



Anthropic Decision Theory



I think,
therefore I am



I am,
therefore... I do?



Why anthropic decisions make sense,
but anthropic probabilities don't.

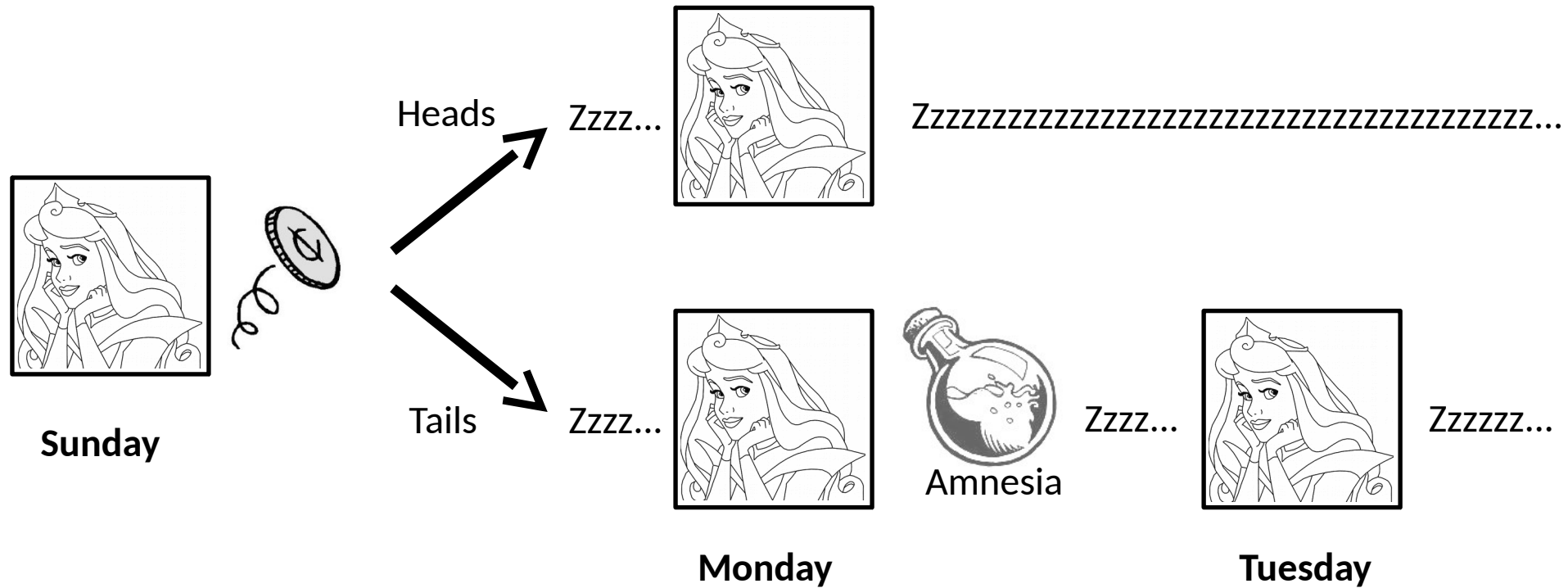
Anthropic questions



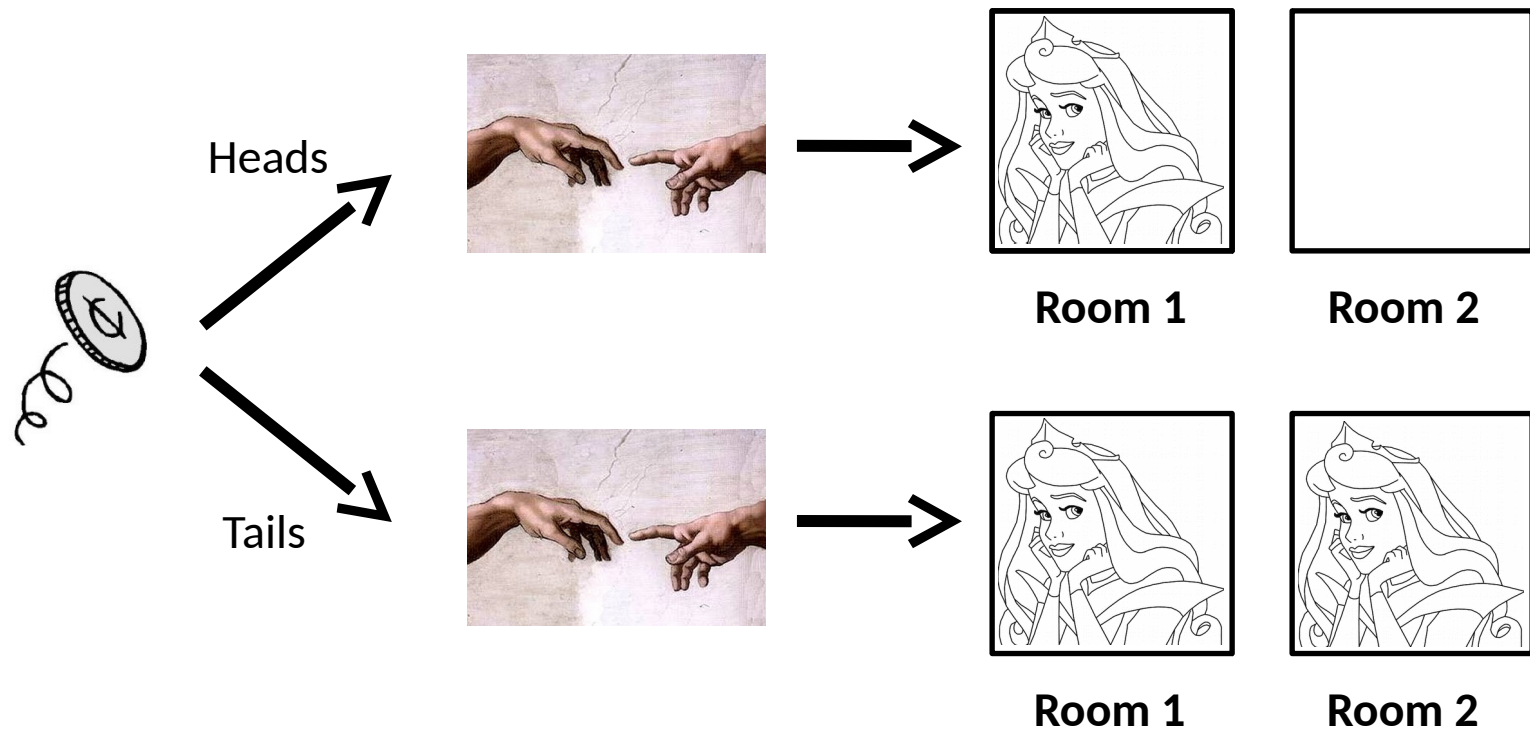
Humanity on Earth implies...

...what about the universe?

Sleeping Beauty I Amnesia



Sleeping Beauty II Incubator



Upon awakening, what is the **probability** of Heads? Of Room1?

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Those are the initial odds.

And you learn nothing new: no update.

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Those are the initial odds.

And you learn nothing new: no update.

- Thirder position: $1/3$ on heads.

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Those are the initial odds.

And you learn nothing new: no update.

- Thirder position: $1/3$ on heads.

Because “(Monday, heads)”, “(Monday, tails)”, and “(Tuesday, tails)” are indistinguishable.

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Those are the initial odds.

And you learn nothing new: no update.

- Thirder position: $1/3$ on heads.

Because “(Monday, heads)”, “(Monday, tails)”, and “(Tuesday, tails)” are indistinguishable.

~~“(Tuesday, heads)”~~ **must** tell you something.

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Self-Sampling Assumption (SSA)

- Thirder position: $1/3$ on heads.

Self-Indication Assumption (SIA)

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

Self-Sampling Assumption (SSA): An observer is randomly selected from the set of all *actually existent* observers in their reference class.

- Thirder position: $1/3$ on heads.

Self-Indication Assumption (SIA)

Standard resolutions: probability

- Halfer position: $1/2$ on heads.

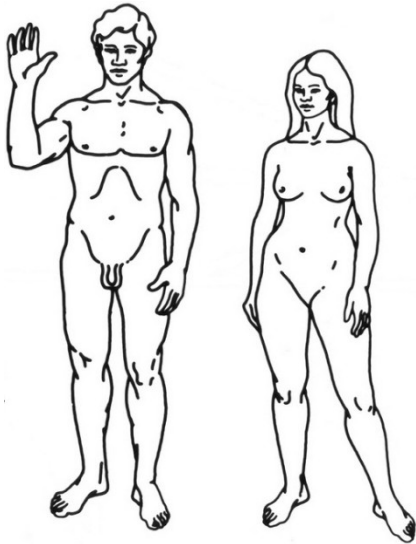
Self-Sampling Assumption (SSA): An observer is randomly selected from the set of all *actually existent* observers in their reference class.

- Thirder position: $1/3$ on heads.

Self-Indication Assumption (SIA): An observer is randomly selected from the set of all *possible* observers.

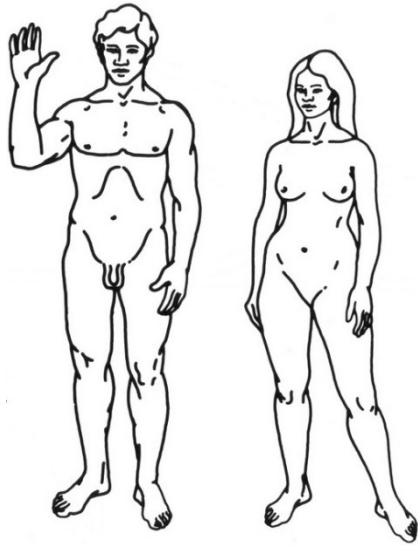
Adam and Eve paradox

SSA prefers
small universes
(present and future)

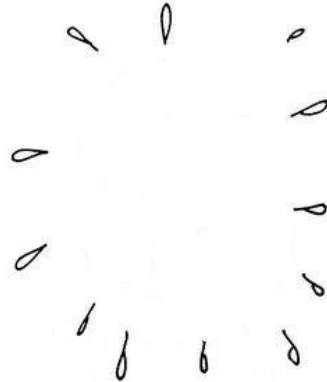


Adam and Eve paradox

SSA prefers
small universes
(present and future)

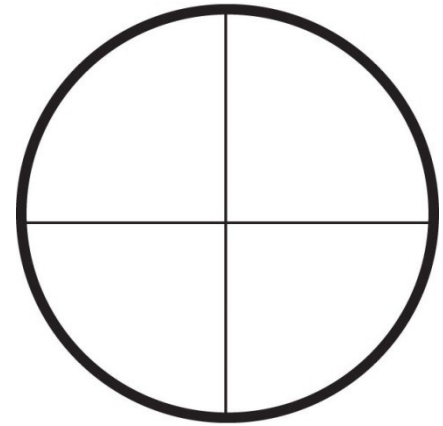
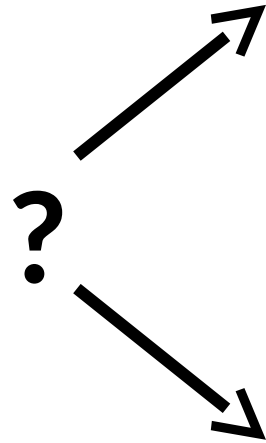
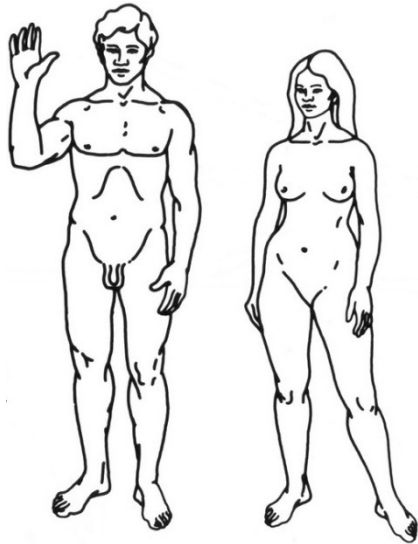


?



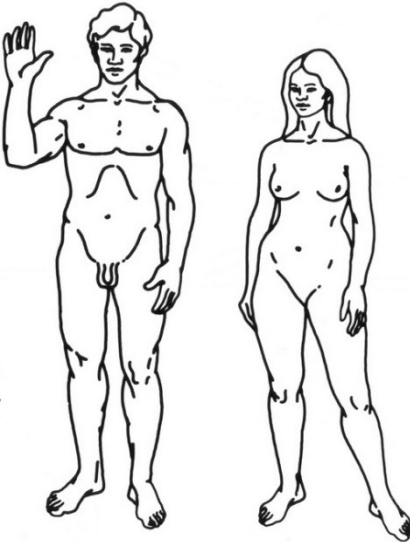
Adam and Eve paradox

SSA prefers
small universes
(present and future)

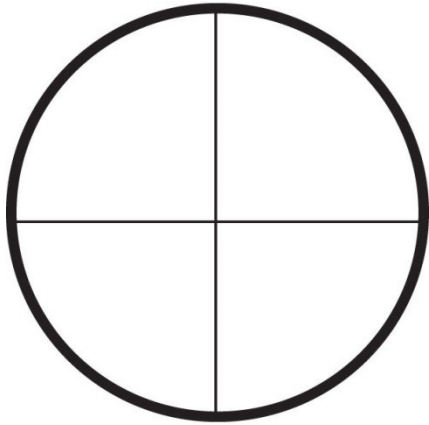
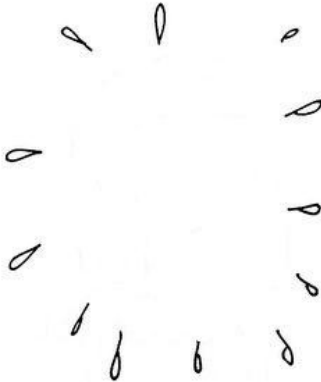
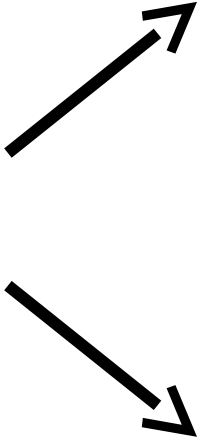


Adam and Eve paradox

SSA prefers
small universes
(present and future)

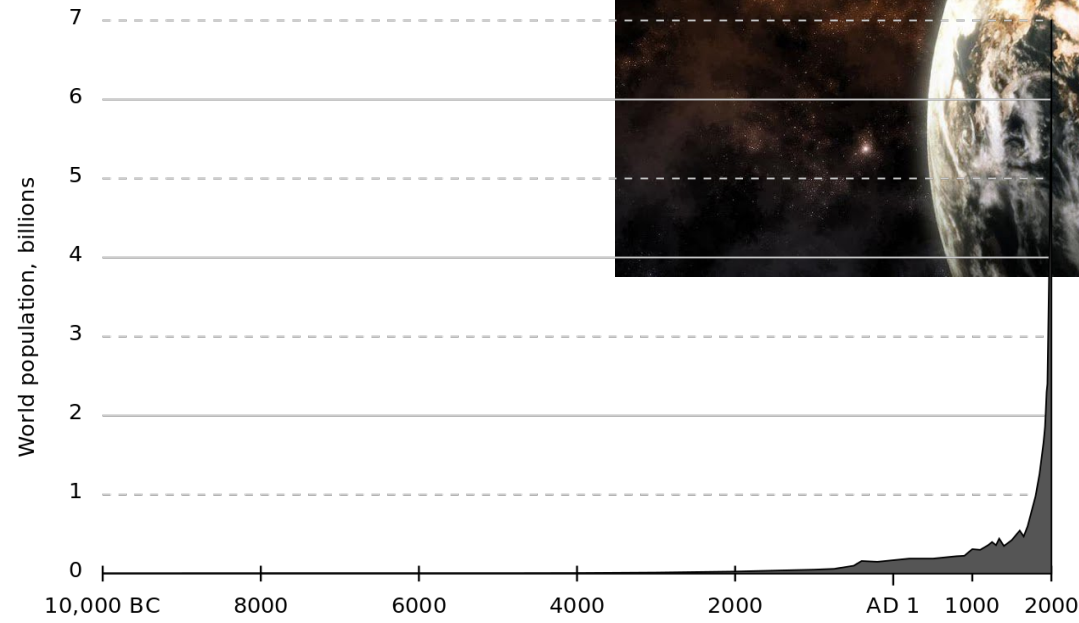


?



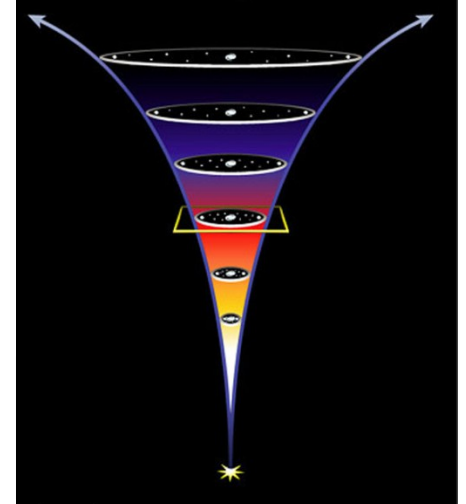
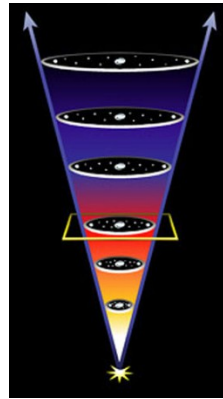
Doomsday argument

SSA prefers
small universes
(present and future)



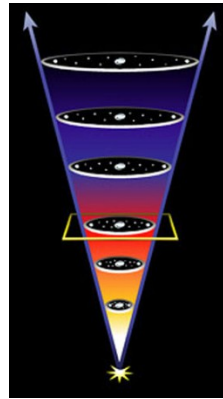
Presumptuous philosopher

SIA prefers
large universes
(present, not future)

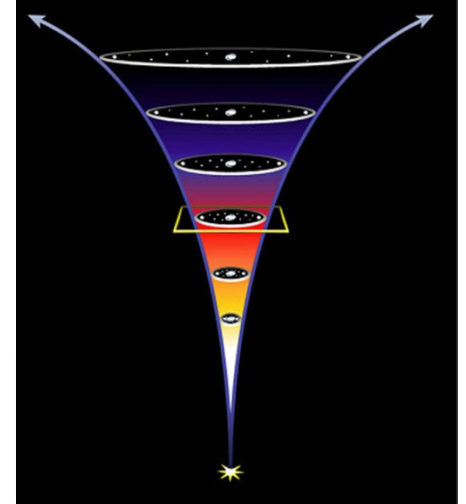


Presumptuous philosopher

SIA prefers
large universes
(present, not future)

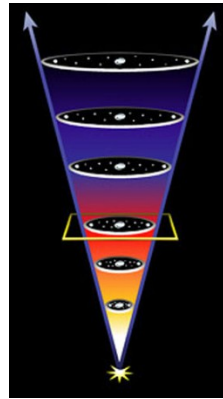


$\Lambda = ?$

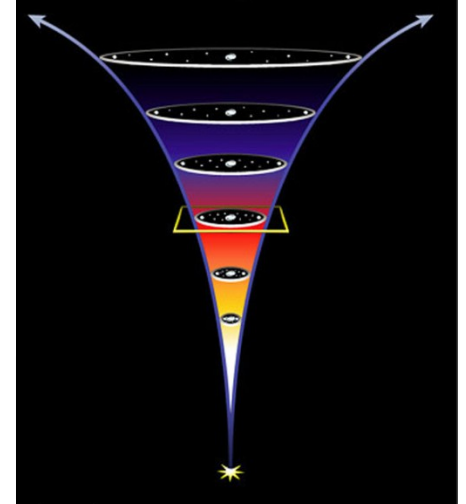


Presumptuous philosopher

SIA prefers
large universes
(present, not future)



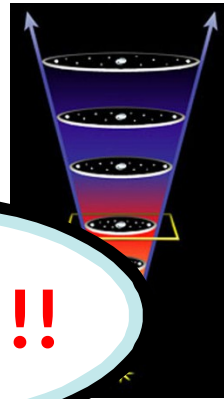
$\Lambda = ?$



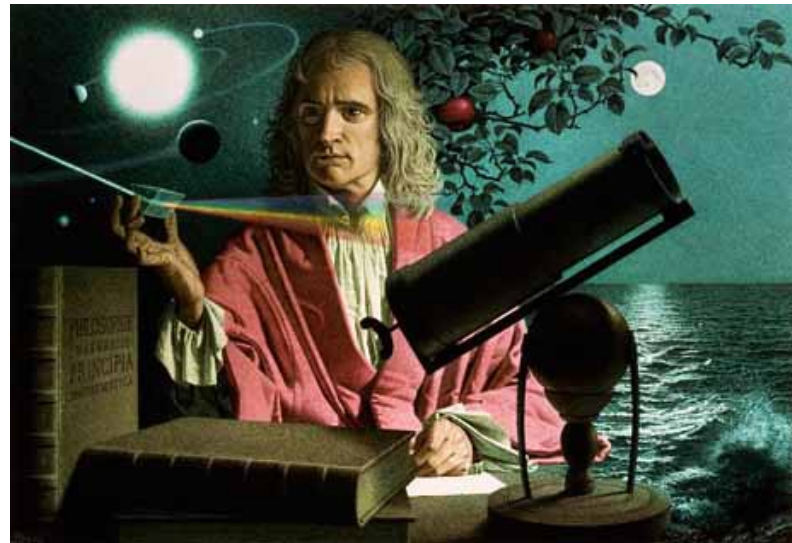
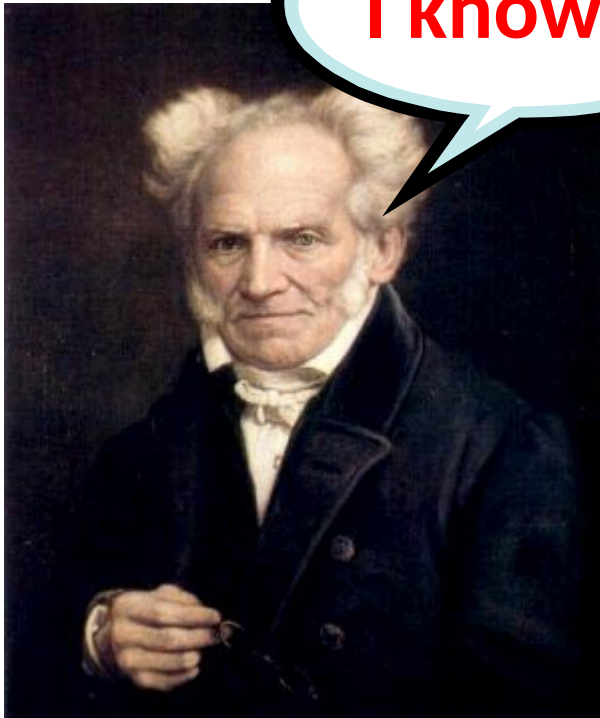
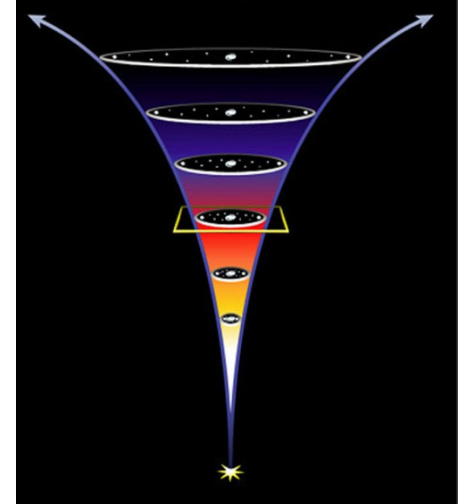
Presumptuous philosopher

SIA prefers
large universes
(present, not future)

I know!!!



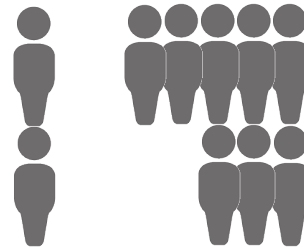
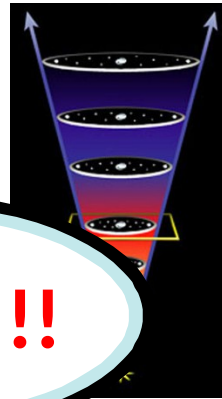
$\Lambda = ?$



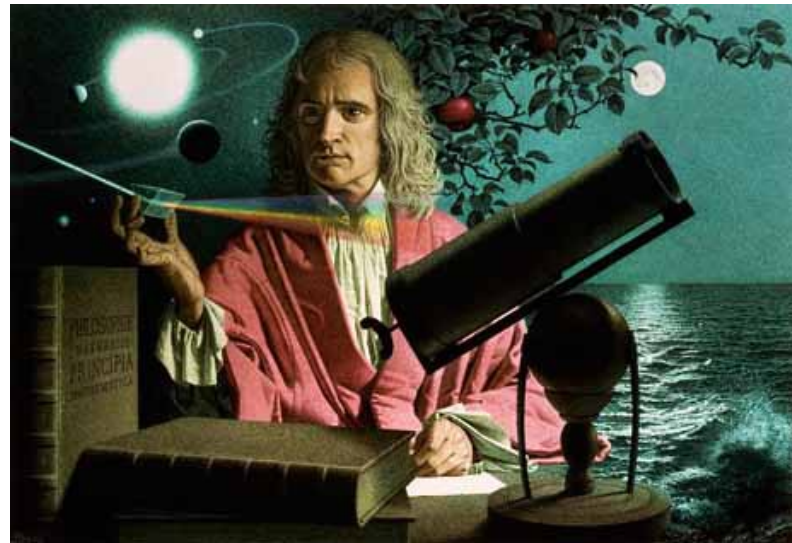
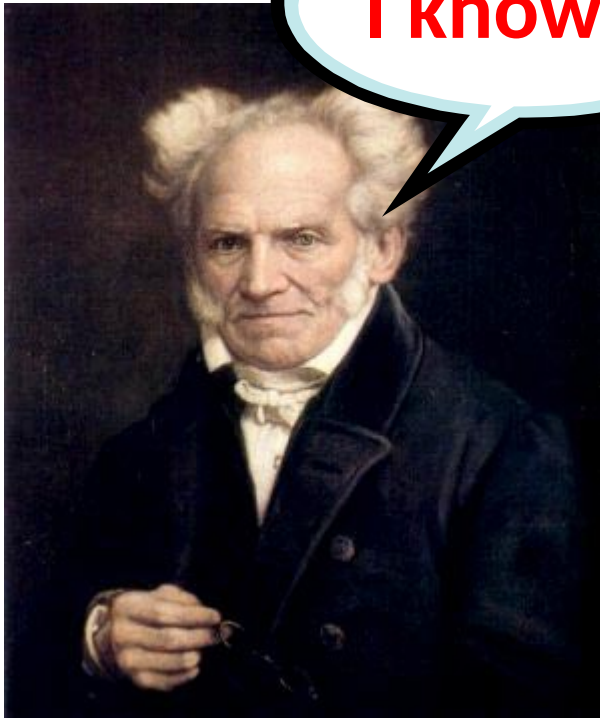
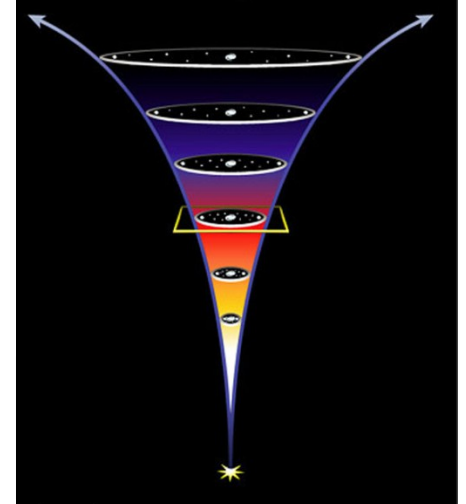
Presumptuous philosopher

SIA prefers
large universes
(present, not future)

I know!!!

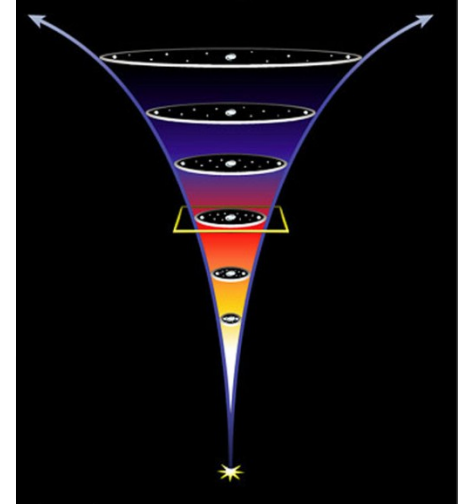
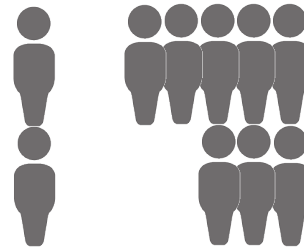
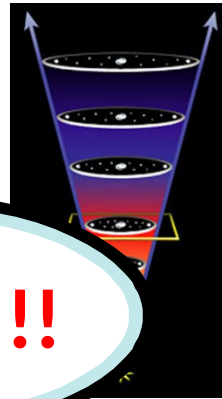


$$\Lambda = ?$$



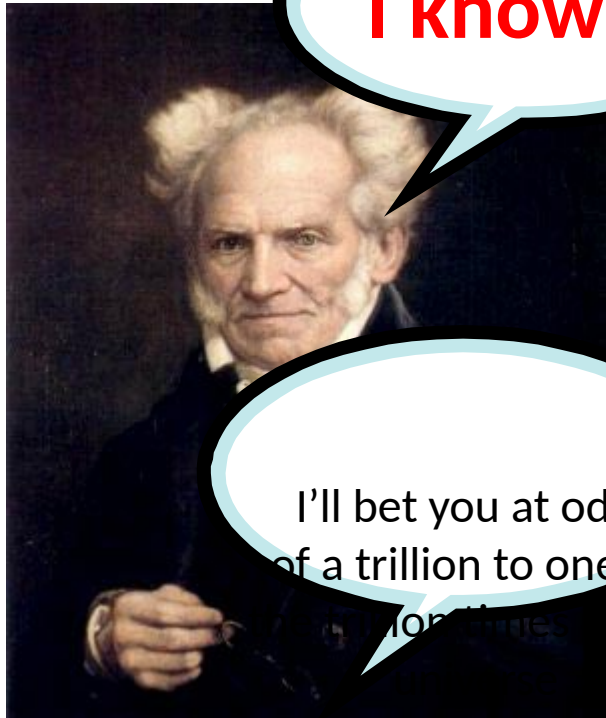
Presumptuous philosopher

SIA prefers
large universes
(present, not future)



$$\Lambda = ?$$

I know!!!

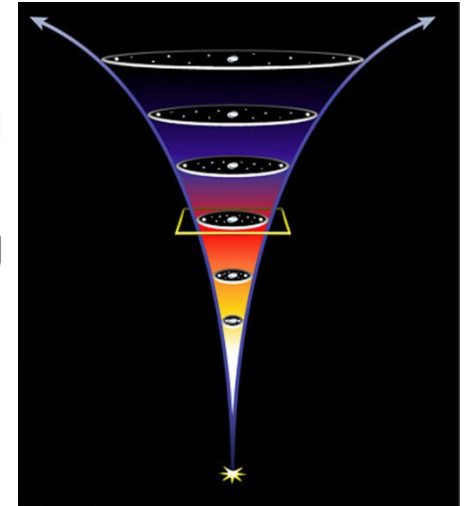
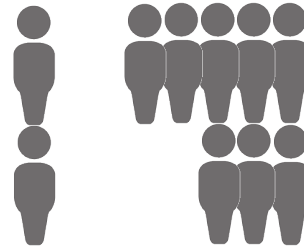
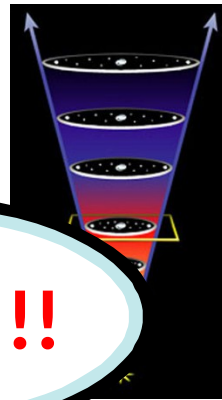


I'll bet you at odds
of a trillion to one on
non things. I gge



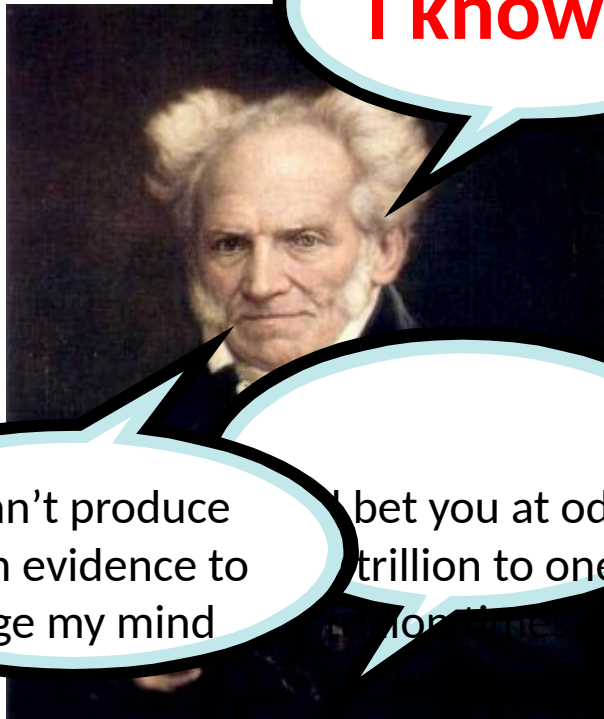
Presumptuous philosopher

SIA prefers
large universes
(present, not future)



$$\Lambda = ?$$

I know!!!



You can't produce
enough evidence to
change my mind

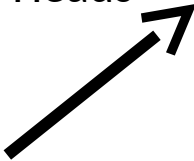
I bet you at odds
trillion to one on
something bigger



Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



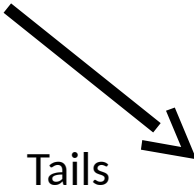
Heads




Room 1



Room 2

Tails




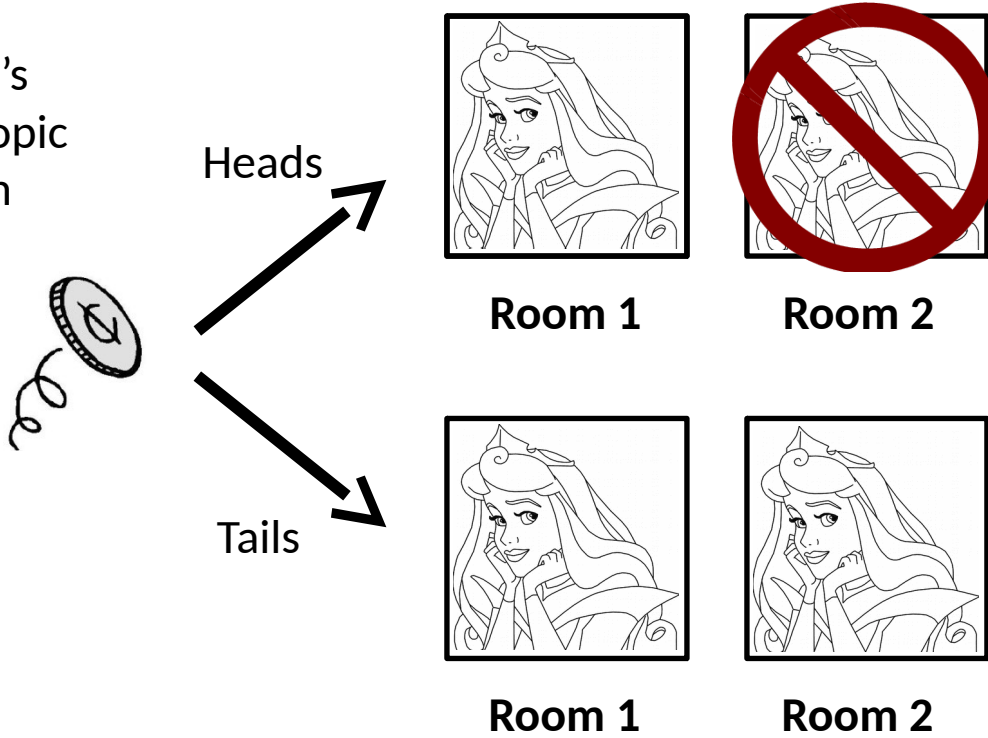
Room 1



Room 2

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Heads

Tails



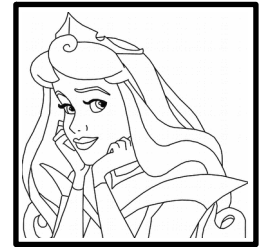
Room 1



Room 2



Room 1

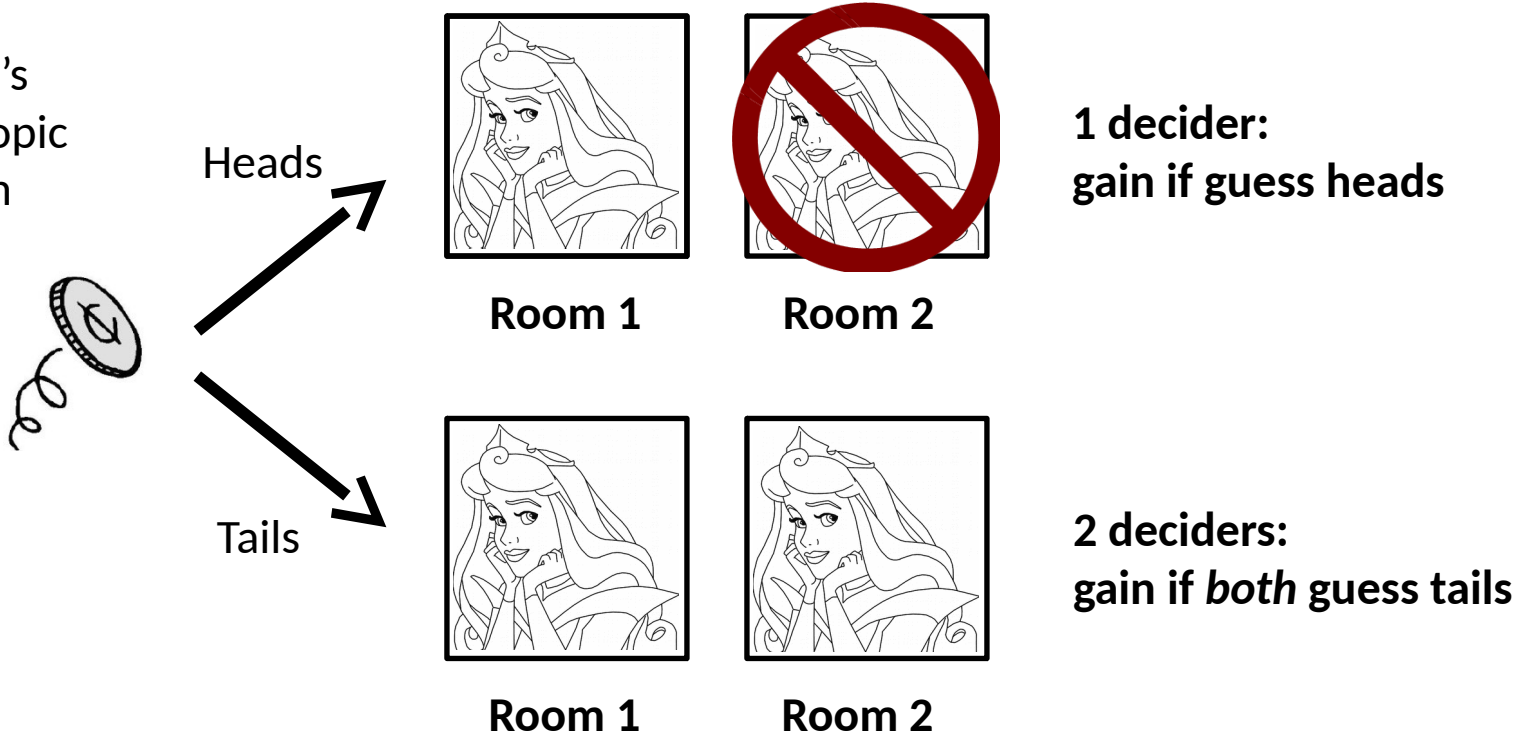


Room 2

**1 decider:
gain if guess heads**

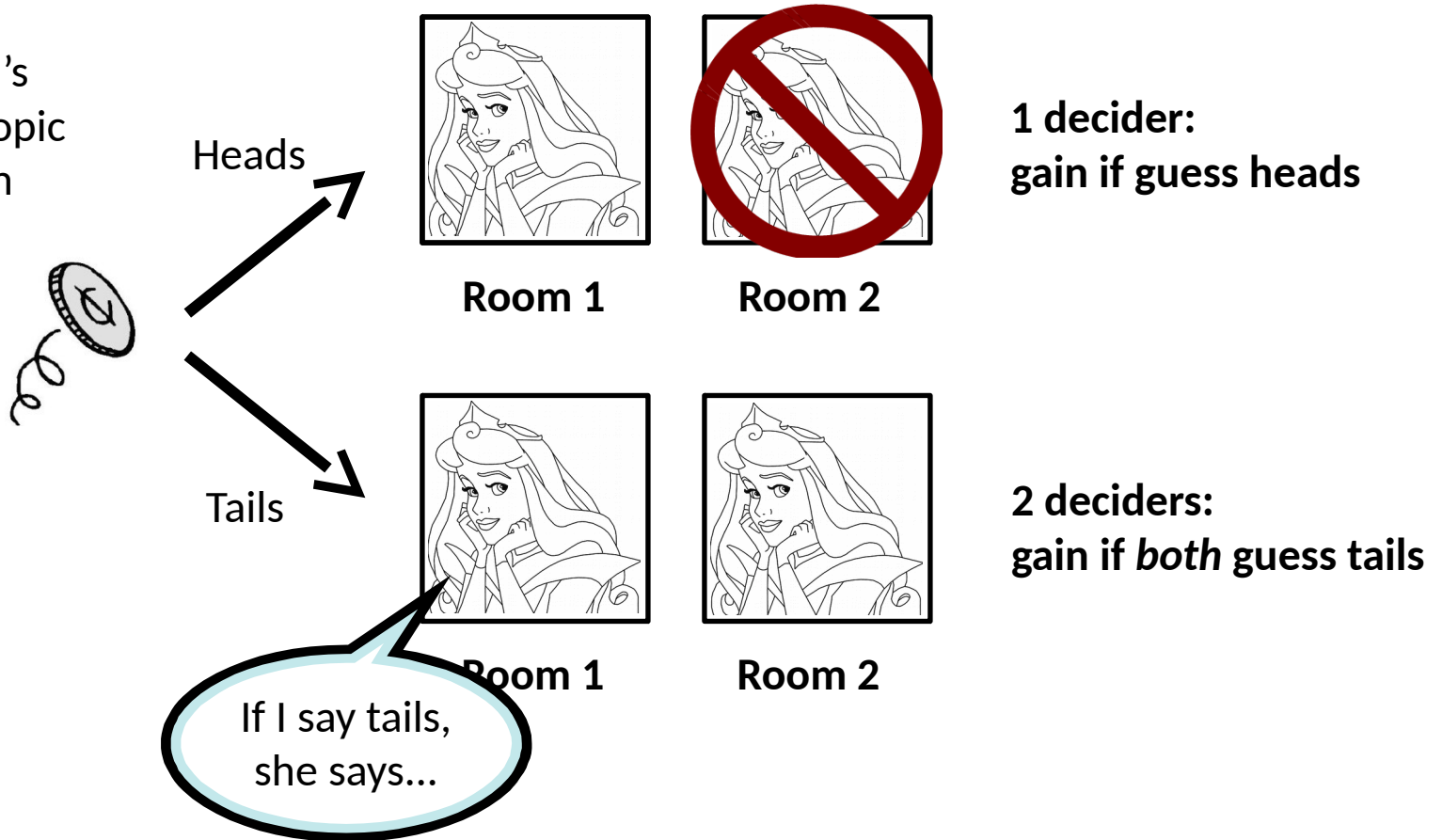
Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



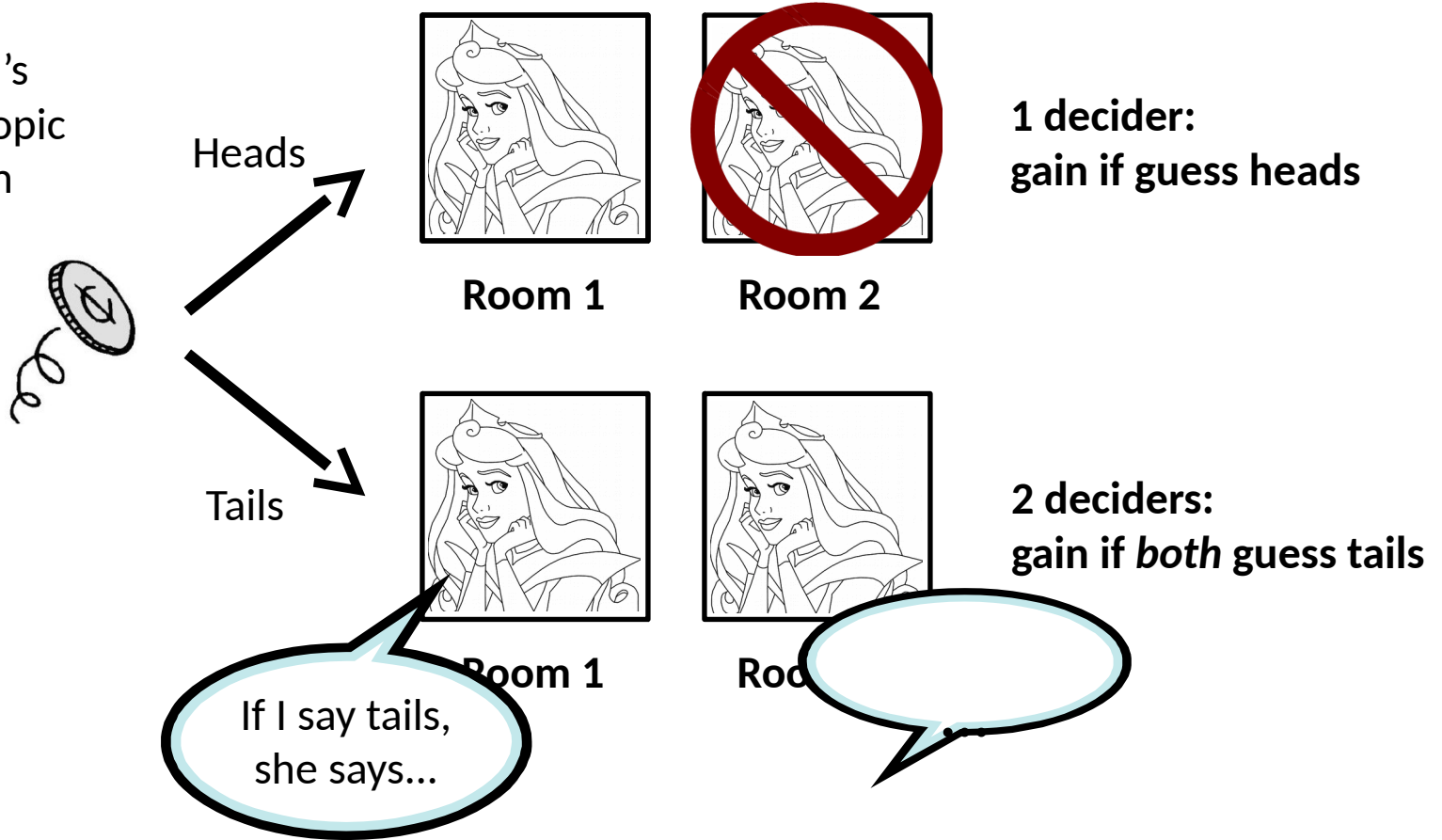
Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



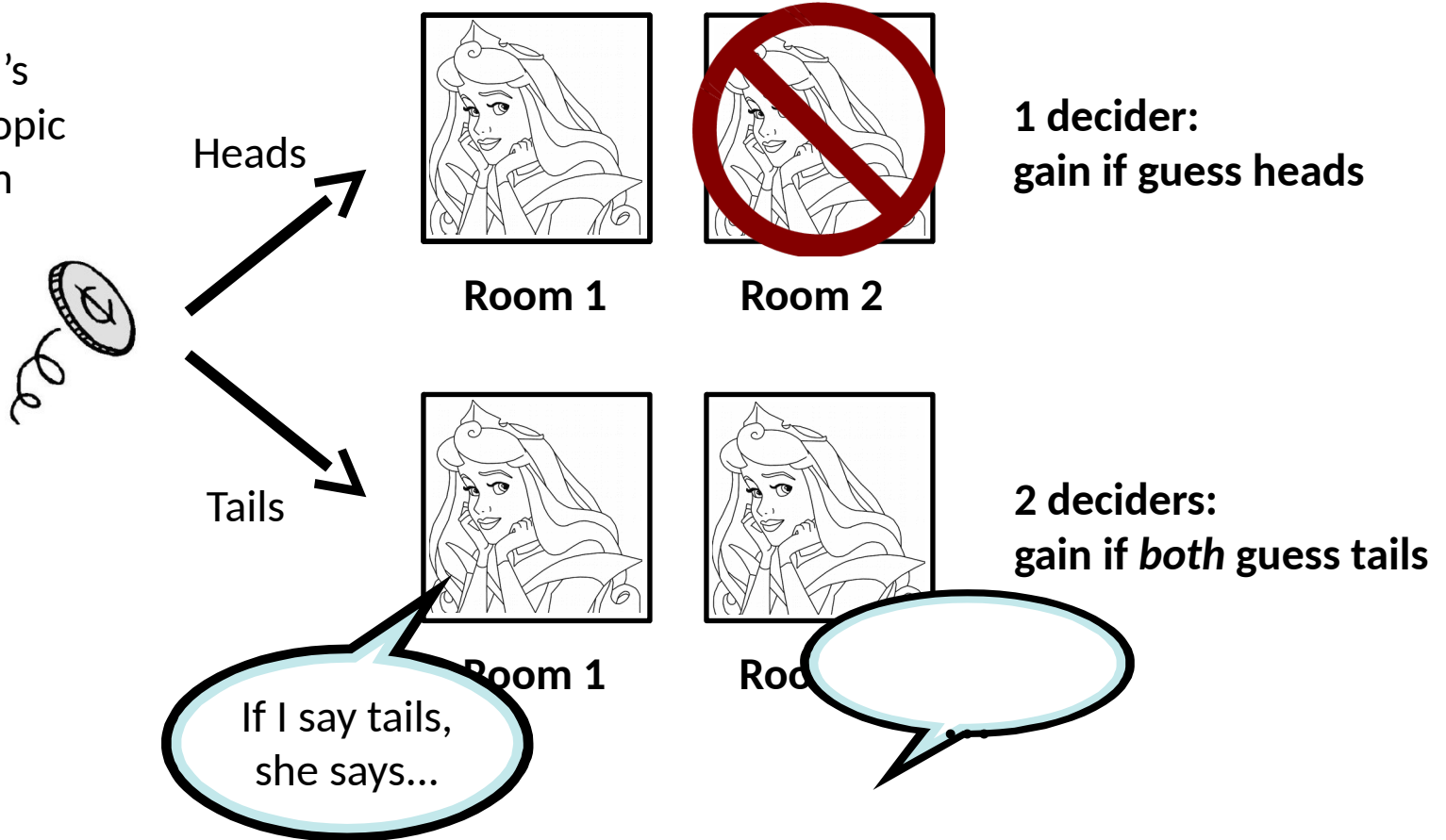
Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



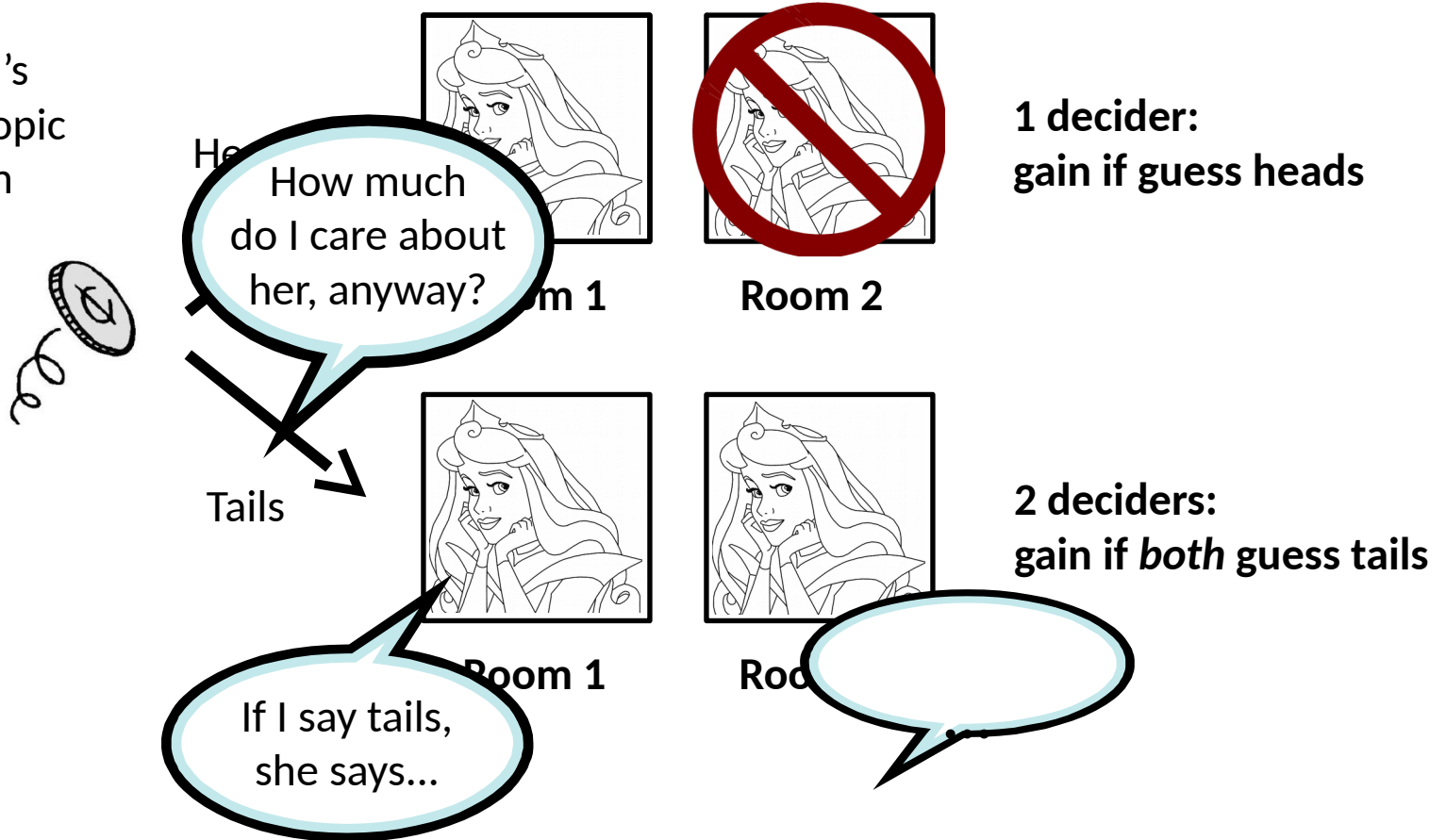
Evidential Decision Theory



Causal Decision Theory

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



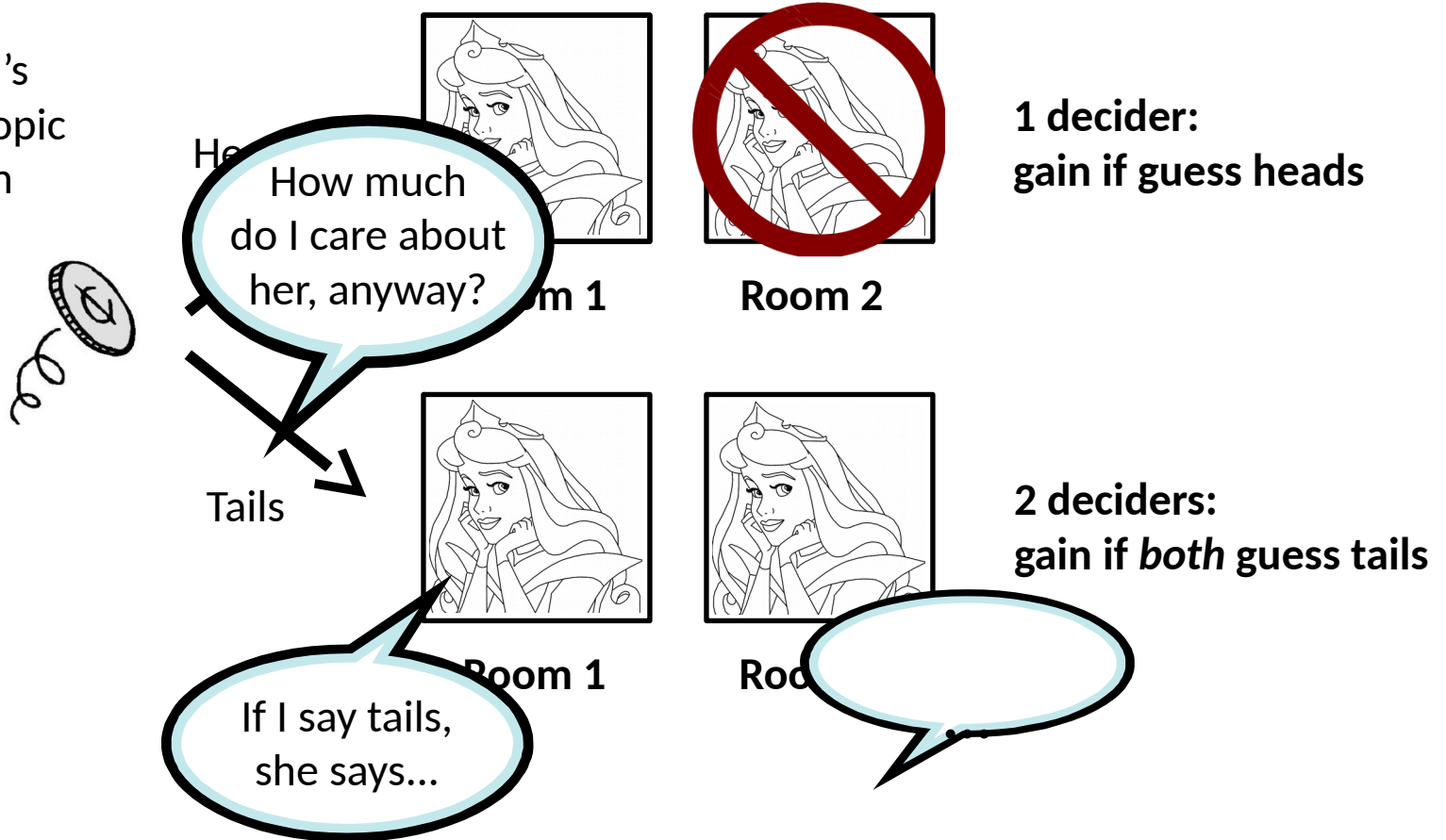
Evidential Decision Theory



Causal Decision Theory

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Evidential Decision Theory



Causal Decision Theory

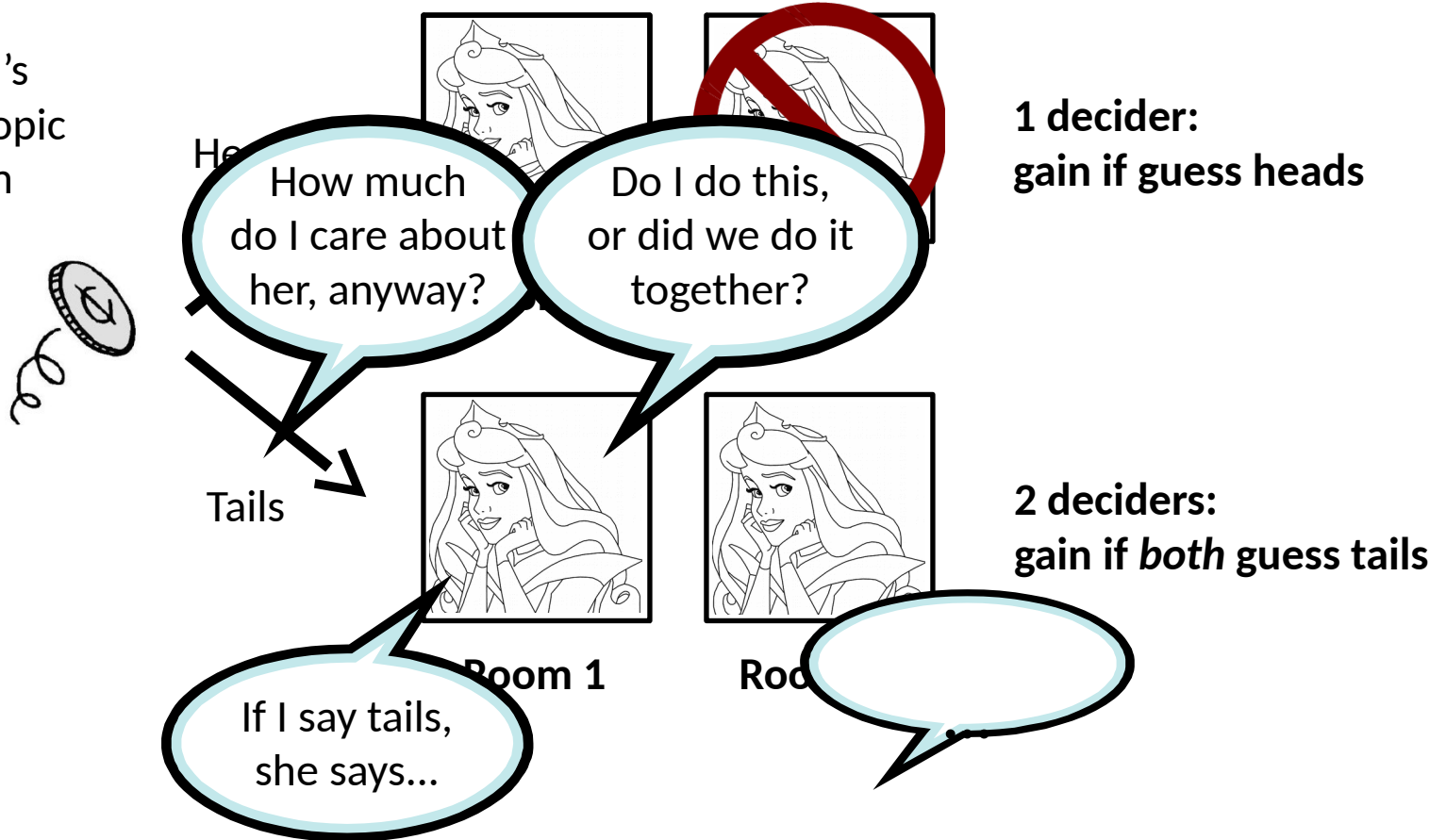
Altruistic



Selfish (precommit?)

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Evidential Decision Theory



Causal Decision Theory

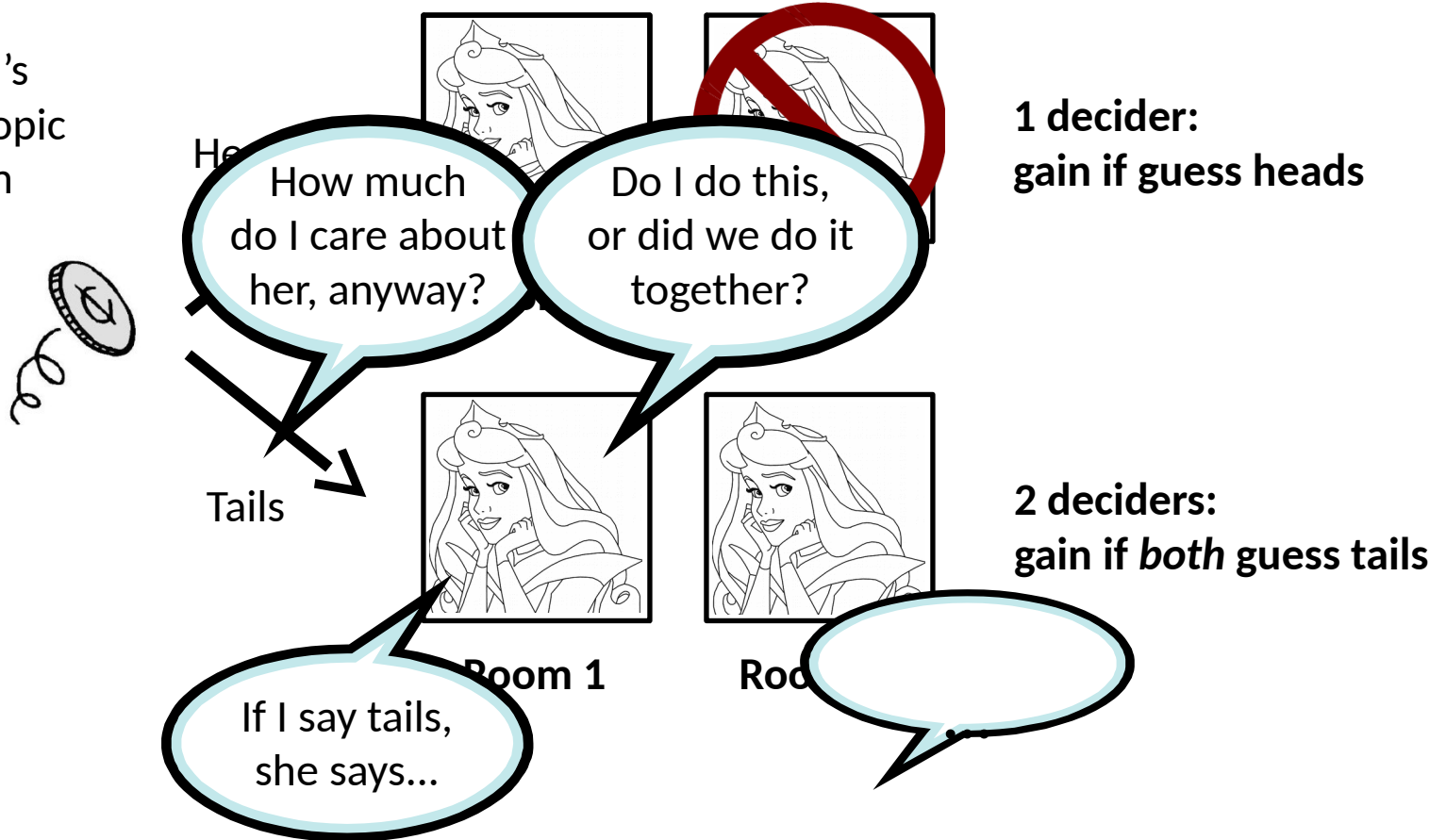
Altruistic



Selfish (precommit?)

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Evidential Decision Theory



Causal Decision Theory

Altruistic



Selfish (precommit?)

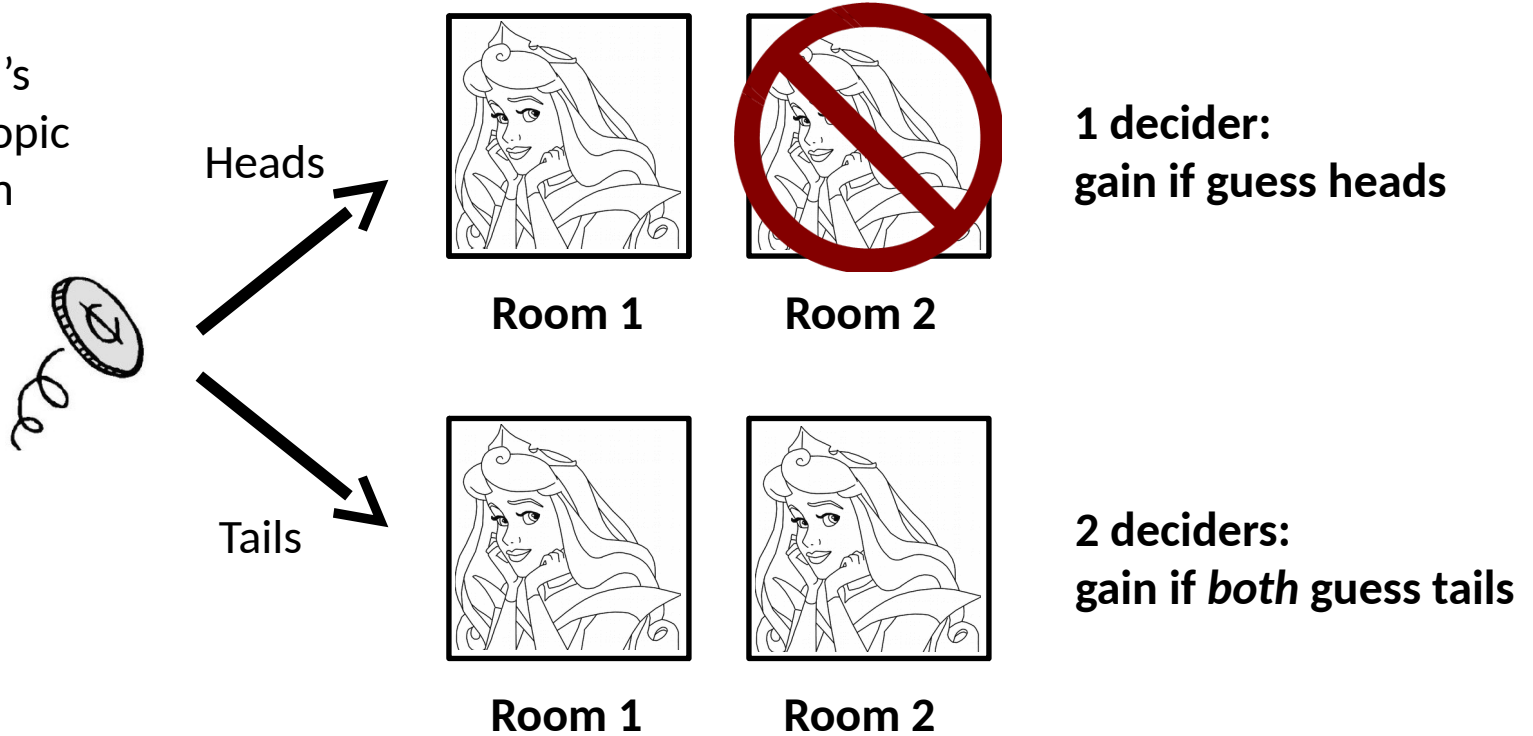
Total responsibility



Partial responsibility

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Evidential Decision Theory



Causal Decision Theory

Altruistic



Selfish (precommit?)

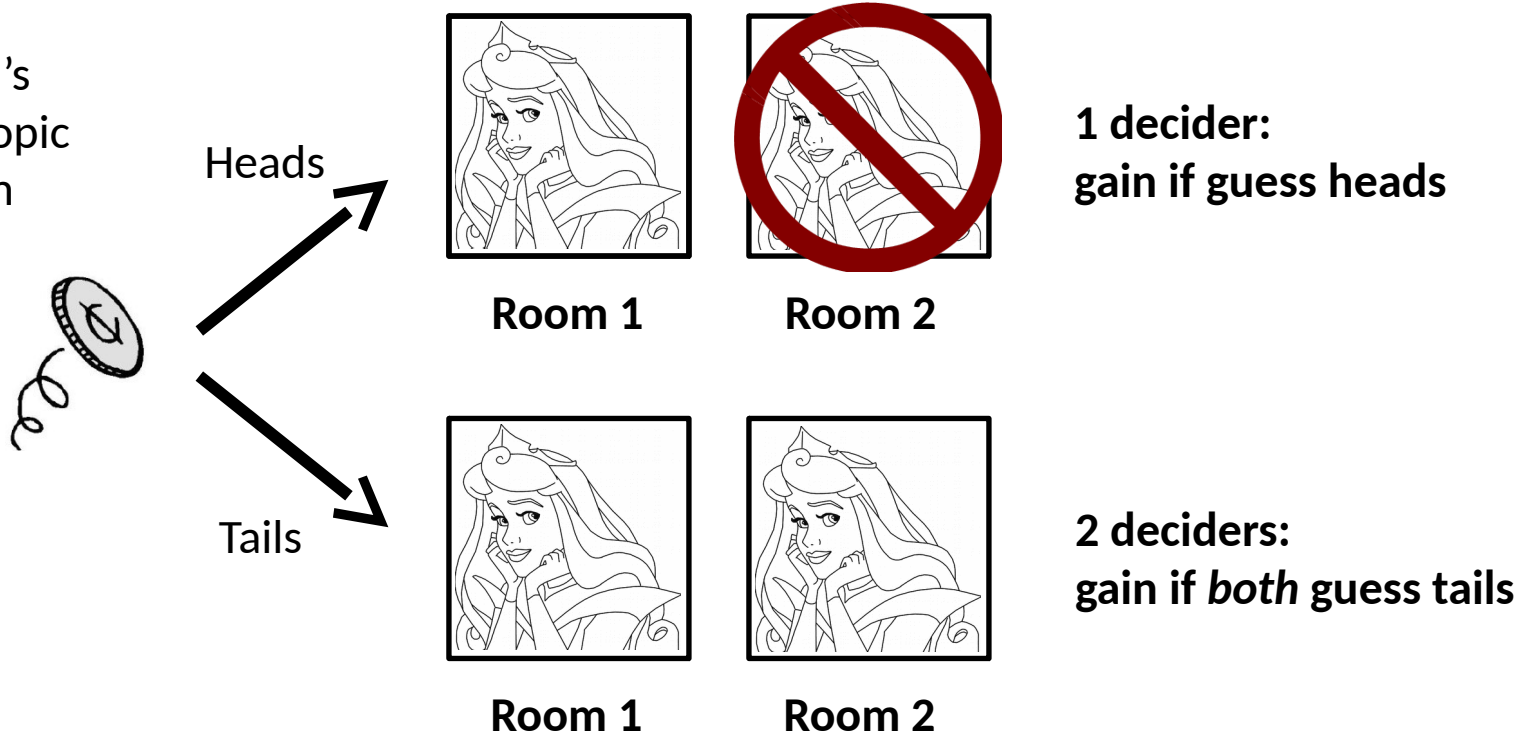
Total responsibility



Partial responsibility

Is anthropics the problem?

Psy-Kosh's
non-anthropic
problem



Evidential Decision Theory



Causal Decision Theory

Altruistic



Selfish (precommit?)

Total responsibility



Partial responsibility

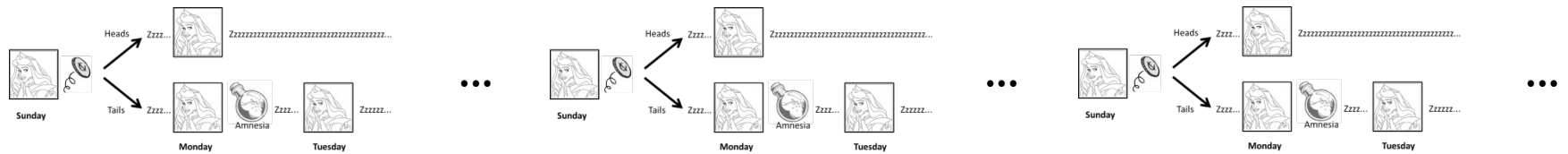
SIA



SSA

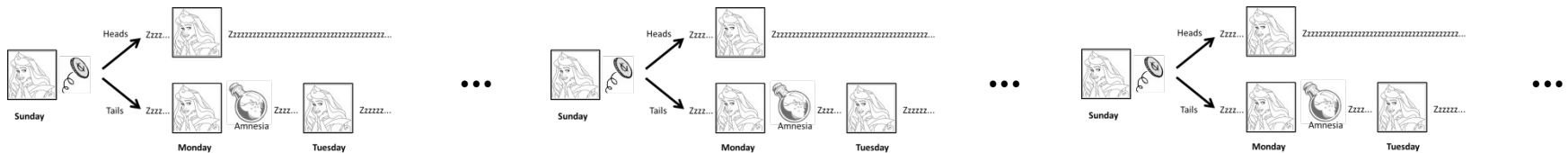
Anthropic probabilities don't really make sense

Frequentism:



Anthropic probabilities don't really make sense

Frequentism:



How many times were you right (SIA)?

VS

How many experiments were you right in (SSA)?

Anthropic probabilities don't really make sense

Bayesianism:

?

?

??

?

?

Anthropic probabilities don't really make sense

Bayesianism:

? ? ? ?? ? ?

Uncertain about the world with you in it (SSA)?

vs

Uncertain about you in the world (SIA)?

Anthropic probabilities don't really make sense

Subjective credences and expectations:



These were forged by evolution in non-anthropocentric situations.

The morals of the talk



Sleeping Beauty ~~problem~~ underdefined –
need Beauty's **values**.

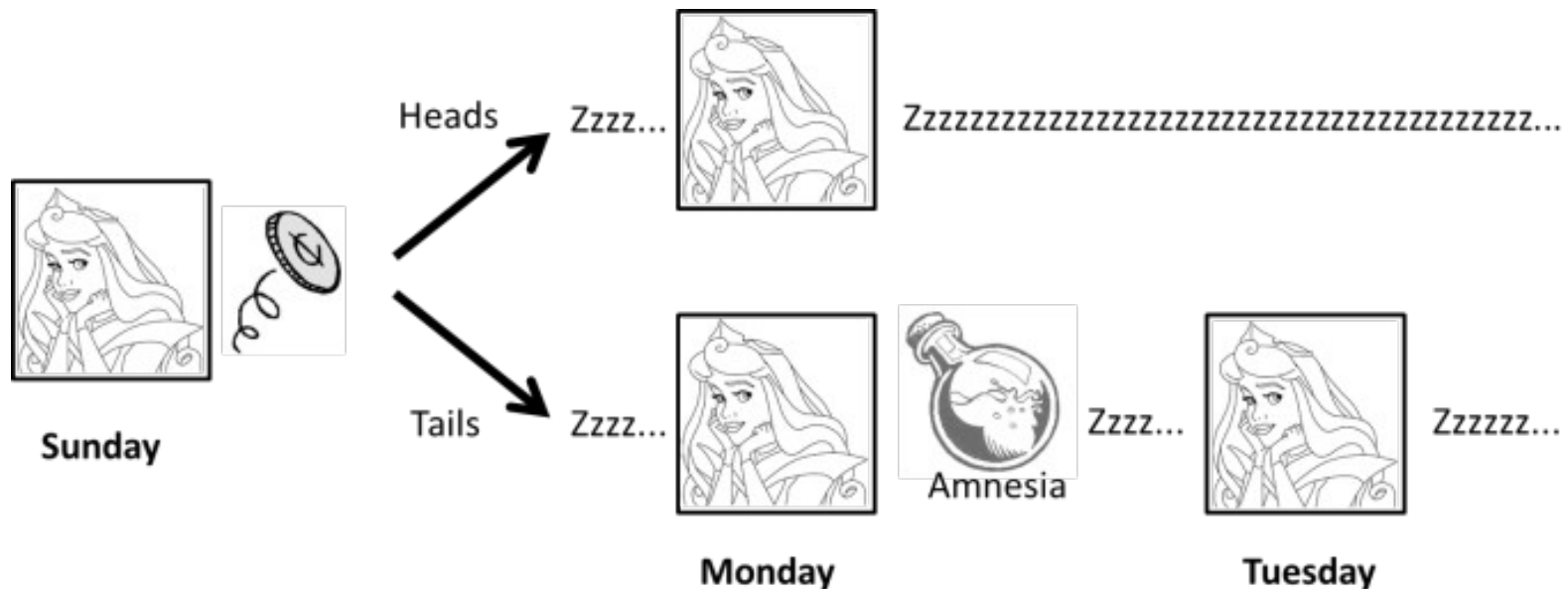
Even with
we can stil



obabilities,
t decision.

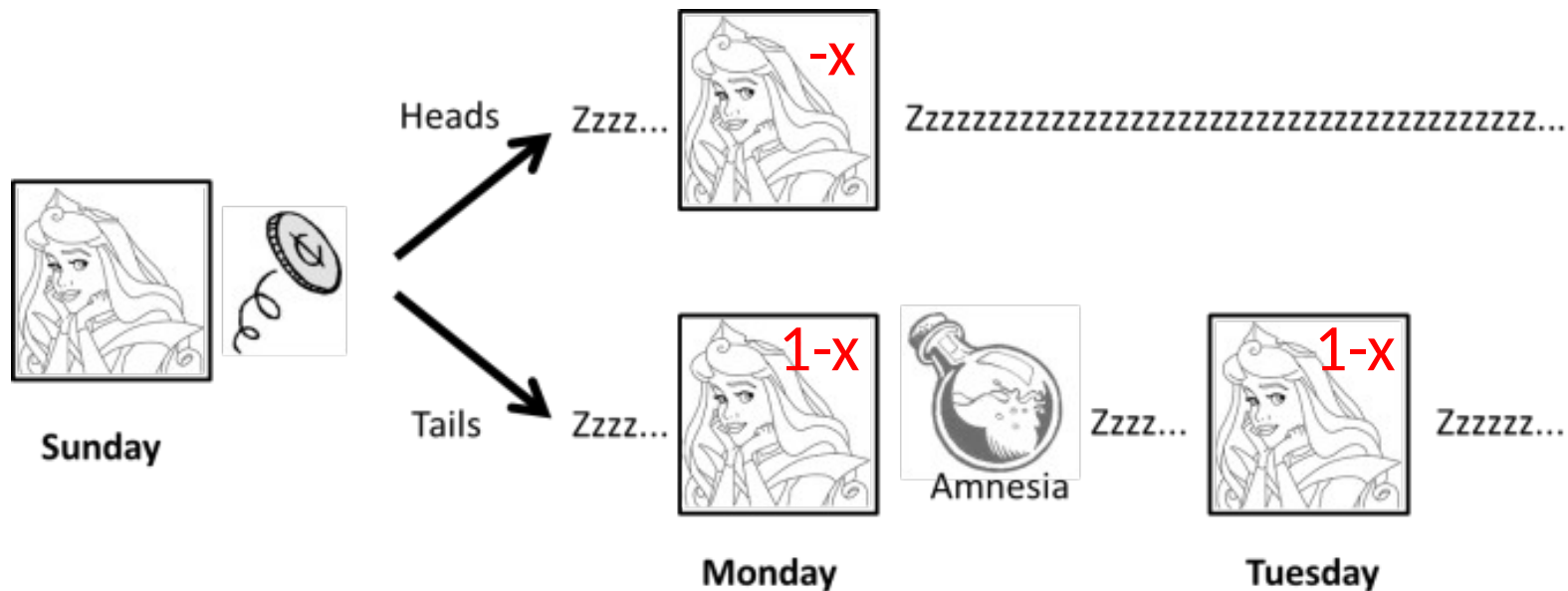
Decisions and values, not probabilities

Upon each awakening, Beauty is offered a coupon at $\pounds X$ that pays $\pounds 1$ if the coin was tails.



Decisions and values, not probabilities

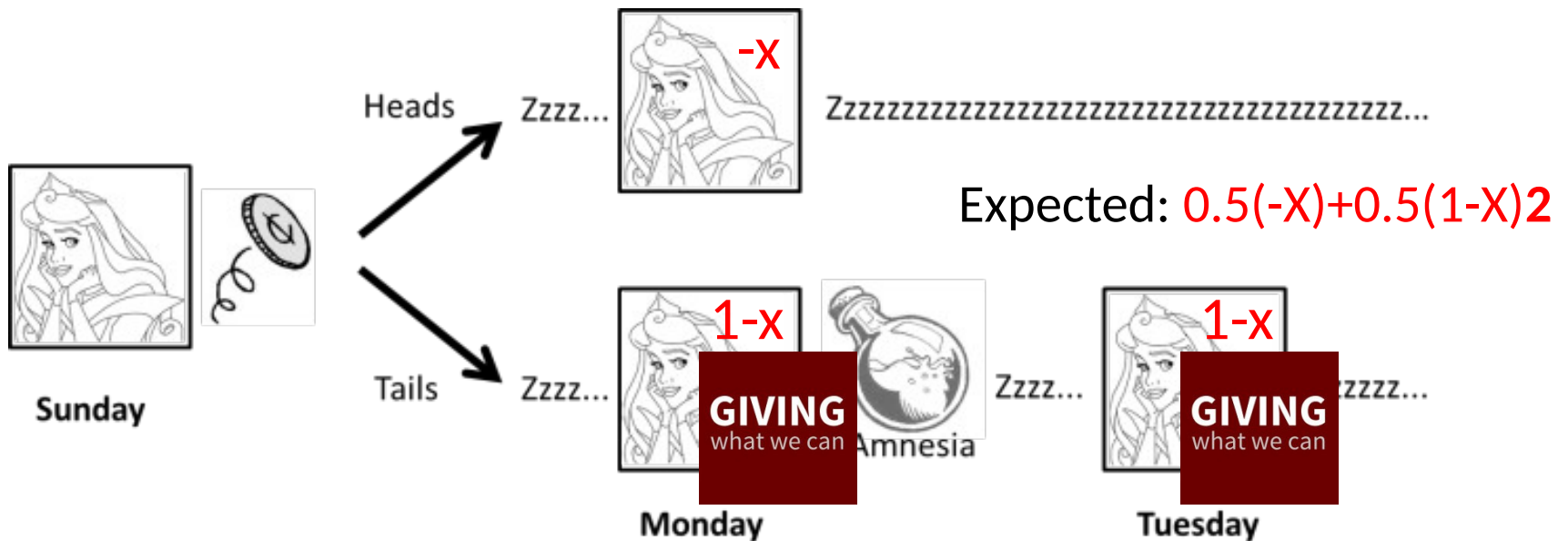
Upon each awakening, Beauty is offered a coupon at $\pounds X$ that pays $\pounds 1$ if the coin was tails.



Decisions and values, not probabilities

What would Sunday Beauty want?

If all cash goes towards a “cause”: $X < \text{£}2/3$



Decisions and values, not probabilities

What would Sunday Beauty want?

If all cash goes towards a “cause”: $X < \text{£}2/3$



Axiom 1: Precommitments are possible.

Decisions and values, not probabilities

What would Sunday Beauty want?

If cash is saved: $X < \frac{2}{3}$



Decisions and values, not probabilities

What would Sunday Beauty want?

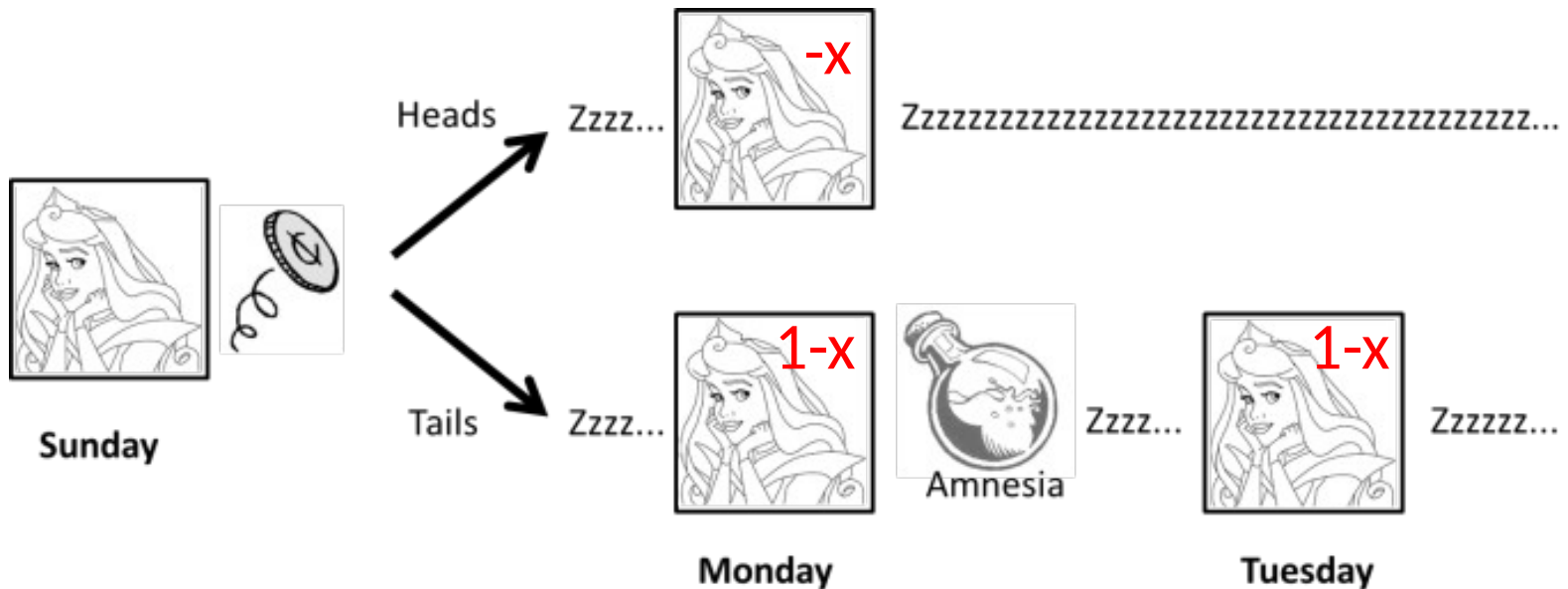
If cash buys chocolate: $X < \text{£}2/3$ or **£1/2**



Axiom 1: Precommitments are possible.

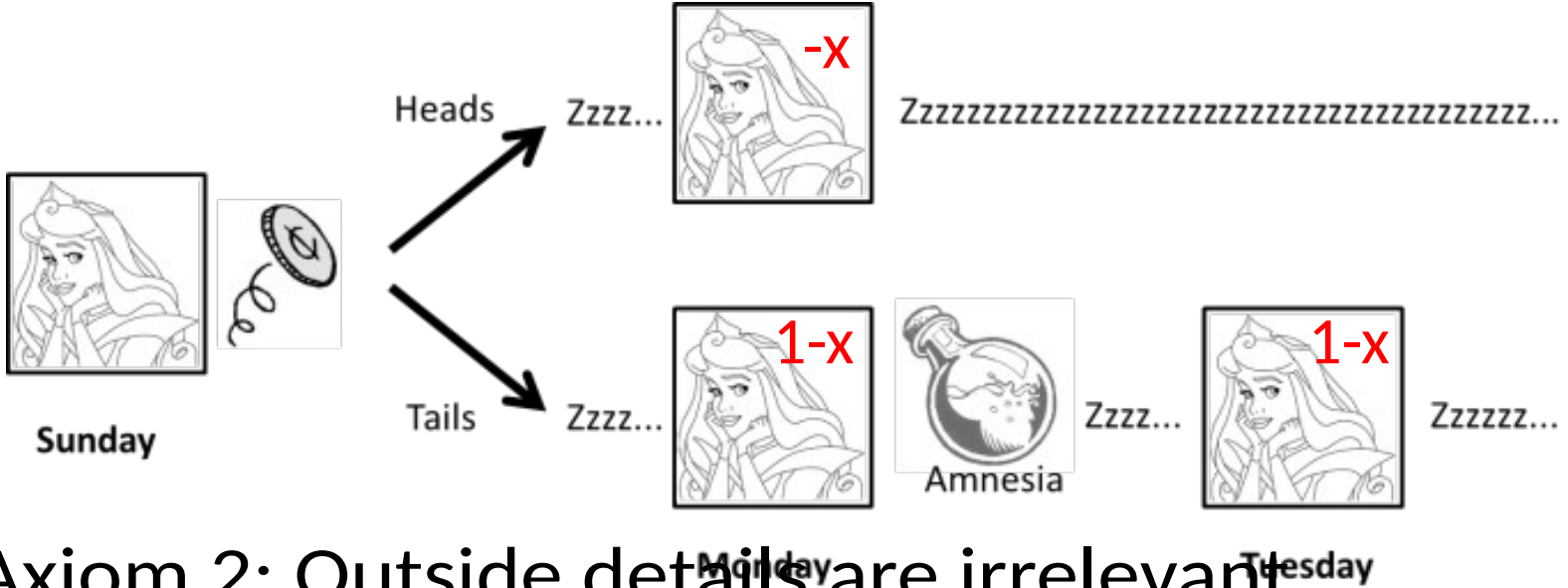
Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian	Copy-altruistic average utilitarian



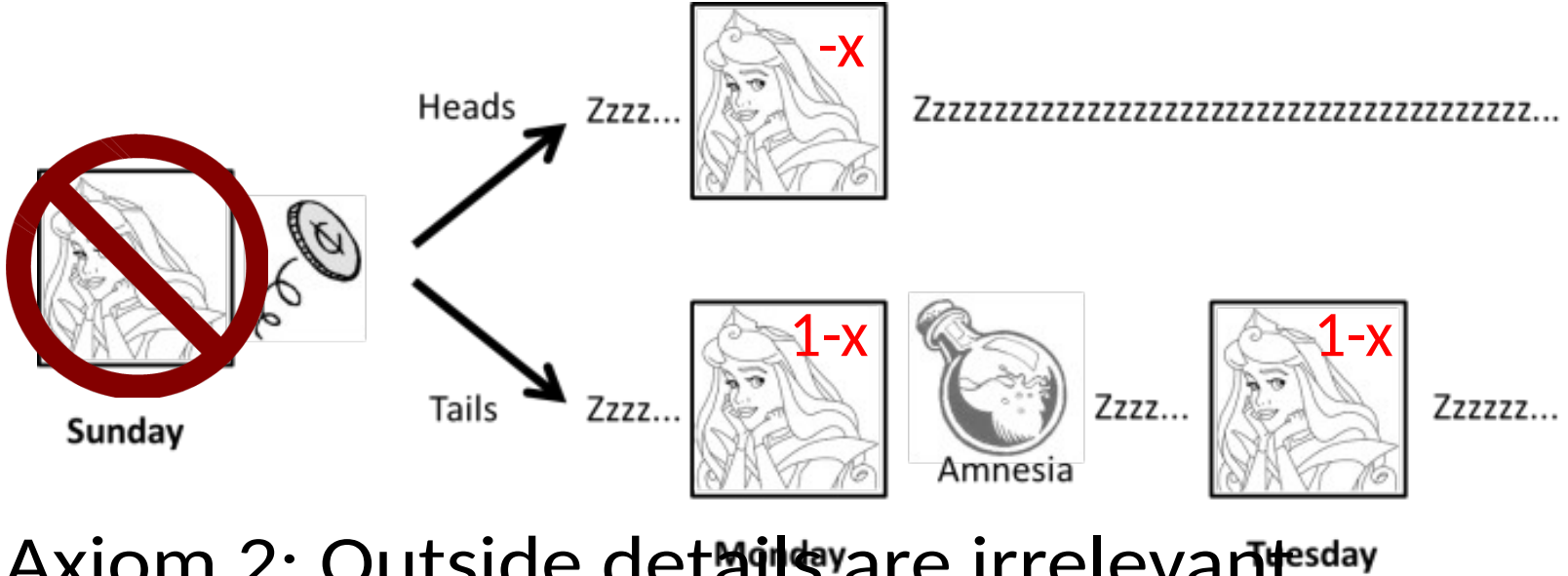
Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian \leftrightarrow	Copy-altruistic average utilitarian



Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian \leftrightarrow	Copy-altruistic average utilitarian



Axiom 2: Outside details are irrelevant.

Decisions and values, not probabilities

SIA-ish

SSA-ish

Non-indexical utility

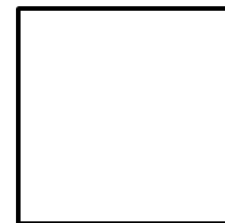
Copy-altruistic total utilitarian \leftrightarrow Copy-altruistic average utilitarian

I ♥ ME



Heads

Tails



Room 1

Room 2

Expected: $0.5(-X) + 0.5(1-X)1$

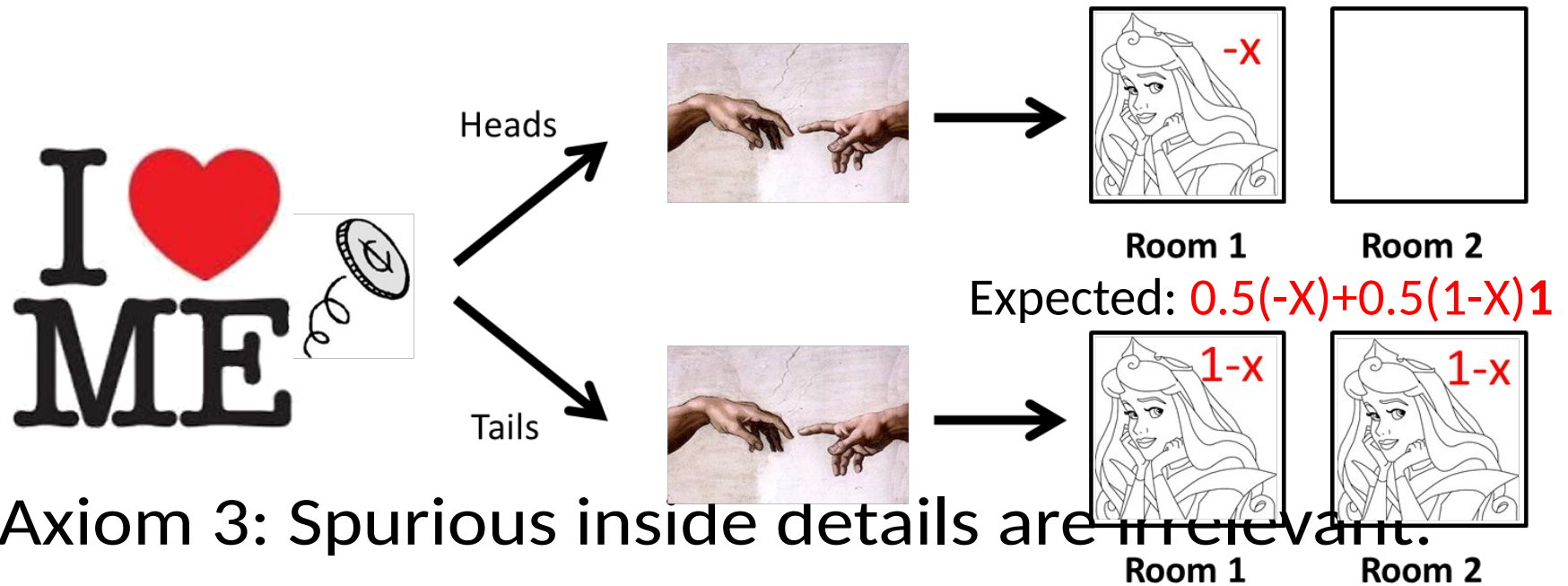


Room 1

Room 2

Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian	Copy-altruistic average utilitarian



Axiom 3: Spurious inside details are irrelevant.

Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian	Copy-altruistic average utilitarian
	Selfish (?)



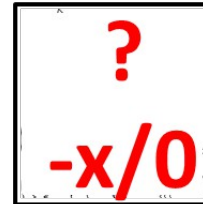
Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian	Copy-altruistic average utilitarian
	Selfish (?)

I ♥ ME



Heads



Room 1



Room 2

Expected: $0.5(-x)/2 + 0.5(1-x)1$

Tails



Room 1



Room 2

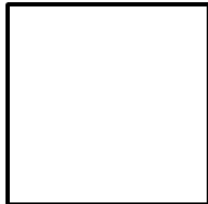
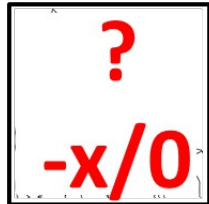
Decisions and values, not probabilities

SIA-ish	SSA-ish
Non-indexical utility	
Copy-altruistic total utilitarian	Copy-altruistic average utilitarian
Selfish (strict???)	Selfish (psychological approach)

I ♥ ME



Heads



Room 1

Room 2

Expected: $0.5(-x)/2 + 0.5(1-x)1$

Tails



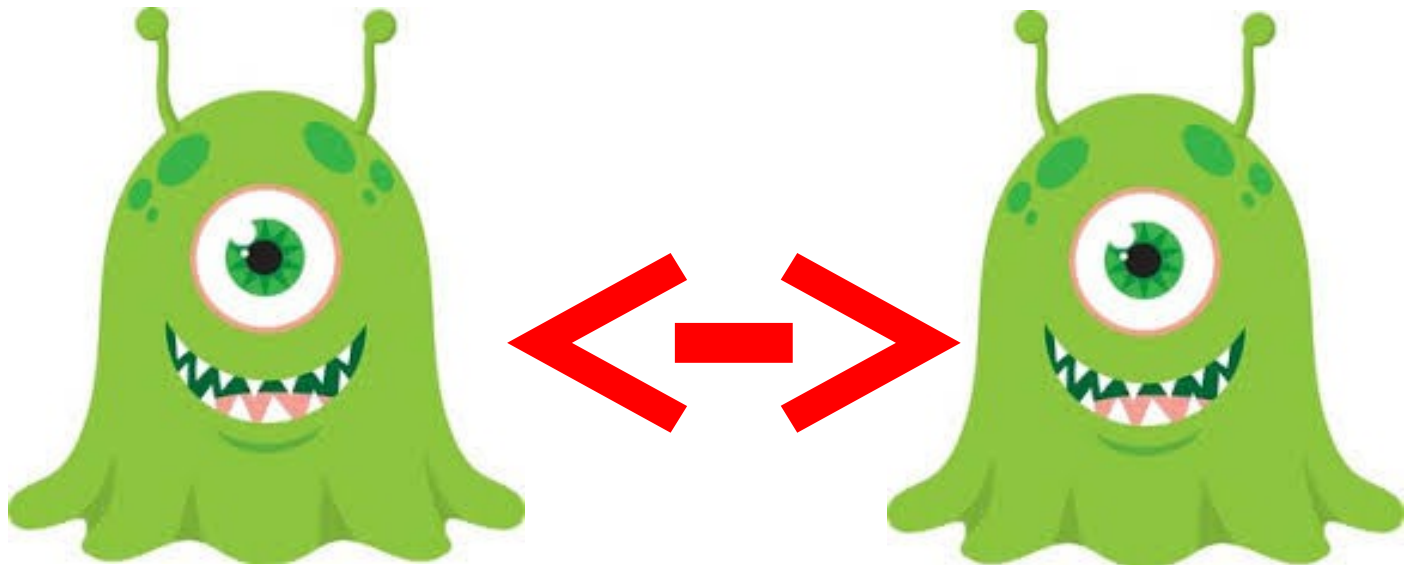
Room 1

Room 2

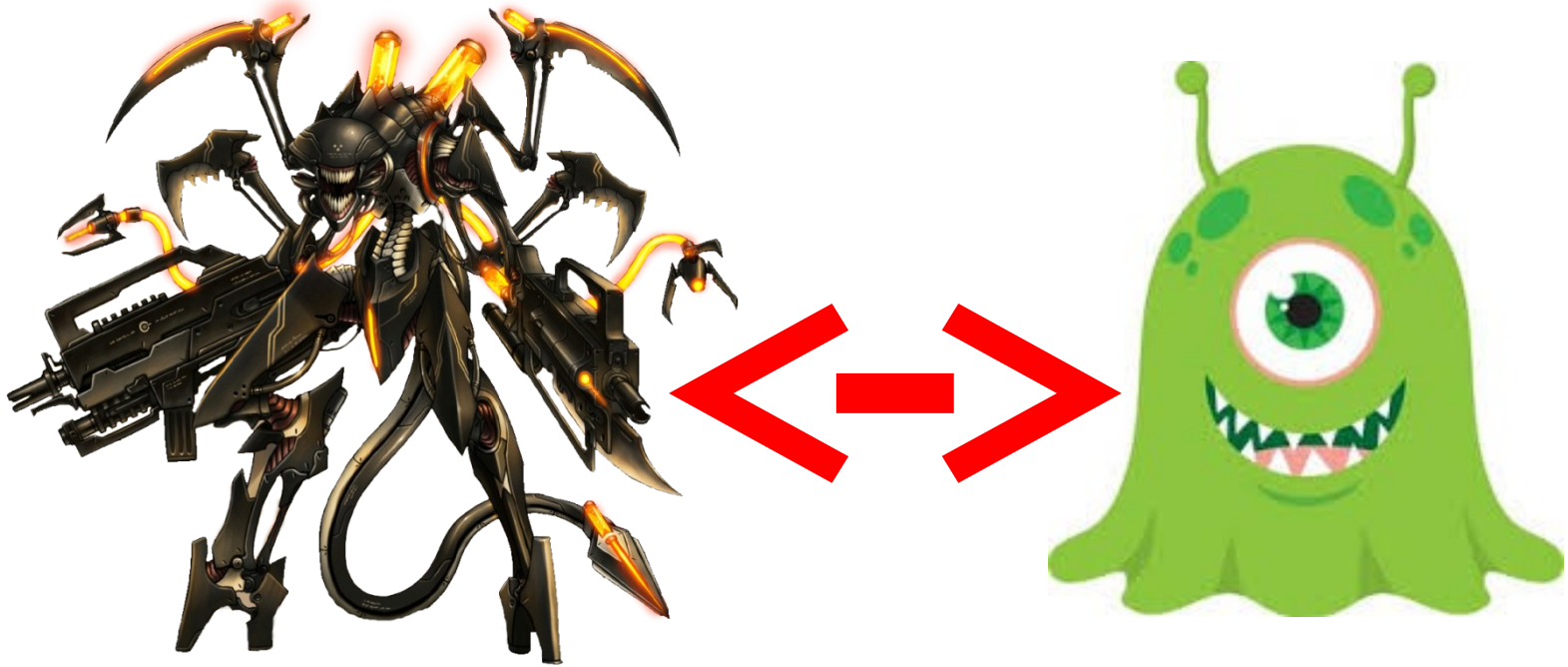
Axioms

- **Axiom 1: Precommitments are possible.**
(gives standard Sleeping Beauty for non-indexical preferences and altruists)
- **Axiom 2: Outside details are irrelevant.**
(gives incubator variant of Sleeping Beauty)
- **Axiom 3: Spurious inside details are irrelevant.**
(gives selfish preferences)

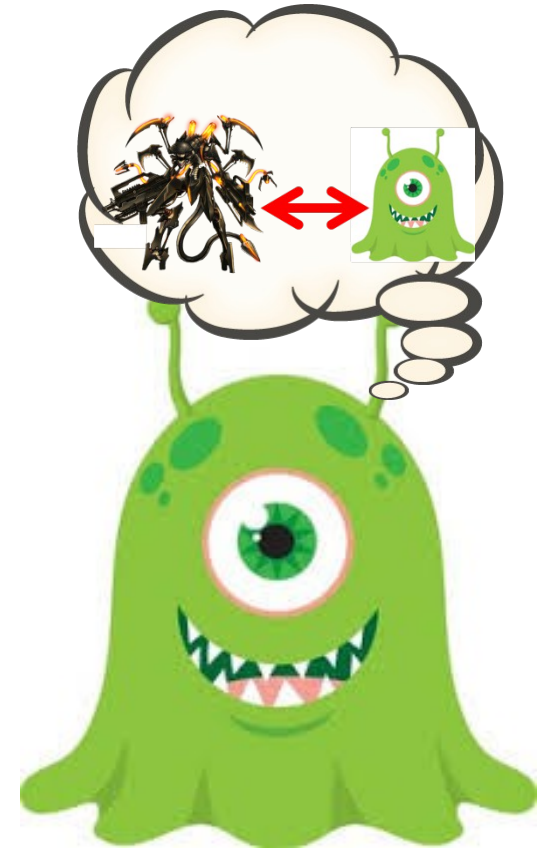
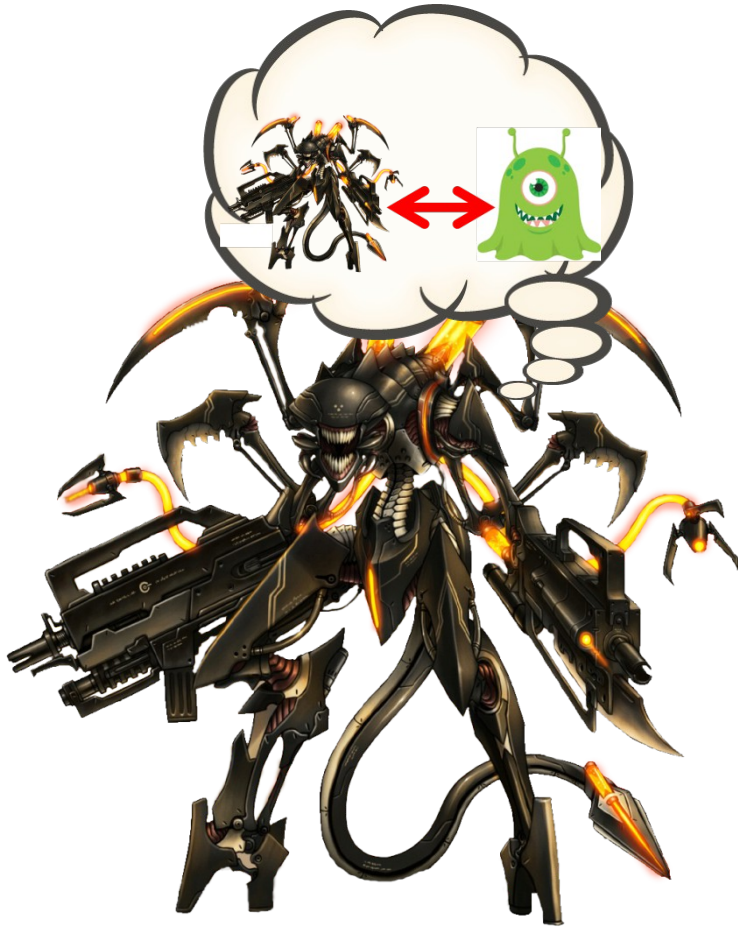
Linked decisions



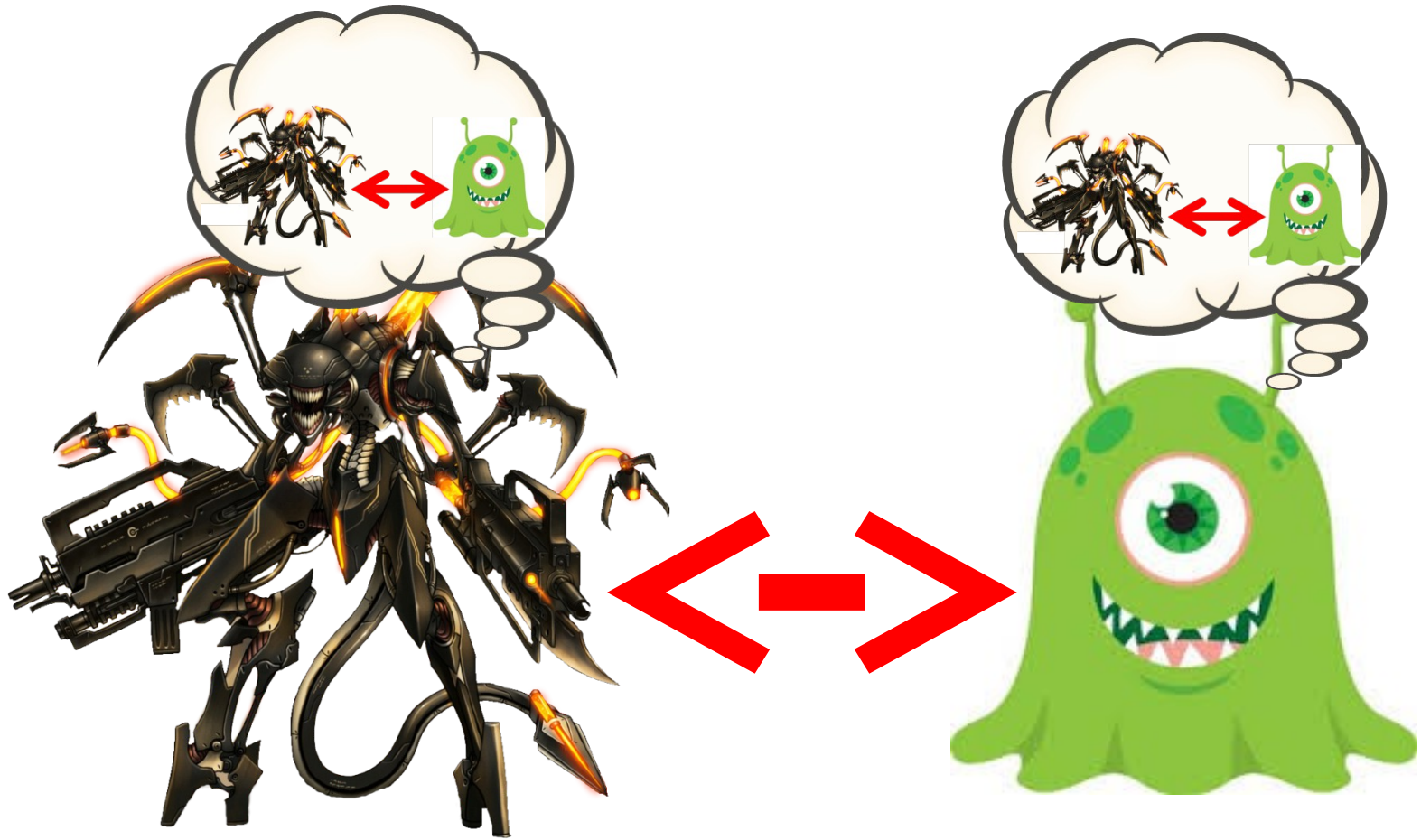
Linked decisions



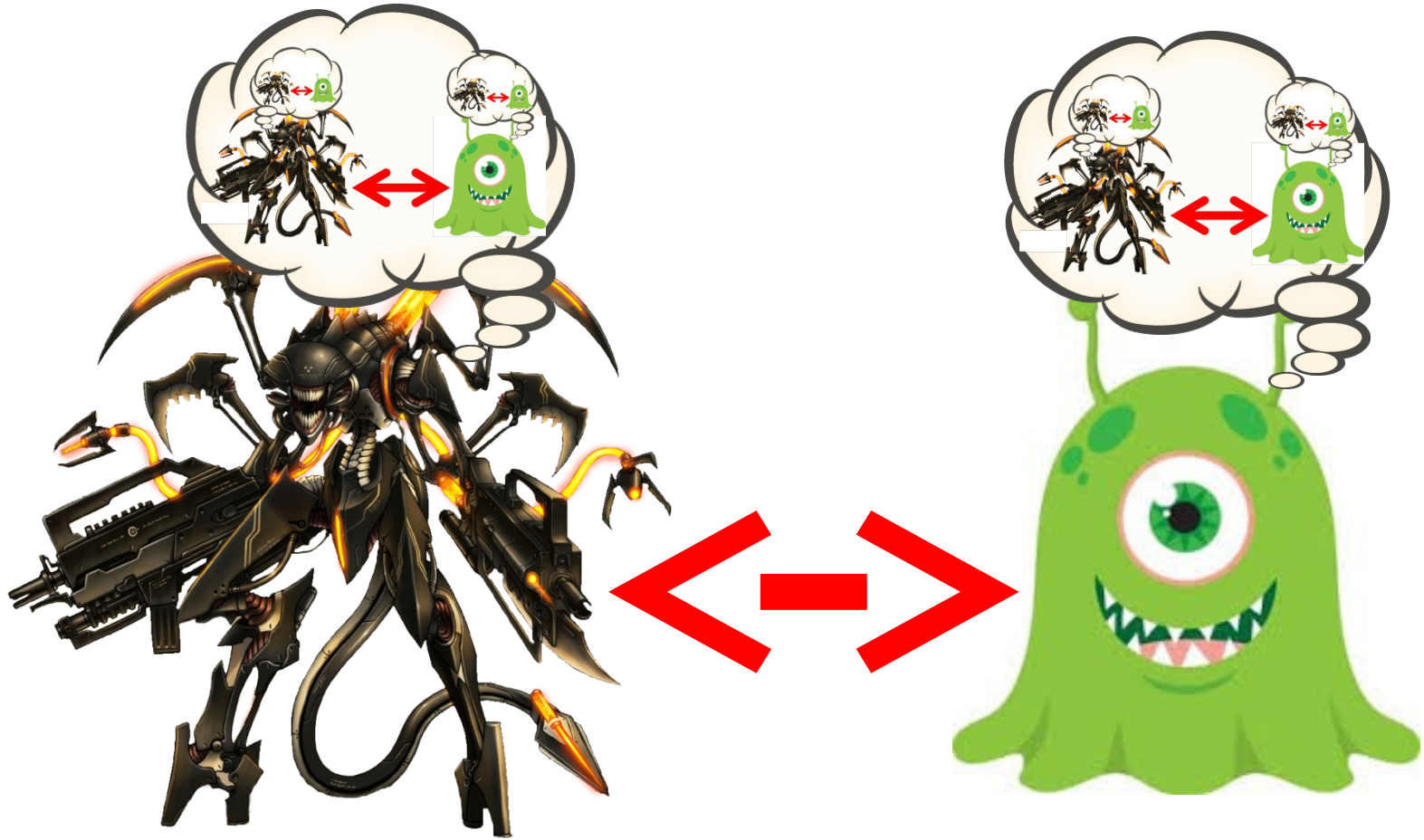
Linked decisions



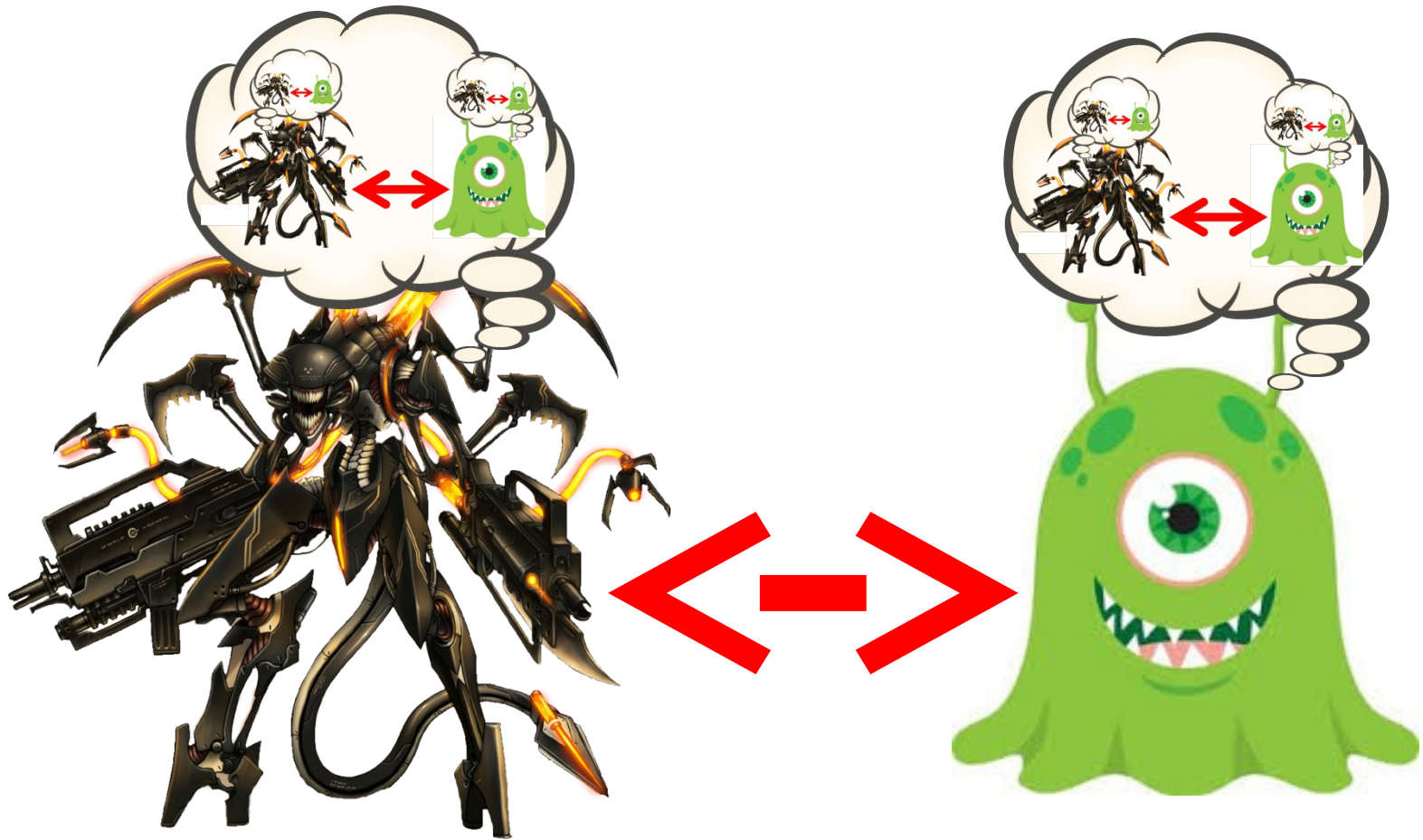
Linked decisions



Linked decisions



Linked decisions



Self-confirming linking

Anthropic Decision Theory

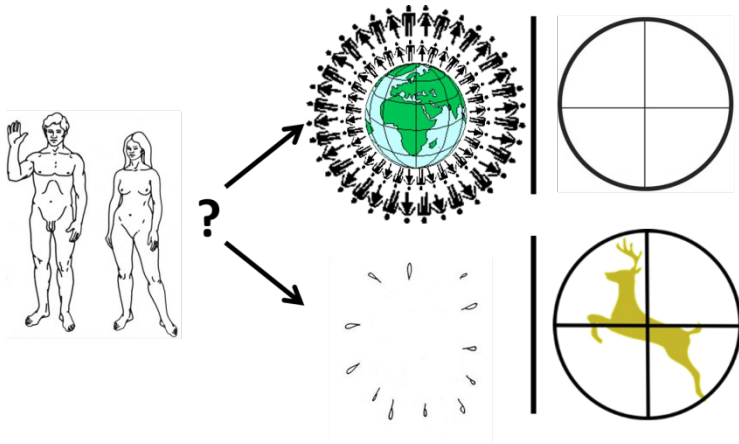
Anthropic decision theory (ADT):

An ADT agent searches for self-confirming linkings (for a given decision).

It then maximises expected utility, using standard (non-anthropropic) probabilities, acting as if it controlled all the agents' linked decisions.

Adam and Eve paradox

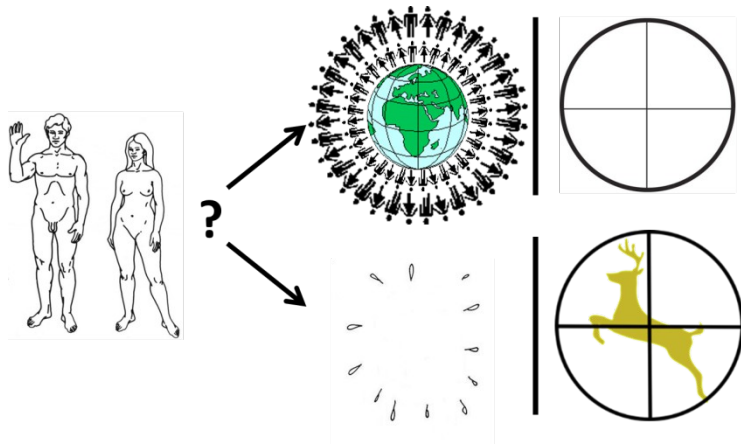
SSA: **Probability** of successful hunt is high.



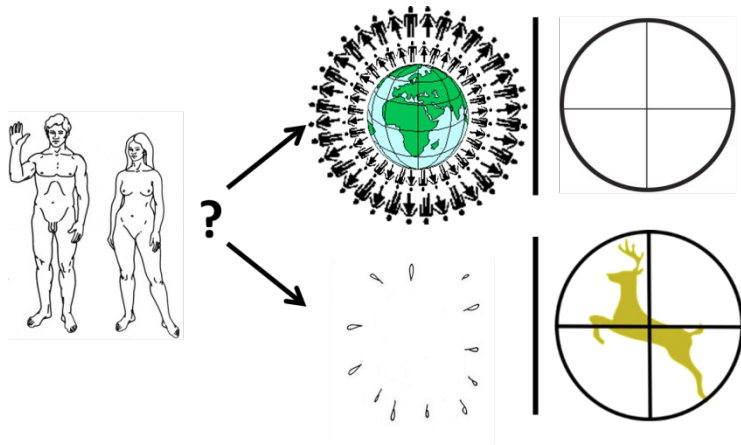
Adam and Eve paradox

SSA: **Probability** of successful hunt is high.

Average utilitarian: If average happiness is the same, **disutility** of failed hunt less if there are more people.



Adam and Eve paradox



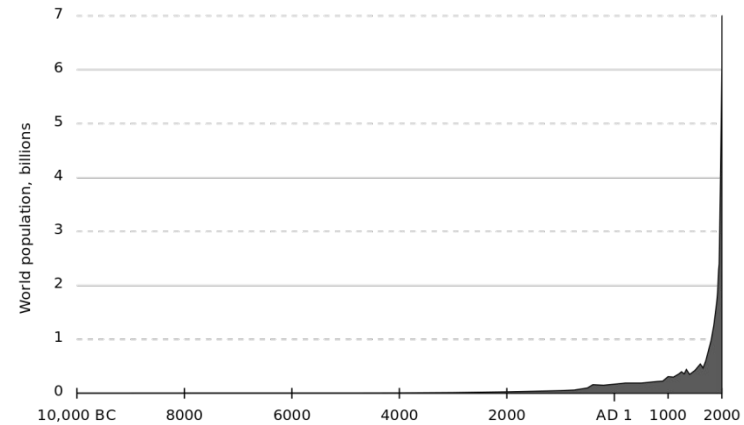
SSA: **Probability** of successful hunt is high.

Average utilitarian: If average happiness is the same, **disutility** of failed hunt less if there are more people.

Selfish + precommitment + ignorance:
In first world, Adam and Eve **suffer**, but I'm unlikely to be them. In second world, Adam and Eve **benefit**, and I'm certain to be one of them.

Doomsday argument

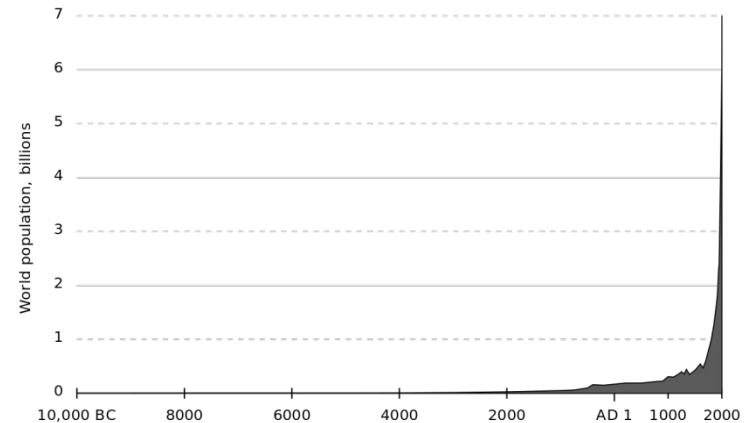
SSA: **Probability** of doom is high. No future generations.



Doomsday argument

SSA: **Probability** of doom is high. No future generations.

What kind of betting behaviour are we looking for? Prefers to consume a windfall now rather than save future generations.



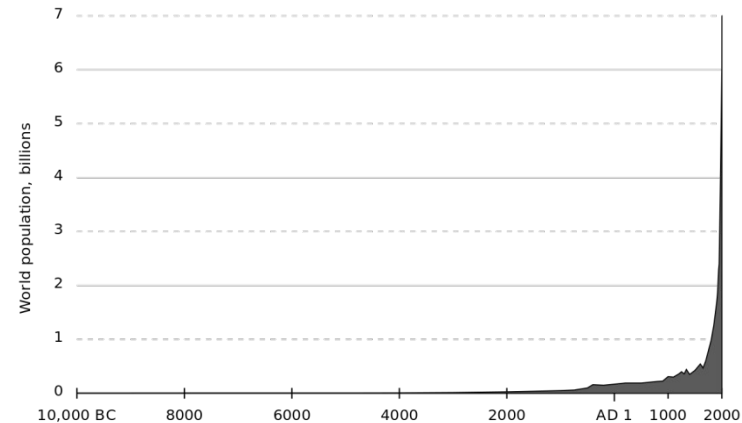
Doomsday argument

SSA: **Probability** of doom is high. No future generations.

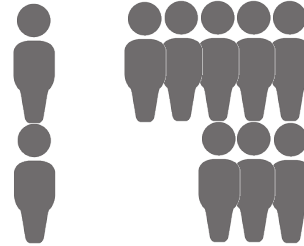
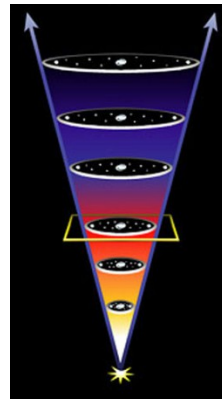
What kind of betting behaviour are we looking for? Prefers to consume a windfall now rather than save future generations.

Average utilitarian: if future generations are of similar average happiness, then better consume windfall ω today than let Ω more people exist.

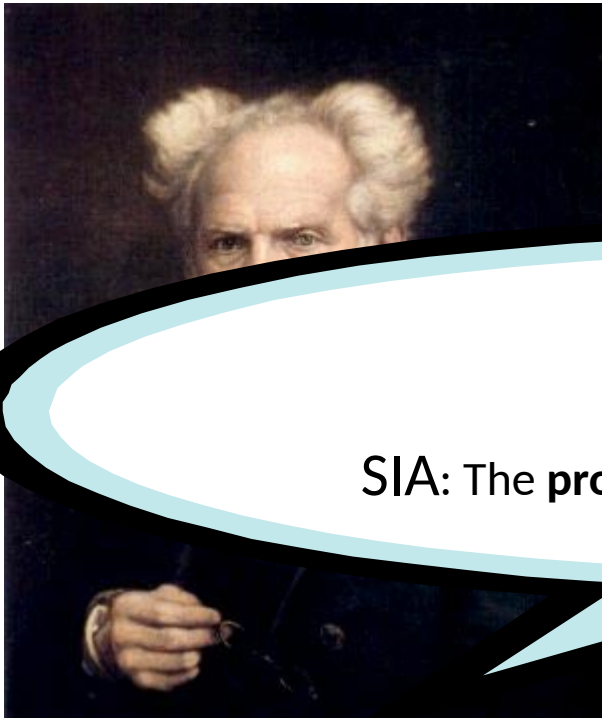
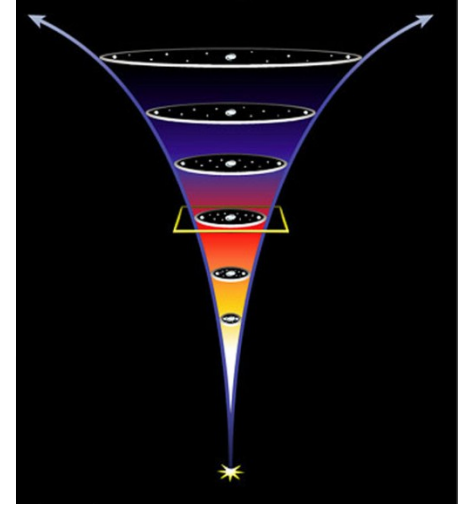
$$\omega/\Omega \approx 0$$



Presumptuous philosopher

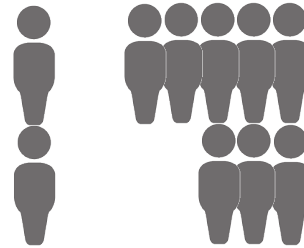
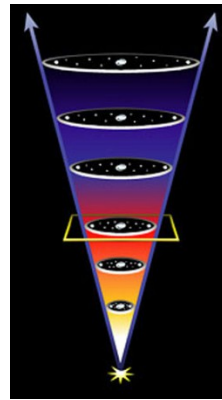


$$\Lambda = ?$$

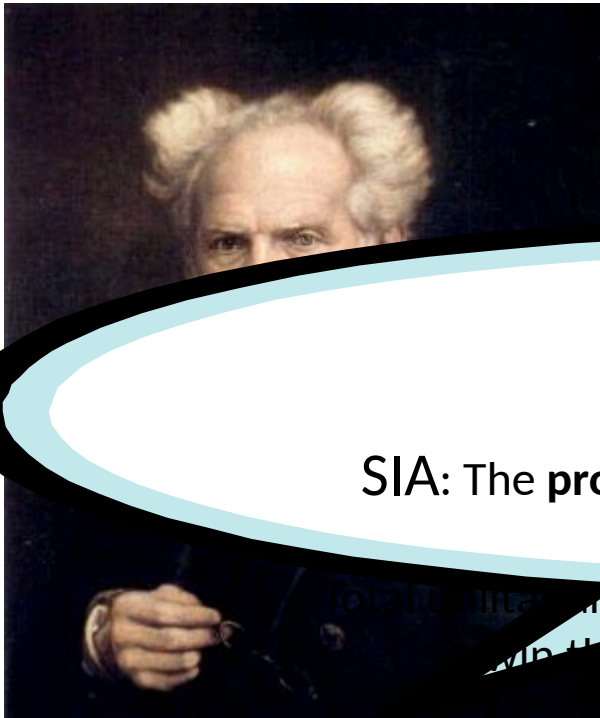
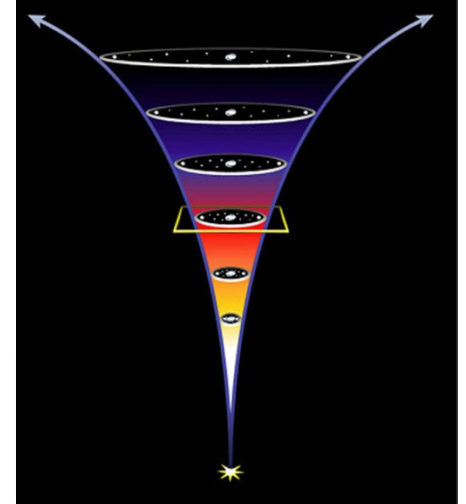


SIA: The **probability** of the large universe is large.

Presumptuous philosopher



$$\Lambda = ?$$



SIA: The **probability** of the large universe is large.

Presumptuous philosopher: in a large universe, many philosophers
will win their bets, and I care about them.