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

Sunday

Heads
Zzzz...

Tails
Zzzz...

Monday

Zzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzz...

Amnesia

Zzzz...

Tuesday

Zzzzzz...
Upon awakening, what is the **probability** of Heads? Of Monday?
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)
SIA prefers large universes (present, not future)
Presumptuous philosopher

SIA prefers large universes (present, not future)
Presumptuous philosopher

SIA prefers large universes (present, not future)
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 the trillion times bigger universe.
Presumptuous philosopher

SIA prefers
large universes
(present, not future)

I know!!!

I’ll bet you at odds
of a trillion to one on
the trillion times bigger
universe

You can’t produce
enough evidence to
change my mind

Λ=?
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

Heads

Room 1

Room 2

Tails

Room 1

Room 2
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

Heads

Room 1

Tails

Room 1

Room 2

Room 2

1 decider: gain if guess heads
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

- Heads
  - Room 1
  - Room 2

- Tails
  - Room 1
  - Room 2

1 decider: gain if guess heads

2 deciders: gain if both guess tails
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

If I say tails, she says...

1 decider: gain if guess heads

2 deciders: gain if *both* guess tails
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

1 decider: gain if guess heads

2 deciders: gain if both guess tails

If I say tails, she says...
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

Heads

Room 1

Room 2

Tails

If I say tails, she says...

1 decider: gain if guess heads

2 deciders: gain if both guess tails

Evidential Decision Theory

Causal Decision Theory
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

How much do I care about her, anyway?

If I say tails, she says...

1 decider: gain if guess heads

2 deciders: gain if both guess tails

Evidential Decision Theory

Causal Decision Theory
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

1 decider: gain if guess heads

2 deciders: gain if both guess tails

If I say tails, she says...

How much do I care about her, anyway?

Evidential Decision Theory

Altruistic

Causal Decision Theory

Selfish (precommit?)
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

1 decider: gain if guess heads

2 deciders: gain if both guess tails

How much do I care about her, anyway?
Do I do this, or did we do it together?

If I say tails, she says...

Evidential Decision Theory
Causal Decision Theory

Altruistic
Selfish (precommit?)
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

1 decider: gain if guess heads

2 deciders: gain if both guess tails

If I say tails, she says...

How much do I care about her, anyway?

Do I do this, or did we do it together?

Room 1

Room 2

Evidential Decision Theory

Altruistic

Causal Decision Theory

Selfish (precommit?)

Total responsibility

Partial responsibility
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

1 decider: gain if guess heads

2 deciders: gain if both guess tails

Evidential Decision Theory

Altruistic

Total responsibility

Causal Decision Theory

Selfish (precommit?)

Partial responsibility
Is anthropics the problem?

Psy-Kosh’s non-anthropic problem

Heads

1 decider: gain if guess heads

Room 1

Tails

2 deciders: gain if both guess tails

Room 1

Room 2

Evidential Decision Theory

Altruistic

Total responsibility

Causal Decision Theory

Selfish (precommit?)

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-anthropic situations.
The morals of the talk

Sleeping Beauty problem is underdefined – need Beauty’s values.

Even without anthropic probabilities, we can still make the right decision.
Decisions and values, not probabilities

Upon each awakening, Beauty is offered a coupon at £X that pays £1 if the coin was tails.
Decisions and values, not probabilities

Upon each awakening, Beauty is offered a coupon at £X that pays £1 if the coin was tails.
Decisions and values, not probabilities

What would Sunday Beauty want?
Decisions and values, not probabilities

What would Sunday Beauty want?
If all cash goes towards a “cause”: $X < \frac{2}{3}$

Expected: $0.5(-X) + 0.5(1-X)^2$
Decisions and values, not probabilities

What would Sunday Beauty want?
If all cash goes towards a “cause”: $X < \frac{\£2}{3}$

Axiom 1: Precommitments are possible.
Decisions and values, not probabilities

What would Sunday Beauty want?
If cash is saved: $X < \£2/3$

Axiom 1: Precommitments are possible.

Expected: $0.5(-X)+0.5(1-X^2)$
Decisions and values, not probabilities

What would Sunday Beauty want?
If cash buys chocolate: \( X < \frac{\£}{2/3} \) or \( \frac{\£}{1/2} \)

Axiom 1: Precommitments are possible.
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td></td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
</tbody>
</table>

### Copy-altruistic total utilitarian

- Heads: Sunday
- Tails: Monday

- Amnesia: 1-x
- Zzzz...

### Copy-altruistic average utilitarian

- Zzzz...
- 1-x
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td></td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
</tbody>
</table>

Axiom 2: Outside details are irrelevant.
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td></td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
</tbody>
</table>

Axiom 2: Outside details are irrelevant.
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td></td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
</tbody>
</table>

Outside details are irrelevant.

If cash buys chocolate: $X < \frac{2}{3}$ or $\frac{1}{2}$
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td>Copy-altruistic total utilitarian ↔ Copy-altruistic average utilitarian</td>
</tr>
</tbody>
</table>

Expected: \(0.5(-X) + 0.5(1-X)\)
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td>Copy-altruistic total utilitarian $\iff$ Copy-altruistic average utilitarian</td>
</tr>
</tbody>
</table>

Axiom 3: Spurious inside details are irrelevant.
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td></td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
<tr>
<td></td>
<td>Selfish (?)</td>
</tr>
</tbody>
</table>
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Selfish (?)</td>
</tr>
</tbody>
</table>

Expected: \[0.5(-X)/2 + 0.5(1-X)1\]
Decisions and values, not probabilities

<table>
<thead>
<tr>
<th>SIA-ish</th>
<th>SSA-ish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-indexical utility</td>
<td></td>
</tr>
<tr>
<td>Copy-altruistic total utilitarian</td>
<td>Copy-altruistic average utilitarian</td>
</tr>
<tr>
<td>Selfish (strict???)</td>
<td>Selfish (psychological approach)</td>
</tr>
</tbody>
</table>

**Expected Calculation:**

\[
\text{Expected: } 0.5(-X)/2 + 0.5(1-X)1
\]
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-anthropic) 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 $\omega$ today than let $\Omega$ more people exist.

$$\frac{\omega}{\Omega} \approx 0$$
Presumptuous philosopher

SIA: The probability of the large universe is large.
Presumptuous philosopher

SIA: The probability of the large universe is large.

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