A Million Dollar Game

Krikamol Muandet

Empirical Inference Department
Max Planck Institute for Intelligent Systems

October 8, 2013

A Million Dollar Game

A

$1,000

B

$0 or $1,000,000

A Million Dollar Game

A **B**

$1,000 $0 or $1,000,000

Predictor

make a prediction

1. take both boxes $\Rightarrow$ $0
2. take only B $\Rightarrow$ $1,000,000

"almost certainly" correct

A Million Dollar Game

A **B**

$1,000 $0 or $1,000,000

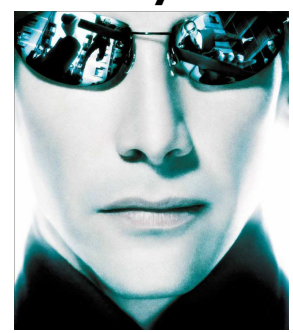
Predictor

make a prediction

1. take both boxes $\Rightarrow$ $0
2. take only B $\Rightarrow$ $1,000,000

"almost certainly" correct

Player



make a decision

1. take both boxes
2. take only B

One-Boxer vs. Two-Boxer

A

$1,000

B

$0 or $1,000,000

Take Both Boxes

Rational Player

A

$1,000

B

$0 or $1,000,000

Take Both Boxes

Rational Player

A **B**

$1,000 $0 or $1,000,000

If prediction is for both A and B to be taken.

- choosing between $1,000 (by taking A and B) and $0 (by taking just B)
- taking both boxes is obviously preferable

Take Both Boxes

Rational Player

A **B**

$1,000 $0 or $1,000,000

If prediction is for both A and B to be taken.

- choosing between $1,000 (by taking A and B) and $0 (by taking just B)
- taking both boxes is obviously preferable

If the prediction is for the player to take only B.

- taking both boxes yields $1,001,000
- taking only B yields only $1,000,000

Take Both Boxes

Rational Player

A **B**

$1,000 $0 or $1,000,000

If prediction is for both A and B to be taken.

- choosing between $1,000 (by taking A and B) and $0 (by taking just B)
- taking both boxes is obviously preferable

If the prediction is for the player to take only B.

- taking both boxes yields $1,001,000
- taking only B yields only $1,000,000

Regardless of what prediction the Predictor has made, taking both boxes yields more money.

Take Only B

Irrational Player

B

$0 or $1,000,000

Take Only B

Irrational Player

B

$0 or $1,000,000

The Predictor is almost never wrong $\Rightarrow$ ignore $0 and $1,001,000

- receive $1,000 (both boxes)
- receive $1,000,000 (only box B)

Take Only B

Irrational Player

B

$0 or $1,000,000

The Predictor is almost never wrong $\Rightarrow$ ignore $0 and $1,001,000

- receive $1,000 (both boxes)
- receive $1,000,000 (only box B)

Taking only box B is better!

Newcomb's Paradox?

- causal decision theory – precommit yourself to take one box $\rightarrow$ box B to be filled
- to take only one box, you must somehow believe that your choice can affect whether box B is empty or full - and that's unreasonable!
- toxin puzzle – one can have a reason to intend to do something without having a reason to actually do it.
- common cause problem (Burgess, Eells, etc.)
- prisoner's dilemma (Lewis)
- time machine (Craig) $\Rightarrow$ free will
- free will and determinism
- grandfather paradox – the "chooser" is not free to choose (contradictory assumptions)

Rational Decision-Making

1. take the action with the greater expected value outcome, i.e., one-box; versus
2. take the action which, conditional on the current state of the world, guarantees you a better outcome than any other action, i.e., two-box.

Real-World

Thanks!