# Models of decision-making based on logical counterfactuals

Vladimir Slepnev

# Overview

- Describe a simple decision problem
- Solve it in an overcomplicated way
- Generalize the approach
- Solve some more problems
- Give an outline of further research

# A simple decision problem

“Would you like some chocolate?”

- Yes → you get some chocolate.
- No → you don’t.

# A simple decision problem

“Would you like some chocolate?”

- Yes → you get some chocolate.
- No → you don’t.

Desiderata for a model:

- Two mathematical objects: U (universe) and A (agent).
- Both U and A should be “completely deterministic”.
- The description of U should “contain” the description of A.
- The descriptions of both U and A should be “completely known” to A.
- A’s decision should be based on “reasoning” about U and A.

# Our proposed model

- U and A are sentences in Peano arithmetic (PA) without free variables.
- The truth value of A indicates whether the agent says "yes" or "no".
- The truth value of U indicates whether the agent gets chocolate or not.

# Our proposed model

- U and A are sentences in Peano arithmetic (PA) without free variables.
- The truth value of A indicates whether the agent says "yes" or "no".
- The truth value of U indicates whether the agent gets chocolate or not.

A mutually recursive definition of U and A:

- $U \leftrightarrow A$

  *“If you say "yes", I will give you chocolate, otherwise I won't.”*
- $A \leftrightarrow Prov(\ulcorner A \rightarrow U \urcorner)$

  *“If I can prove that saying “yes” leads to chocolate, then I say “yes”, otherwise “no”.”*

All self-references occur within Gödel number quotes, therefore such U and A exist, by the Diagonal Lemma.

# Analysis

$$U \leftrightarrow A$$
$$A \leftrightarrow Prov(\ulcorner A \rightarrow U \urcorner)$$

It’s easy to prove that U and A are both true.

# Analysis

$$U \leftrightarrow A$$
$$A \leftrightarrow Prov(\ulcorner A \to U \urcorner)$$

It's easy to prove that U and A are both true.

What if we changed the problem a little? Reward “no” with chocolate:

$$U \leftrightarrow \neg A$$
$$A \leftrightarrow Prov(\ulcorner A \to U \urcorner)$$

Now A is false (as long as PA is consistent), and U is again true.

It feels like A is trying to make U true, in order to get some chocolate :-)

# But does it generalize?

- Many possible outcomes
- Many possible actions
- Many possible worlds
- Probabilistic strategies
- Reacting to observations
- Multiple instances of yourself
- Multiple competing agents
- Various kinds of uncertainty
- ...

# Newcomb’s problem

- There are two closed boxes in front of me.
- I can take either box 1 and box 2 (“two-box”), or only box 2 (“one-box”).
- Before the experiment, a perfect predictor predicted my action.
- The information from the prediction was used to fill the boxes.
- Box 1 always contains $1000.
- Box 2 contains $1000000 iff the predictor predicted that I would one-box.

# Newcomb’s problem

We will define these sentences in PA:

- A is true iff the agent one-boxes.
- P is true iff the predictor predicted that the agent would one-box.
- $B_1$ is true iff the agent gets the $1000 from box 1.
- $B_2$ is true iff the agent gets the $1000000 from box 2.

We will use these equations:

- $P \leftrightarrow A$
- $B_1 \leftrightarrow \neg A$
- $B_2 \leftrightarrow P$
- $A \leftrightarrow ?$

# Newcomb's problem

- $P \leftrightarrow A$

  *"The predictor predicts that I one-box iff I actually one-box."*

- $B_1 \leftrightarrow \neg A$

  *"I get the contents of the first box iff I two-box."*

- $B_2 \leftrightarrow P$

  *"I get the contents of the second box iff the predictor predicted that I would one-box."*

- $A \leftrightarrow ?$

  *"If I can get the contents of both boxes by one-boxing, then I one-box;*
  *otherwise, if I can get both boxes by two-boxing, then I two-box;*
  *otherwise, if I can get only box 2 by one-boxing, then I one-box;*
  *otherwise, if I can get only box 2 by two-boxing, then I two-box;*
  *otherwise, if I can get only box 1 by one-boxing, then I one-box;*
  *otherwise I two-box."*

# Newcomb's problem

The completed equations:

$$
\begin{aligned}
&P \leftrightarrow A \\
&B_1 \leftrightarrow \neg A \\
&B_2 \leftrightarrow P \\
&A \leftrightarrow (Prov(\ulcorner A \rightarrow B_1 \wedge B_2 \urcorner) \vee \\
&\qquad (\neg Prov(\ulcorner \neg A \rightarrow B_1 \wedge B_2 \urcorner) \wedge \\
&\qquad (Prov(\ulcorner A \rightarrow \neg B_1 \wedge B_2 \urcorner) \vee \\
&\qquad (\neg Prov(\ulcorner \neg A \rightarrow \neg B_1 \wedge B_2 \urcorner) \wedge \\
&\qquad (Prov(\ulcorner A \rightarrow B_1 \wedge \neg B_2 \urcorner))))))
\end{aligned}
$$

It's easy to prove that A is true, $B_1$ is false, and $B_2$ is true.

Thus, our approach favors one-boxing.

# Absent-minded driver problem

- To get home from work, you need to pass two identical intersections.
- At each intersection you can either continue or exit.
- At the first intersection you need to continue.
- At the second intersection you need to exit.
- You're absent-minded and can't remember which intersection you're at.
- To allow probabilistic choices, you observe a coinflip at each intersection.
- What strategy gives you the best chance of getting home?

(Slightly modified from Piccione and Rubinstein, 1997)

# Absent-minded driver problem

We will define these sentences in PA:

- $A_1$ is true iff you continue in case of heads
- $A_2$ is true iff you continue in case of tails
- $U_{11}$ is true iff you get home in case of (heads, heads)
- Similar for $U_{12}$, $U_{21}$, $U_{22}$

$U_{11} \leftrightarrow U_{22} \leftrightarrow \bot$

$U_{12} \leftrightarrow A_1 \land \neg A_2$

$U_{21} \leftrightarrow \neg A_1 \land A_2$

$A_1 \leftrightarrow ?$

$A_2 \leftrightarrow ?$

# Absent-minded driver problem

$A_1 \leftrightarrow ?$
$A_2 \leftrightarrow ?$

*"If making $A_1$ and $A_2$ true will make all $U_{ij}$ true, then I'll make $A_1$ and $A_2$ true;*
*otherwise, if making $A_1$ true and $A_2$ false will make all $U_{ij}$ true, then I'll make $A_1$ true and $A_2$ false;*
*{...}*
*otherwise, if making $A_1$ and $A_2$ true will make exactly three of $U_{ij}$ true, then I'll make $A_1$ and $A_2$ true;*
*{...}"*

The equations begin like this:

$$A_1 \leftrightarrow Prov(\ulcorner A_1 \wedge A_2 \rightarrow U_{11} \wedge U_{12} \wedge U_{21} \wedge U_{22} \urcorner) \vee \ldots$$
$$A_2 \leftrightarrow Prov(\ulcorner A_1 \wedge A_2 \rightarrow U_{11} \wedge U_{12} \wedge U_{21} \wedge U_{22} \urcorner) \vee \ldots$$

# Other proposed models

Using Gödel-Löb provability logic instead of PA:

- Use □ instead of *Prov*
- Use modal fixed points instead of the Diagonal Lemma
- Equivalent to the PA approach, because GL is adequate for PA (Solovay)
- Decidable!

Using computer programs that look for proofs, instead of arithmetic formulas:

- Chronologically, the first approach we came up with
- If programs have access to provability oracles, this is also equivalent to PA
- If programs enumerate proofs up to a fixed size, it's "almost" equivalent
- Undecidable in general

# Further work

From decision theory to game theory

- What if there are multiple agents proving things about each other?
- What do you want other agents to prove about you?
- How does “proof warfare” influence cooperation, bargaining, blackmail...

From perfect certainty to uncertainty

- How do you handle uncertainty about mathematical facts?
- How do you handle uncertainty about your description of yourself?
- How do you handle uncertainty about your values?

# Questions?

Thank you :-)

vladimir.slepnev@gmail.com

http://lesswrong.com/user/cousin_it/submitted

http://agentfoundations.org/submitted?id=Vladimir_Slepnev