Newcomb's Problem: One Box xor Two Boxes, Which is Rational?

Selmer Bringsjord

Department of Cognitive Science

Department of Computer Science

Lally School of Management

Rensselaer AI & Reasoning Lab

RPI

Selmer.Bringsjord@gmail.com

Are Humans Rational?

10/28/19

Rensselaer AI and Reasoning Lab

The Paradoxes: Our Coverage, & RPI Context

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

[**The Knowability Paradox**: Active area of sponsored r&d for RAIR Lab. Will be in a documentary film on AI based in part on a visit to RAIR Lab. Offered 10/21/19 (w/ optional content.)]

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

[**The Knowability Paradox**: Active area of sponsored r&d for RAIR Lab. Will be in a documentary film on AI based in part on a visit to RAIR Lab. Offered 10/21/19 (w/ optional content.)]

Newcomb's Problem: Active area of *on-spec* r&d for RAIR Lab. Gist of Selmer's proposed solution conveyed *today* (10/28/19).

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

[**The Knowability Paradox**: Active area of sponsored r&d for RAIR Lab. Will be in a documentary film on AI based in part on a visit to RAIR Lab. Offered 10/21/19 (w/ optional content.)]

Newcomb's Problem: Active area of *on-spec* r&d for RAIR Lab. Gist of Selmer's proposed solution conveyed *today* (10/28/19).

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

[**The Knowability Paradox**: Active area of sponsored r&d for RAIR Lab. Will be in a documentary film on AI based in part on a visit to RAIR Lab. Offered 10/21/19 (w/ optional content.)]

Today

Today

Newcomb's Problem: Active area of *on-spec* r&d for RAIR Lab. Gist of Selmer's proposed solution conveyed *today* (10/28/19).

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

[**The Knowability Paradox**: Active area of sponsored r&d for RAIR Lab. Will be in a documentary film on AI based in part on a visit to RAIR Lab. Offered 10/21/19 (w/ optional content.)]

Today

Today

Newcomb's Problem: Active area of *on-spec* r&d for RAIR Lab. Gist of Selmer's proposed solution conveyed *today* (10/28/19).

The Lottery Paradox: Solved by RAIR Lab! Next class (10/31/19). And the St. Petersburg Paradox too?

The Paradoxes: Our Coverage, & RPI Context

The Liar: You're on your own :). Presented 10/21/19.

The Barber: Flat-out solved, period (via ZFC foundation for math); take *Intro to Logic*. Presented 10/21/19.

[**The Knowability Paradox**: Active area of sponsored r&d for RAIR Lab. Will be in a documentary film on AI based in part on a visit to RAIR Lab. Offered 10/21/19 (w/ optional content.)]

Today Today

Newcomb's Problem: Active area of *on-spec* r&d for RAIR Lab. Gist of Selmer's proposed solution conveyed *today* (10/28/19).

The Lottery Paradox: Solved by RAIR Lab! Next class (10/31/19). And the St. Petersburg Paradox too?

The Paradoxes of Time Travel: Grandfather & "Looping Painter": Painter is an active research area, sponsored by Air Force; Naveen Sundar G., S Bringsjord. Covered 11/4/19.

The Setup …

Original Version

Original Version

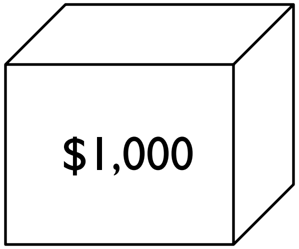
You face two boxes.

Original Version

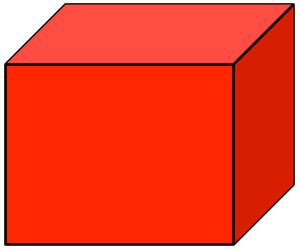
You face two boxes.
b1 is transparent, and contains $1,000.

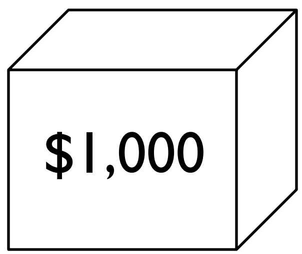
Original Version

You face two boxes.
b1 is transparent, and contains $1,000.
b2, on the other hand, is opaque.

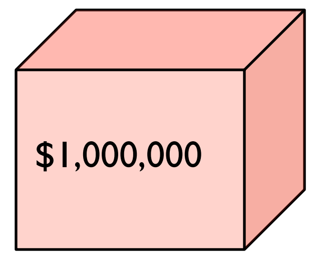


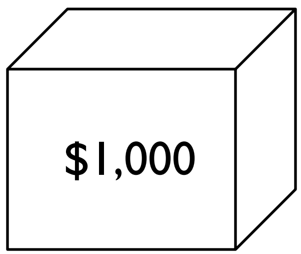
b2 either contains...



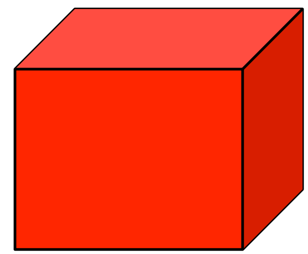


b2 either contains...

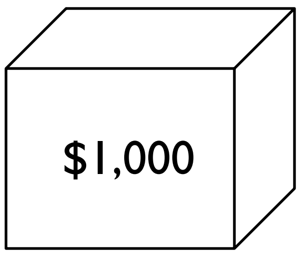




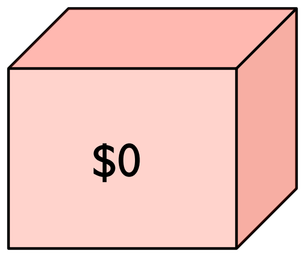
or...




$1,000
or...
$0



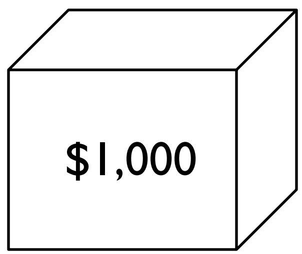
or...

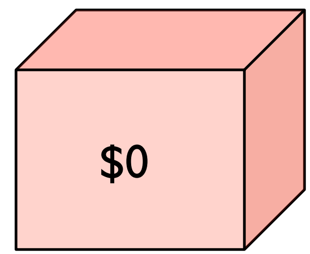


And whether or not there's $1M in b2 is a 50/50 proposition.

You can either pick box b2, or b2 *and* b1. What's rational to do?

or...





And whether or not there's $1M in b2 is a 50/50 proposition.

The Optimality Principle

When choosing between alternative actions a_1 and a_2, rationality dictates choosing that action that maximizes expected value, computed by multiplying the value of each outcome that can result from each action by the probability that it will occur, adding the results together, and selecting the action associated with the higher utility.

The Optimality Principle

When choosing between alternative actions a_1 and a_2, rationality dictates choosing that action that maximizes expected value, computed by multiplying the value of each outcome that can result from each action by the probability that it will occur, adding the results together, and selecting the action associated with the higher utility.

(This principle is taught to students in every introductory economics or decision-theory class, and is at least usually a key thing to follow in the pursuit of rational behavior.)

Easy w/o the Predictor:

	b2 empty	b2 filled
take b1 & b2	$1,000	$1,001,000
take only b2	0	$1,000,000

	b2 empty	b2 filled
take b1 & b2	$1,000	$1,001,000
take only b2	0	$1,000,000

Use The Optimality Principle!

	b2 empty	b2 filled
take b1 & b2	$1,000	$1,001,000
take only b2	0	$1,000,000

Use The Optimality Principle!

	b2 empty	b2 filled
take b1 & b2	\$1,000	\$1,001,000
take only b2	0	\$1,000,000

take b1 & b2 = 1(1000) + .5(1,000,000) = 1000 + 500,000 = 501,000

Use The Optimality Principle!

	b2 empty	b2 filled
take b1 & b2	$1,000	$1,001,000
take only b2	0	$1,000,000

take b1 & b2 = 1(1000) + .5(1,000,000) = 1000 + 500,000 = 501,000

take only b2 = 1(0) + .5(1,000,000) = 0 + 500,000 = 500,000

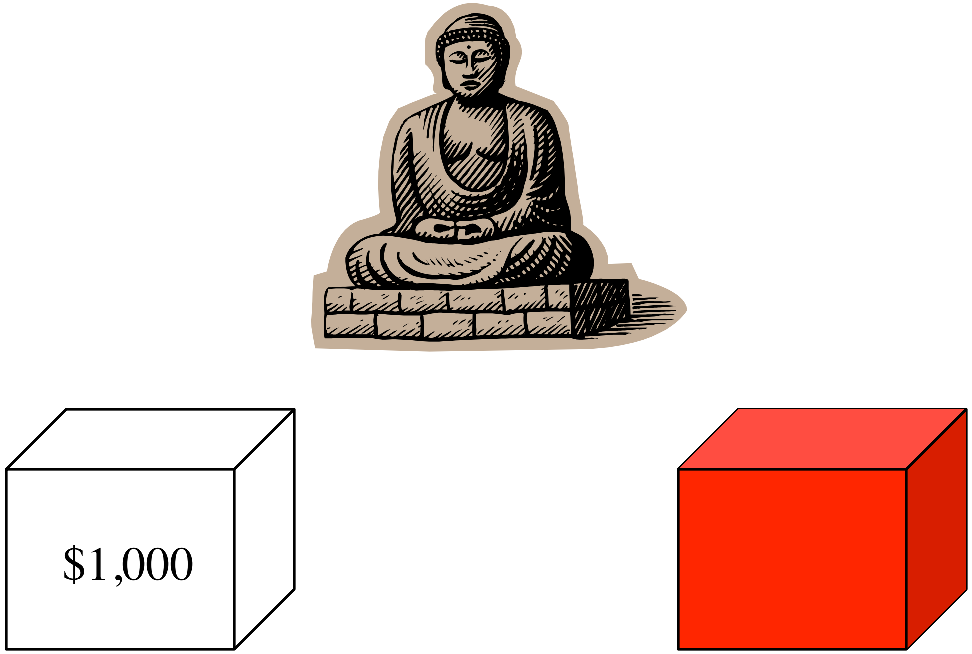
Use The Optimality Principle!

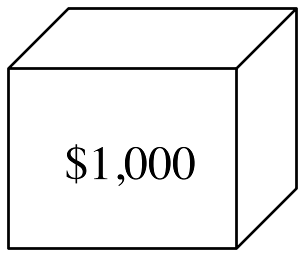
	b2 empty	b2 filled
take b1 & b2	\$1,000	\$1,001,000
take only b2	0	\$1,000,000

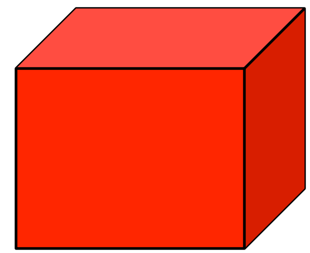
take b1 & b2 = 1(1000) + .5(1,000,000) = 1000 + 500,000 = 501,000

take only b2 = 1(0) + .5(1,000,000) = 0 + 500,000 = 500,000

$$\sum_{i=1}^{n} p_i \times u(O_i^a)$$


$1,000





The catch is that a being with preternaturally accurate powers of prediction (on the basis, say, of brain scans), scanned your brain before your decision point, and if he predicted that you would take both boxes, he left b2 empty, while if he predicted you'd take only b2, he put the $1,000,000 in it. ('Preternatural accuracy' can be unpacked by statistical facts as stupendous as you wish. E.g., the being can be batting 1000 in previous predictions about future human actions.)

So where's the paradox? …

So where's the paradox? …

Argument #1 (Max R)

The super-being is super because in the past he has proved to be nearly invariably correct in his predictions about what humans are going to do. So it's highly probable that his prediction is going to be correct in my case. Given this, if I take b2, it's exceedingly likely that he will have predicted that I would do so, and it is thus highly likely that I will thus receive $1,000,000. On the other hand, if I take both b1 and b2, it is almost certain that he will once again have predicted that that is what I would do, and hence I will receive only $1,000, and I won't get rich. So, by the Optimality Principle I should take b2!

Different Calculation (e.g.)
Based on the Optimality Principle

	b2 empty	b2 filled
take b1 & b2	$1,000	$1,001,000
take only b2	0	$1,000,000

Different Calculation (e.g.) Based on the Optimality Principle

	b2 empty	b2 filled
take b1 & b2	$1,000	$1,001,000
take only b2	0	$1,000,000

take b1 & b2 = 1(1,000) + .1(1,000,000) = 1,000 + 100,000 = 101,000

Different Calculation (e.g.) Based on the Optimality Principle

	b2 empty	b2 filled
take b1 & b2	\$1,000	\$1,001,000
take only b2	0	\$1,000,000

take b1 & b2 = 1(1,000) + .1(1,000,000) = 1,000 + 100,000 = 101,000

take only b2 = 1(0) + .9(1,000,000) = 0 + 900,000 = 900,000

Different Calculation (e.g.) Based on the Optimality Principle

	b2 empty	b2 filled
take b1 & b2	\$1,000	\$1,001,000
take only b2	0	\$1,000,000

take b1 & b2 = 1(1,000) + .1(1,000,000) = 1,000 + 100,000 = 101,000

take only b2 = 1(0) + .9(1,000,000) = 0 + 900,000 = 900,000

$$\sum_{i=1}^{n} p_i \times u(O_i^a)$$

The Dominance Principle

If in some case S your doing action a_1 rather than action a_2 will secure a larger payoff, and if in case $\sim S$ your doing a_1 rather than a_2 will likewise secure a larger payoff, you should do a_1 regardless of whether or not S holds.

Argument #2

The prediction has been made (a week ago, a month ago, then years ago, ...), and what's in the boxes before me isn't going to change. So, either it's just $1,000 in b1 (Case 1), or that plus $1,000,000 in b2 (Case 2). Either way, if I take both boxes I will be the richer for it: If Case 1 holds, I get $1,000 instead of nothing; if Case 2 holds, I get $1,001,000 instead of $1,000,000. So by the Dominance Principle I should take both boxes.

Uh oh!!

The Setup

Moke: A drug, quite pleasurable without any negative side-effects if taken in moderation.

Moking: To take moke.

The Genetic Twist: There's a hidden gene, M-G, present in many people, which causes a desire in these people to moke, and also (in separate etiology/causality) statistically predisposes these people to getting blood clots that can sometimes travel to the brain.

Uh oh:

Uh oh:

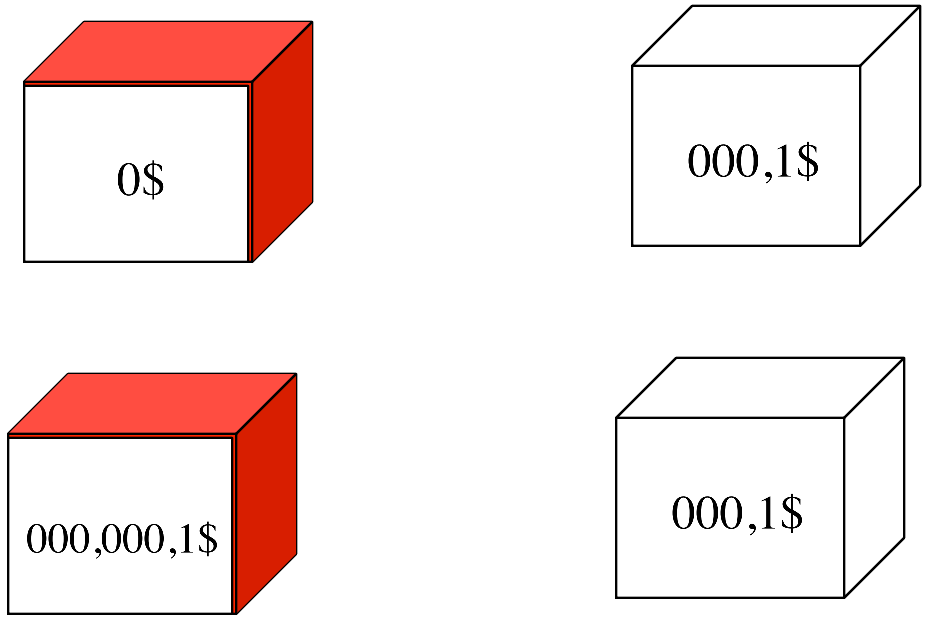
Reasoning based on the Optimality Principle implies that a rational person shouldn't moke. But that seems stupid, since moking doesn't *cause* blood clots, and you already have the M-G gene or not, whether or not you moke! Since moking is quite pleasant, you should go ahead and enjoy the activity of moking (by the Dominance Principle).

Selmer:
"A t*hird* argument rules,
and trumps the other two…"

As Nozick points out all the way back in 1969 when introducing the Newcomb Problem to the world, suppose someone ...

Case 1

Case 2



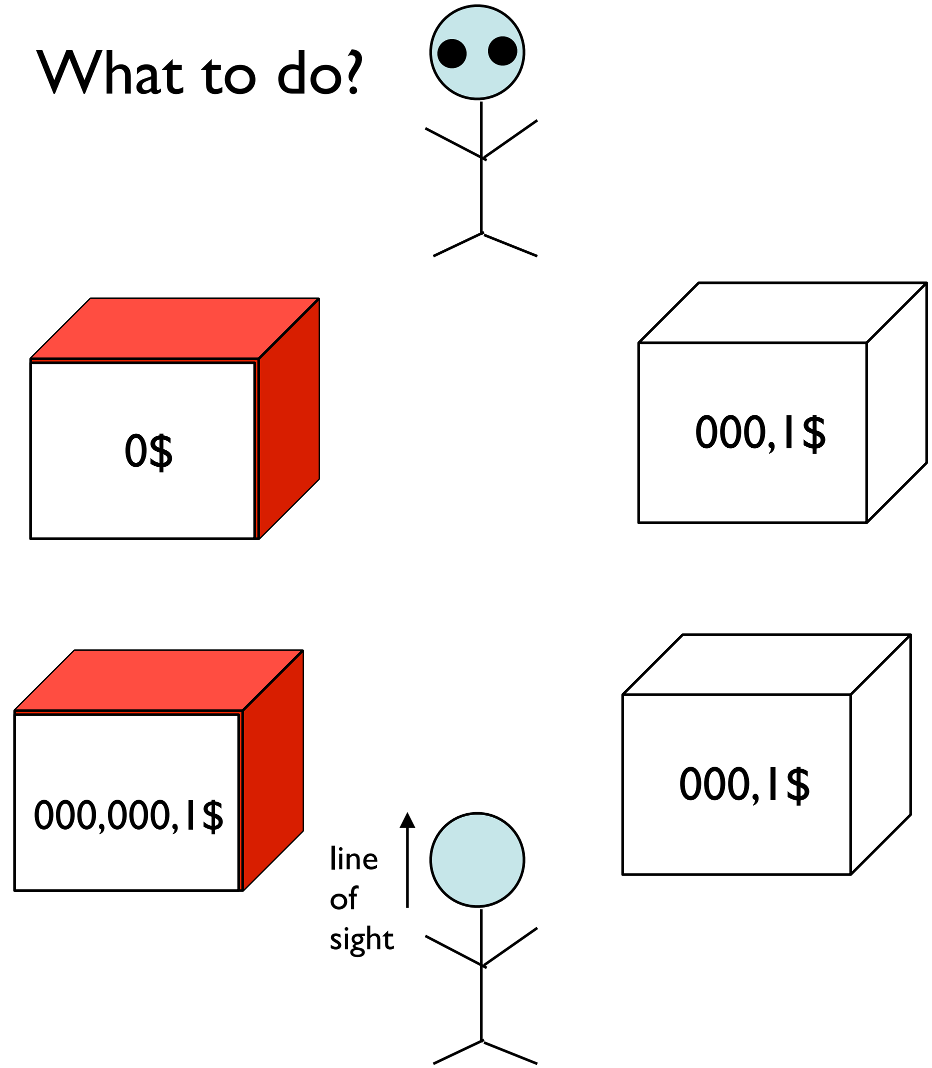
can see inside b2. Wouldn't they be (internally) shouting to you: Take both boxes?!

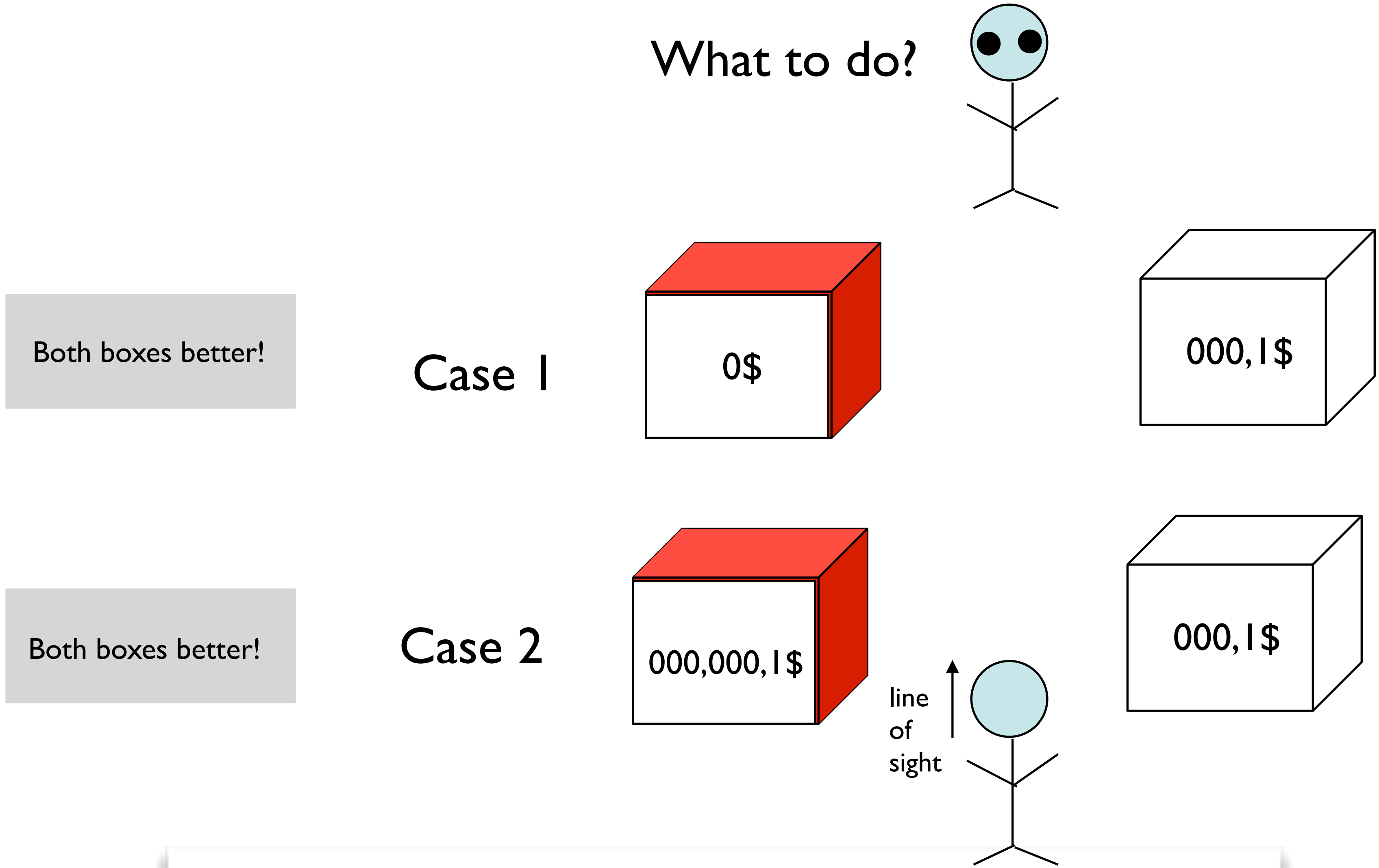
Both boxes better!

Case 1

Both boxes better!

Case 2





One ought to choose at *t*, from among *n* competing options, the one which is utility-best from the perspective of a god-like agent who knows all the relevant "occurrent" information at *t*, and also knows all the consequences of selecting each of the options.

Further Reading

- A seminal paper on Newcomb's Problem appeared fairly recently in the journal *Synthese*:
  - Pollock, J. (2010) "A Resource-Bounded Agent Addresses the Newcomb Problem" *Synthese* **176.1**: 57–82.
  - This truly excellent paper, ultimately a defense of two-boxing, is available, in preprint form, at:
    - http://johnpollock.us/ftp/PAPERS/Newcomb%20Problem.pdf.
- Newcomb's Problem was originally introduced in 1969 by Robert Nozick. Full references are provided in Pollock's paper.