Formal Ethics: Some Ideas

how can formalism help ethics?

Can Başkent

can@canbaskent.net www.canbaskent.net/logic @topologically

Workshop on Longtermism
Global Priorities Institute, University of Oxford, 18-19 March 2019

Today's Plan

1. Longtermism and Dynamic Preferences
2. Knowledge and Belief in Games for Longtermism
3. Few More (unbaked) Pointers

Longtermism and Dynamic Preferences

Motivation

If subjective preferences bear some weight in understanding longtermism, what is the best way to formalise them?

Motivation

If subjective preferences bear some weight in understanding longtermism, what is the best way to formalise them?

I will focus on a system that is based on *histories* and can be *updated*: past behaviour may help relevant for the current preferences, and players should be allowed to revise their preferences.

Such a system should have infinite histories and preferences defined on the to test the theoretical boundaries of longtermism.

(joint work with Guy McCusker)

The Formal Structure

We need a set of players A and a set of events/moves E.

A history h is a sequence of events from E. A time-stamp t will denote a point in history.

A player i will be associated a subset of events $E_i \subseteq E$. This allows subjectivity.

For each agent, we can define a set of histories that are indistinguishable from an epistemic point of view.

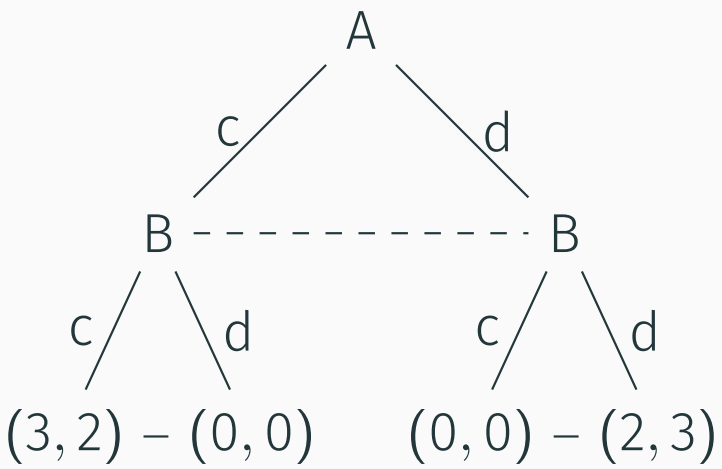
> "A Knowledge Based Semantics of Messages", Parikh and Ramanujam, *Journal of Logic, Language and Information*, vol. 12(4), pp. 453-467, 2003.

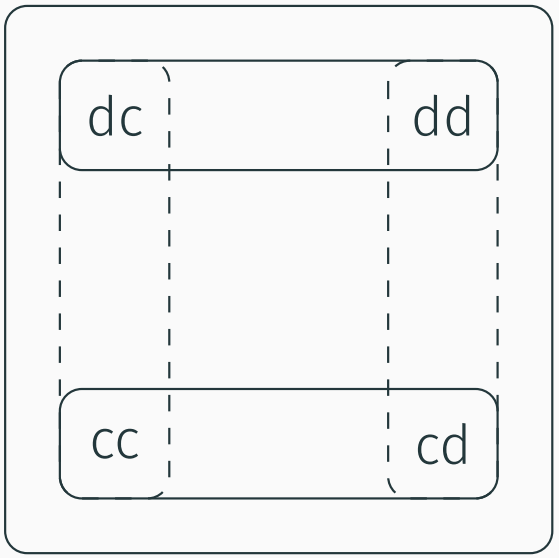
The Formal Structure

Let $h' \preceq_i h$ if the history h is (weakly)-preferable to h for player i.

We can then evaluate Boolean, epistemic ($K_i\varphi$), temporal $\bigcirc\varphi$ and preferential ($\Box_i\varphi$) formulas over history-time pairs (h, t).

Dynamic Preferences: An Example





For player *A* and *B*, consider two actions: cooking c and dancing d.

The solid line defines the knowledge set of Player *A* whereas the dashed line defines that of *B*.

Dynamic Preferences: An Example

In this coordination game, two players *A* and *B* want to attend the same event together.

They have two choices: going to a cooking class (c) or dancing (d). Player *A* prefers the cooking class, whereas Player *B* prefers dancing. But, both prefer attending the same activity rather than different ones.

A game theoretical conundrum occurs, if we are in the situation that *A* and *B* made plans to meet up to attend an event together, but they cannot remember where. If they cannot communicate, what should they do?

Dynamic Preferences: An Example

For player A, we have:

$$\text{cd} \preceq_A \text{dc} \preceq_A \text{dd} \preceq_A \text{cc}$$

and

$$\text{cd} \sim_A \text{cc} \qquad\qquad \text{dd} \sim_A \text{dc}.$$

A's incentive is to go to cooking class. But, the game is an imperfect information game.

Let us assume that A learns that B is on her way to dancing, after a common friend tells her. This eliminates A's preference of going to cooking class.

Dynamic Preferences: An Example

Then, A's highest preference becomes going dancing.

When A learns about B's d move, she revises her preferences to leave only

$$\text{cd} \preceq_A \text{dd}.$$

In this case, her best move becomes d. She no longer prefers going to cooking class over dancing, consequently the preference relation between them is eliminated.

The preference update is controlled by a sentence in the language: "B makes a d move".

Dynamic Preferences

The formula $[\varphi]\psi$ will express that "after a preference update by φ, ψ holds".

The updated preference order $\preceq_i^*$ is defined as

$$\preceq_i^* := \preceq_i \setminus \{(h, h') : h, t \models_M \varphi \text{ and } h', t \models_M \neg\varphi \text{ for any } t\}.$$

This is an efficient and complete formal system.

Longtermism and Dynamic Preferences

A next step is to evaluate the potential change of equilibria in games after preference updates.

Read backwards: What preference changes, over time, will make it easier and more efficient to reach equilibria?

This is an important point for longtermism.

Knowledge and Belief in Games for Longtermism

Paradoxes are game changers!

Paradoxes have changed how we understand set theory and logic.
Now, they are changing the way we approach games.

Paradoxes are game changers!

Paradoxes have changed how we understand set theory and logic. Now, they are changing the way we approach games.

And, they will change, again, how we solve ethical puzzles.

Absolute Longtermism

Definition

Absolute Longtermism is a paradigm where the formal ethical framework allows arbitrary number of (potentially infinitary) players and moves.

Using logical equivalence, it is easy to see that we are allowed to have arbitrary number of beliefs, assumptions and knowledge in a game of absolute longtermism.

Absolute Longtermism

Informally, absolute longtermism is a paradigm where we allow infinitary players with infinitary beliefs.

Absolute Longtermism

Informally, absolute longtermism is a paradigm where we allow infinitary players with infinitary beliefs.

Challenge

What are the mathematical limitations of absolute longtermist games?

Yablo's Paradox

Yablo's Paradox, according to its author, is a non-self referential paradox of arbitrary many sentences.

Yablo considers the following sequence of sentences.

$$S_1 : \forall k > 1, S_k \text{ is untrue},$$
$$S_2 : \forall k > 2, S_k \text{ is untrue},$$
$$S_3 : \forall k > 3, S_k \text{ is untrue},$$
$$\vdots$$

"Paradox without Self-Reference", S. Yablo, *Analysis*, vol. 53, pp. 251-2, 1993.

Yablo's Paradox

The proof is elementary (and fun).

Let S_n be true for an arbitrary n.

Then, for all $k > n$, S_k is untrue. Particularly, S_{n+1} is untrue.

On the other hand, since for all $k > n + 1$, S_k is untrue, S_{n+1} turns out to be true.

Contradiction. Thus, S_n cannot be true for any n.

However, in that case, each S_i is true as all $n > i$ is untrue.

Hence, the set of sentences $S_1, S_2, \ldots$ is impossible.

Yablo's Paradox

By using *reductio*, Yablo argues that the above set of sentences is contradictory. Here, the infinitary nature of the paradox is essential as each finite set of S_n is satisfiable.

The scheme of this paradox is not new. To the best of my knowledge, the first analysis of this paradox was suggested in 1953 by Yuting.

> "Paradox of the Class of All Grounded Classes", Sh. Yuting, *The Journal of Symbolic Logic*, vol. 18, p. 114, 1953.

A Yabloesque Paradox in Epistemic Games

Consider the following set of assumptions where numerals represent game theoretical players and assumption is the strongest belief.

A_1 : 1 believes that $\forall k > 1$, k's assumption A_l about $\forall l > k$ is untrue,

A_2 : 2 believes that $\forall k > 2$, k's assumption A_l about $\forall l > k$ is untrue,

A_3 : 3 believes that $\forall k > 3$, k's assumption A_l about $\forall l > k$ is untrue,

$\vdots$

"A Yabloesque Paradox in Epistemic Game Theory", CB, *Synthese*, 2018, vol. 195(1), pp. 441-464.

Results

Theorem

The Yabloesque sentence is inconsistent.

> “A Yabloesque Paradox in Epistemic Game Theory”, CB, *Synthese*, 2018, vol. 195(1), pp. 441-464.

An Interpretation

Imagine a queue of players, where players are conveniently named after numerals, holding beliefs about each player behind them, but not about themselves. In this case, each player i believes that each player $k > i$ behind them has an assumption about each other player $l > k$ behind them and i believes that each k's assumption is false.

This statement is perfectly perceivable for games, and involves a specific configuration of players' beliefs and assumptions, which can be expressible in the language.

And it is not self-referential. No player has a belief about herself.

An Ethical Interpretation

The paradox formally answers the following questions:

- what can we know about the true limits of longtermist games?,
- what can we know about the longterm effects of our ethical and epistemic beliefs?

The paradox also tells us what assumptions need to be made in order to have a (consistent) model for absolute longtermism.

Is longermism possible then?

Few More (unbaked) Pointers

Logic and Longtermism

Modal logic has powerful tools to express various properties.

Linear Temporal Logic, for example, can express

- `safety properties` which says something bad never happens in the future ($\Box\neg F$),
- `liveness properties` which says something good keeps happening in the future ($\Box(F \to \Diamond F')$).

This is interesting. It shows theoretically how longtermism can bridge to program evaluation.

Long term evaluation of a moral action can be computed using certain predicates in the language of modal logic.

A longtermist action must have a liveness property.

Digital Ethics as Longtermism

In the age of data, arguably, longtermism can be supported by big-data.

All data are created equally, all data matter equally, and what makes it “big-data” is that *all* data need to be considered.

This is an amazing playground for longtermism, relating to major debates on privacy, piracy and data ownership.

Thank you!

can@canbaskent.net canbaskent.net/logic @topologically