathbb{I} \times O$ is the set of histories in which $O$ occurs

# Finite Worlds

A "finite" world is just a world which we are confident in evaluating

Let $\mathbb{O}_F \subseteq \mathbb{O}$ be a set of aspects of the universe outside our causal sphere such that, when conditioning on $\mathbb{I} \times \mathbb{O}_F$, we can confidently rank actions $p$ and $q$

For example, $\mathbb{I} \times \mathbb{O}_F$ might be the event that

- there are no people outside our sphere of influence
- there are less than a trillion people outside our sphere of influence
- there is no sentient life outside our sphere of influence

Intuitively, $\mathbb{I} \times \mathbb{O}_F$ is the set of histories for which the universe has finite "value"

# Finite Worlds

Formally, we will say that $p$ and $q$ are comparable given $\mathbb{O}_F$

**Definition.** Two actions $p, q \in \Delta\Omega$ are *comparable given* $\mathbb{O}_F$ if the decision maker can rank $p$ and $q$ conditional on $\mathbb{I} \times \mathbb{O}_F$, i.e.,

$$p(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \succeq q(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \quad \text{or} \quad q(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \succeq p(\cdot \mid \mathbb{I} \times \mathbb{O}_F).$$

Let $\mathbb{O}_\infty = \mathbb{O} \setminus \mathbb{O}_F$

Intuitively, $\mathbb{I} \times \mathbb{O}_F$ are the finite worlds and $\mathbb{I} \times \mathbb{O}_\infty$ are the infinite worlds

# Finite and Infinite Worlds

More specifically,

$\mathbb{I} \times \mathbb{O}_F$ are the worlds in which

- sentient matter only occupies finite space and time, and
- the rest of the possibly infinite universe is filled with non-sentient matter

$\mathbb{I} \times \mathbb{O}_\infty$ are the rest of the worlds

# Cautious Ignorance

**Definition.** Given any two actions $p, q \in \Delta\Omega$ which are comparable given $\mathbb{O}_F$, a decision maker is *cautiously ignorant* about $p$ and $q$ given $\mathbb{O}_\infty$ if

$$p(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \succeq q(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \implies p(\cdot \mid \mathbb{I} \times \mathbb{O}_\infty) \not\prec^* q(\cdot \mid \mathbb{I} \times \mathbb{O}_\infty).$$

Cautious ignorance reflects the idea that

- if a decision maker is confident about how to morally rank two actions $p$ and $q$ given $\mathbb{I} \times \mathbb{O}_F$, then
- although she may still be uncertain about how to rank $p$ and $q$ given $\mathbb{I} \times \mathbb{O}_\infty$, she should be confident that her preference **will not flip**

She is not completely ignorant about $\succeq^*$, but rather "cautiously" ignorant about $\succeq^*$

# Cautious Ignorance: Example

If

$$\underset{\mathbb{I}}{3 \times ☺} \quad \underset{\mathbb{O}_F}{5 \times ☺} \quad \succeq \quad \underset{\mathbb{I}}{2 \times ☺} \quad \underset{\mathbb{O}_F}{5 \times ☺}$$

then it is **not** the case that

$$\underset{\mathbb{I}}{3 \times ☺} \quad \underset{\mathbb{O}_\infty}{\infty \times ☺} \quad \prec^* \quad \underset{\mathbb{I}}{2 \times ☺} \quad \underset{\mathbb{O}_\infty}{\infty \times ☺}$$

# Axiom 2 (Independence)

Axiom 2 is vNM independence applied to the divine preference $\succeq^*$

**Axiom 2 (Independence).** For any $p, q, r \in \Delta\Omega$ and $\alpha \in (0, 1]$,

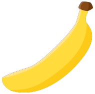
$$p \succ^* q \iff \alpha p + (1-\alpha) r \succ^* \alpha q + (1-\alpha) r$$

and

$$p \sim^* q \iff \alpha p + (1-\alpha) r \sim^* \alpha q + (1-\alpha) r.$$

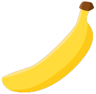
# Axiom 2 (Independence): Intuition

Suppose you tell me that you'd like me to get you an apple from the grocery store, but if there are no apples, you'd like me to get you a banana

 $\succ$ 

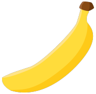
Suppose I tell you there is a chance the store is closed, in which case I won't be able to get you anything

$\alpha,$  $1-\alpha,$ 

$\succ$ $\alpha,$  $1-\alpha,$ 

# Axiom 2 (Independence): Intuition

$\alpha,$  $1-\alpha,$ 

$\succ$ $\alpha,$  $1-\alpha,$ 


Does your preference about what I should get you *if the store is open* depend on the probability *the store is closed*?

The independence axiom says it does not

# Proposition

**Proposition.** Suppose Axiom 1 (Compatibility) and Axiom 2 (Independence) hold. For any $p, q \in \Delta\Omega$, if

- $p$ and $q$ have no effect on $\mathbb{O}$,
- $p$ and $q$ are comparable given $\mathbb{O}_F$, and
- we are cautiously ignorant about $p$ and $q$ given $\mathbb{O}_\infty$,

then

$$p(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \succ q(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \implies p \succ^* q \text{ or they incomparable}$$

and

$$p(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \sim q(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \implies p \sim^* q \text{ or they incomparable.}$$

# Axiom 3 (Permissibility Dominance)

**Axiom 3 (Permissibility Dominance).** If some actions are surely permissible and others may or may not be permissible, a decision maker should default to the actions which are surely are permissible.

Denote the relation that respects permissibility dominance by $\succeq_0$

# Theorem (Infinite Ignorance)

**Theorem (Infinite Ignorance).** Suppose Axiom 1, 2, and 3 hold. For any $p, q \in \Delta\Omega$, if

- $p$ and $q$ have no effect on $\mathbb{O}$,
- $p$ and $q$ are comparable given $\mathbb{O}_F$, and
- we are cautiously ignorant about $p$ and $q$ given $\mathbb{O}_\infty$,

then

$$p \succeq_0 q \iff p(\cdot \mid \mathbb{I} \times \mathbb{O}_F) \succeq q(\cdot \mid \mathbb{I} \times \mathbb{O}_F).$$

When comparing two actions $p$ and $q$, it suffices to compare $p$ and $q$ **conditioning on a finite world**

# Infinite Ignorance: Intuition and Proof

- If the universe is finite, suppose saving a life is strictly better than not saving a life.
- If the universe is infinite, saving a life must not be strictly worse than not saving a life by cautious ignorance.
- There is some probability $\alpha > 0$ the universe is finite.
- Saving a life results in a strictly better universe than not saving a life with probability $\alpha$ and a not strictly worse universe with probability $1 - \alpha$.
- By independence, saving a life is either strictly better than or incomparable to not saving a life.
- By permissibility dominance, from our vantage point, saving a life is strictly better than not saving a life.

# Infinite Paralysis vs Infinite Ignorance

**Infinite Paralysis.** Suppose every feasible action causes a finite change in value. If there is any positive probability the universe contains infinite moral value, then we should be morally indifferent among all our feasible actions.

**Infinite Ignorance.** Suppose every feasible action causes a finite change in value. If there is any positive probability that the universe contains finite moral value, then we should evaluate our feasible actions conditional on the universe containing finite moral value.

# With and Without Explicit Probabilities

We did this analysis with probabilities as a primitive, where actions were probability distributions over histories (a vNM framework)

We may also do this analysis *without* probabilities as a primitive, where actions are sets of histories and subjective probabilities are merely a property of preferences over such actions (a Bolker-Jeffrey framework)

Both analyses follow the same logical path, have the same conclusions, and can be found in the paper

# Thank you for listening!

Questions, comments, or concerns?