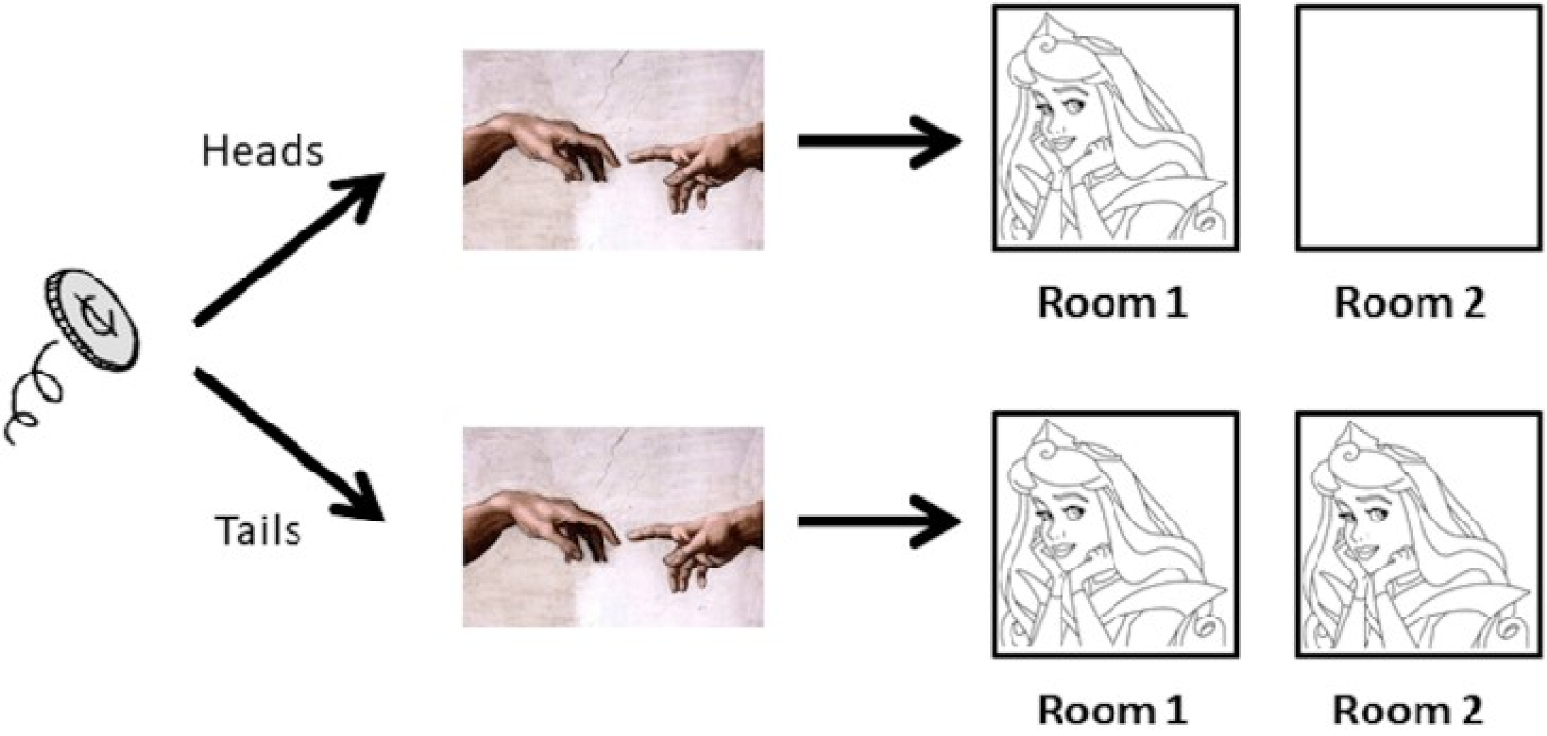

Heads
Tails
Room 1
Room 2
Room 1
Room 2

Anthropic Decision Theory

Anthropic decision theory I: Sleeping beauty and selflessness

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

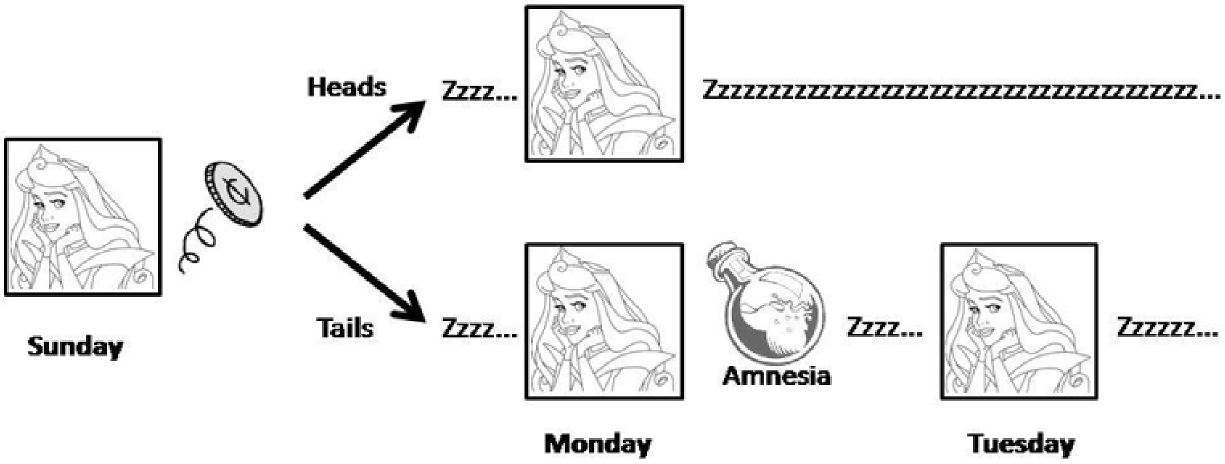
A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this and subsequent posts 1 2 3 4 5 6.

Many thanks to Nick Bostrom, Wei Dai, Anders Sandberg, Katja Grace, Carl Shulman, Toby Ord, Anna Salamon, Owen Cotton-barratt, and Eliezer Yudkowsky.

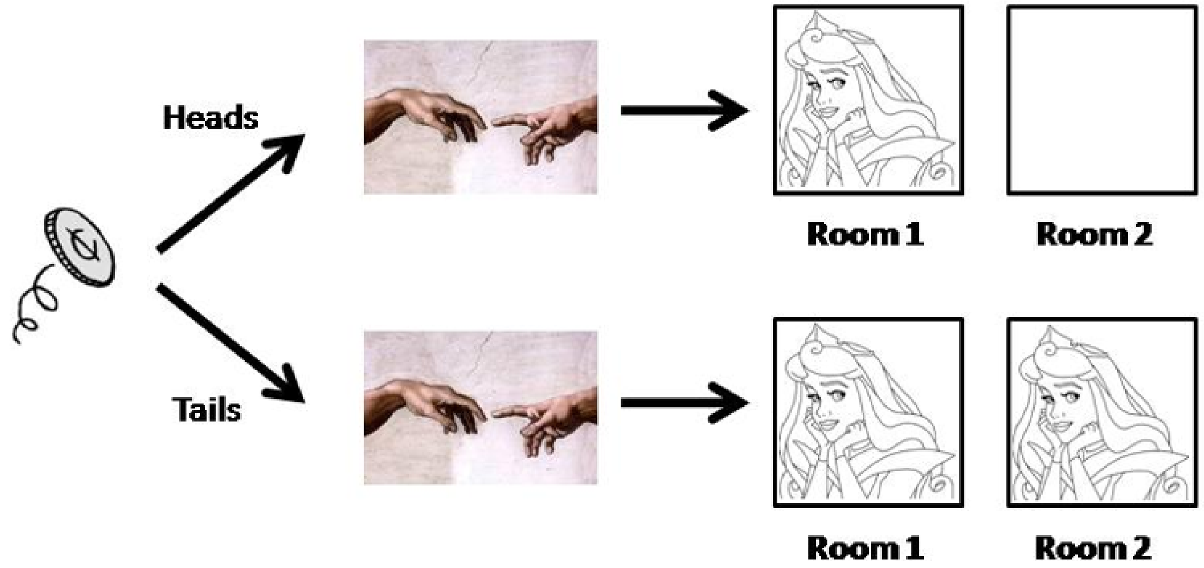
The Sleeping Beauty problem, and the incubator variant

The Sleeping Beauty problem is a major one in anthropics, and my paper establishes anthropic decision theory (ADT) by a careful analysis it. Therefore we should start with an explanation of what it is.

In the standard setup, Sleeping Beauty is put to sleep on Sunday, and awoken again Monday morning, without being told what day it is. She is put to sleep again at the end of the day. A fair coin was tossed before the experiment began. If that coin showed heads, she is never reawakened. If the coin showed tails, she is fed a one-day amnesia potion (so that she does not remember being awake on Monday) and is reawakened on Tuesday, again without being told what day it is. At the end of Tuesday, she is put to sleep for ever. This is illustrated in the next figure:



The incubator variant of the problem, due to Nick Bostrom, has no initial Sleeping Beauty, just one or two copies of her created (in different, identical rooms), depending on the result of the coin flip. The name `incubator' derived from the machine that was to do the birthing of these observers. This is illustrated in the next figure:



The question then is what probability a recently awoken or created Sleeping Beauty should give to the coin falling heads or tails and it being Monday or Tuesday when she is awakened (or whether she is in Room 1 or 2).

Selfishness, selflessness and altruism

I will be using these terms in precise ways in ADT, somewhat differently from how they are usually used. A selfish agent is one whose preferences are only about their own personal welfare; a pure hedonist would be a good example. A selfless agent, on the other hand is one that cares only about the state of the world, not about their own personal welfare - or anyone else's. They might not be nice (patriots are - arguably - selfless), but they do not care about their own welfare as a terminal goal.

Altruistic agents, on the other hand, care about the welfare of everyone, not just themselves. These can be divided into total utilitarians, and average utilitarians (there are other altruistic motivations, but they aren't relevant to the paper). In summary:

Selfish	"Give me that chocolate bar"
Selfless	"Save the rainforests"
Average Utilitarian	"We must increase per capita GDP"
Total Utilitarian	"Every happy child is a gift to the world"

Anthropic Decision Theory II: Self-Indication, Self-Sampling and decisions

A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this, subsequent, and previous posts 1 2 3 4 5 6.

In the last post, we saw the Sleeping Beauty problem, and the question was what probability a recently awoken or created Sleeping Beauty should give to the coin falling heads or tails and it being Monday or Tuesday when she is awakened (or whether she is in Room 1 or 2). There are two main schools of thought on this, the Self-Sampling Assumption and the Self-Indication Assumption, both of which give different probabilities for these events.

The Self-Sampling Assumption

The self-sampling assumption (SSA) relies on the insight that Sleeping Beauty, before being put to sleep on Sunday, expects that she will be awakened in future. Thus her awakening grants her no extra information, and she should continue to give the same credence to the coin flip being heads as she did before, namely 1/2.

In the case where the coin is tails, there will be two copies of Sleeping Beauty, one on Monday and one on Tuesday, and she will not be able to tell, upon awakening, which copy she is. She should assume that both are equally likely. This leads to SSA:

- All other things being equal, an observer should reason as if they are randomly selected from the set of all actually existent observers (past, present and future) in their reference class.

There are some issues with the concept of 'reference class', but here it is enough to set the reference class to be the set of all other Sleeping Beauties woken up in the experiment.

Given this, the probability calculations become straightforward:

- **P_{SSA}(Heads) = 1/2**
- **P_{SSA}(Tails) = 1/2**
- P_{SSA}(Monday|Heads) = 1
- P_{SSA}(Tuesday|Head) = 0
- P_{SSA}(Monday|Tails) = 1/2
- P_{SSA}(Tuesday|Tails) = 1/2

By Bayes' theorem, these imply that:

- **P_{SSA}(Monday) = 3/4**
- **P_{SSA}(Tuesday) = 1/4**

The Self-Indication Assumption

There is another common way of doing anthropic probability, namely to use the self-indication assumption (SIA). This derives from the insight that being woken up on Monday after a heads, being woken up on Monday after a tails, and being woken up on Tuesday are all subjectively indistinguishable events, which each have a probability 1/2 of happening, therefore we should consider them equally probable. This is formalised as:

- All other things being equal, an observer should reason as if they are randomly selected from the set of all possible observers.

Note that this definition of SIA is slightly different from that used by Bostrom; what we would call SIA he designated as the combined SIA+SSA. We shall stick with the definition above, however, as it is coming into general use. Note that there is no mention of reference classes, as one of the great advantages of SIA is that any reference class will do, as long as it contains the observers in question.

Given SIA, the three following observer situations are equiprobable (each has an 'objective' probability 1/2 of happening), and hence SIA gives them equal probabilities of 1/3:

- P_{SIA}(Monday ∩ Heads) = 1/3
- P_{SIA}(Monday ∩ Tails) = 1/3
- P_{SIA}(Tuesday ∩ Tails) = 1/3

This allows us to compute the probabilities:

- **P_{SIA}(Monday) = 2/3**
- **P_{SIA}(Tuesday) = 1/3**
- **P_{SIA}(Heads) = 1/3**
- **P_{SIA}(Tails) = 2/3**

SIA and SSA are sometimes referred to as the thirder and halfer positions respectively, referring to the probability they give for Heads.

Probabilities and decisions

SIA and SSA give probabilities in anthropic situations, but aren't enough to give decisions. Consider the case where Sleeping Beauty has to vote on some policy, that is only implemented if voted for by all existent copies. When there are multiple copies, being identical, they will vote the same way.
Which poses the question as to how much impact each copy has on the final outcome. Do they have an individual impact, i.e. they are responsible for one n-th of the outcome if there are n copies voting? Or are they responsible for the total impact, since the result requires unanimity, and (since the copies are identical) if they voted the other way, so would the other copies?
In that situation, SIA with individual impact gives the same decision as SSA with total impact (SIA prefers worlds with large number of people in them, which also magnify

the size of total impact). So probabilities are not enough, on their own, to solve anthropic problems. Hence we will be focusing not on anthropic probabilities but on anthropic decisions. It is astonishing that one can solve the later, without making use of the former.

Anthropic Decision Theory III: Solving Selfless and Total Utilitarian Sleeping Beauty

A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this, subsequent, and previous posts 1 2 3 4 5 6.

Consistency

In order to transform the Sleeping Beauty problem into a decision problem, assume that every time she is awoken, she is offered a coupon that pays out £1 if the coin fell tails. She must then decide at what cost she is willing to buy that coupon.

The very first axiom is that of temporal consistency. If your preferences are going to predictably change, then someone will be able to exploit this, by selling you something now that they will buy back for more later, or vice versa. This axiom is implicit in the independence axiom in the von Neumann-Morgenstern axioms of expected utility, where non-independent decisions show inconsistency after partially resolving one of the lotteries. For our purposes, we will define it as:

- Temporal Consistency: If an agent at two different times has the same knowledge and preferences, then the past version will never give up anything of value in order to change the decision of the future version.

This is appropriate for the standard Sleeping Beauty problem, but not for the incubator variant, where the different copies are not future or past versions of each other. To deal with that, we extend the axiom to:

- Consistency: If two copies of an agent have the same knowledge and preferences, then the one version will never give up anything of value in order to change the decision of the other version.

Note that while 'same preferences' is something we could expect to see for the same agent at different times, it is not something the case for copies, who are generally assumed to be selfish towards each other. Indeed, this whole issue of selflessness, altruism and selfishness will be of great import for the agent's behaviour, as we shall now see.

Selfless Sleeping Beauty

Assume that Sleeping Beauty has an entirely selfless utility function. To simplify matters, we will further assume her utility is linear in cash (though cash for her is simply a tool to accomplish her selfless goals). Before Sleeping Beauty is put to sleep the first time, she will follow the following reasoning:

> "In the tails world, future copies of myself will be offered the same deal twice. Any profit they make will be dedicated to my selfless goal, so from my perspective, profits (and losses) will be doubled in the tails world. If my future copies will buy the coupon for £x, there would be an expected £0.5(2(-x + 1) + 1(-x + 0)) = £(1-3/2x) going towards my goal. Hence I would want my copies to buy whenever x<2/3."

Then by the temporal consistency axiom, this is indeed what her future copies will do. Note that Sleeping Beauty is here showing a behaviour similar to the SIA probabilities -- she is betting on 2:1 odds that she is in the world with two copies.

Selfless Incubator Sleeping Beauty

In the incubator variant, there is no initial Sleeping Beauty to make decisions for her future copies. Thus consistency is not enough to resolve the decision problem, even for a selfless Sleeping Beauty. To do so, we will need to make use of our second axiom:

- Outside agent: If there exists a collection of identical agents (which may be the same agent at different times) with same knowledge and preferences, then another copy of them with the same information would never give up anything of value to make them change their decisions.

With this axiom, the situation reduces to the above selfless Sleeping Beauty, by simply adding in the initial Sleeping Beauty again as 'another copy'. Some might feel that that axiom is too strong, that invariance under the creation or destruction of extra copies is something that cannot be simply assumed in anthropic reasoning. An equivalent axiom could be:

- Total agent: If there exists a collection of identical agents (which may be the same agent at different times) with same knowledge and preferences, then they will make their decisions as if they were a single agent simultaneously controlling all their (correlated) decisions.

This axiom is equivalent to the other, with the total agent taking the role of the outside agent. Since all the agents are identical, going through exactly the same reasoning process to reach the same decision, the total agent formulation may be more intuitive. They are, from a certain perspective, literally the same agent. This is the decision that the agents would reach if they could all coordinate with each other, if there were a way of doing this without them figuring out how many of them there were.

Altruistic total utilitarian Sleeping Beauty

An altruistic total utilitarian will have the same preferences over the possible outcomes in a Sleeping Beauty situation: the outcomes in the tails world is doubled, as any gain/loss happens twice, and the altruist adds up the effect of each gain/loss. Hence the altruistic total utilitarian Sleeping Beauty will make the same decisions as the selfless one.

Copy-altruistic total utilitarian Sleeping Beauty

The above argument does not require that Sleeping Beauty be entirely altruistic, only that she be altruistic towards all her copies. Thus she may have selfish personal preferences ("I prefer to have this chocolate bar, rather than letting Snow White get it"), as long as these are not towards her copies ("I'm indifferent as to whether I or Sleeping Beauty II gets the chocolate bar"). And then she will make the same decision in this problem as if she was entirely altruistic.

This post looked at situation implying SIA-like behaviour; tomorrow's post will look at cases where SSA-like behaviour is the right way to go.

Anthropic Decision Theory IV: Solving Selfish and Average-Utilitarian Sleeping Beauty

A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this, subsequent, and previous posts 1 2 3 4 5 6.

In the previous post, I looked at a decision problem when Sleeping Beauty was selfless or a (copy-)total utilitarian. Her behaviour was reminiscent of someone following SIA-type odds. Here I'll look at situations where her behaviour is SSA-like.

Altruistic average utilitarian Sleeping Beauty

In the incubator variant, consider the reasoning of an Outside/Total agent who is an average utilitarian (and there are no other agents in the universe apart from the Sleeping Beauties).

> "If the various Sleeping Beauties decide to pay £x for the coupon, they will make -£x in the heads world. In the tails world, they will each make £(1-x) each, so an average of £(1-x). This give me an expected utility of £0.5(-x+(1-x))= £(0.5-x), so I would want them to buy the coupon for any price less than £0.5."

And this will then be the behaviour the agents will follow, by consistency. Thus they would be behaving as if they were following SSA odds, and putting equal probability on the heads versus tails world.

For a version of this that makes senses for the classical Sleeping Beauty problem, one could imagine that she to be awaknened a week after the experiment. Further imagine she would take her winnings and losses during the experiment in the form of chocolate, consumed immediatly. Then because of the amneia drug, she would only remember one instance of this in the tails world. Hence if she valued memory of pleasure, she would want to be average utilitarian towards the pleasures of her different versions, and would follow SSA odds.

Reference classes and copy-altruistic agents

Standard SSA has a problem with reference classes. For instance, the larger the reference class becomes, the more the results of SSA in small situations become similar to SIA. The above setup mimics the effect: if there is a very large population of outsider individuals that Sleeping Beauty is altruistic towards, then the gains to two extra copies will tend to add, rather than average: if Ω is large, then 2x/(2+Ω) (averaged gain to two created agents each gaining x) is approximately twice x/(1+Ω) (averaged gain to one created agent gaining $x$), so she will behave more closely to SIA odds.

This issue is not present for *copy-altruistic* average utilitarian Sleeping Beauties, as she doesn't care about any outsiders.

Selfish Sleeping Beauty

In all of the above example, the goals of one Sleeping Beauty were always in accordance with the goals of her copies or the past and future versions of herself. But what happens when this fails? What happens when the different versions are entirely selfish towards each other? Very easy to understand in the incubator variant (the different created copies feel no mutual loyalty), it can also be understood in the standard Sleeping Beauty problem if she is a hedonist with a high discount rates.

Since the different copies do have different goals, the consistency axioms no longer apply. It seems that we cannot decide what the correct decision is in this case. There is, however, a tantalising similarity between this case and the altruistic average utilitarian Sleeping Beauty. The setups (including probabilities) are the same. By `setup' we mean the different worlds, their probabilities, the number of agents in each world, and the decisions faced by these agents. Similarly, the possible 'linked' decisions are the same. See future posts for a proper definition of linked decisions; here it just means that all copies will have to make the same decision, being identical, so there is one universal 'buy coupon' or 'reject coupon'. And, given this linking, the utilities derived by the agents is the same for either outcome in the two cases.

To see this, consider the selfish situation. Each Sleeping Beauty will make a single decision, whether to buy the coupon at the price offered. Not buying the coupon nets her £0 in all worlds. Buying the coupon at price £x nets her -£x in the heads world, and £(1-x) in the tails world. The linking is present but has no impact on these selfish agents: they don't care what the other copies decide.

This is exactly the same for the altruistic average utilitarian Sleeping Beauties. In the heads world, buying the coupon at price £x nets her -£x worth of utility. In the tails world, it would net the current copy £(1-x) worth of individual utility. Since the copies are identical (linked decision), this would happen twice in the tails world, but since she only cares about the average, this grants both copies only £(1-x) worth of utility in total. The linking is present, and has an impact, but that impact is dissolved by the average utilitarianism of the copies.

Thus the two situations have the same setup, the same possible linked decisions and the same utility outcomes for each possible linked decision. It would seem there is nothing relevant to decision theory that distinguishes these two cases. This gives us the last axiom:

- Isomorphic decisions: If two situations have the same setup, the same possible linked decisions and the same utility outcomes for each possible linked decision, and all agents are aware of these facts, then agents should make the same decisions in both situations.

This axiom immediately solves the selfish Sleeping Beauty problem, implying that agents there must behave as they do in the altruistic average utilitarian Sleeping Beauty problem, namely paying up to £0.50 for the coupon. In this way, the selfish agents also behave as if they were following SSA probabilities, and believed that heads and tails were equally likely.

Summary of results

We have broadly four categories of agents, and they follow two different types of decisions (SIA-like and SSA-like). In the Sleeping Beauty problem (and in more general problems), the categories decompose as:

1. Selfless agents who will follow SIA-type odds.
2. (Copy-)Altruistic total utilitarians who will follow SIA-type odds.
3. (Copy-)Altruistic average utilitarians who will follow SSA-type odds.
4. Selfish agents who will follow SSA-type odds.

For the standard Sleeping Beauty problem, the first three decisions derived from consistency. The same result can be established for the incubator variants using the Outside/Total agent axioms. The selfish result, however, needs to make use of the Isomorphic decisions axiom.

EDIT: A good question from Wei Dai illustrates the issue of precommitment for selfish agents.

Anthropic Decision Theory V: Linking and ADT

A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this, subsequent, and previous posts 1 2 3 4 5 6.

Now that we've seen what the 'correct' decision is for various Sleeping Beauty Problems, let's see a decision theory that reaches the same conclusions.

Linked decisions

Identical copies of Sleeping Beauty will make the same decision when faced with same situations (technically true until quantum and chaotic effects cause a divergence between them, but most decision processes will not be sensitive to random noise like this). Similarly, Sleeping Beauty and the random man on the street will make the same decision when confronted with a twenty pound note: they will pick it up. However, while we could say that the first situation is linked, the second is coincidental: were Sleeping Beauty to refrain from picking up the note, the man on the street would not so refrain, while her copy would.

The above statement brings up subtle issues of causality and counterfactuals, a deep philosophical debate. To sidestep it entirely, let us recast the problem in programming terms, seeing the agent's decision process as a deterministic algorithm. If agent α is an agent that follows an automated decision algorithm A, then if A knows its own source code (by quining for instance), it might have a line saying something like:

> Module M: If B is another algorithm, belonging to agent β, identical with A ('yourself'), assume A and B will have identical outputs on identical inputs, and base your decision on this.

This could lead, for example, to α and β cooperating in a symmetric Prisoner's Dilemma. And there is no problem with A believing the above assumption, as it is entirely true: identical deterministic algorithms on the same input do produce the same outputs. With this in mind, we give an informal definition of a linked decision as:

> **Linked decisions**: Agent α's decisions are linked with agent β's, if both can prove they will both make the same decision, even after taking into account the fact they know they are linked.

An example of agents that are *not* linked would be two agents α and β, running identical algorithms A and B on identical data, except that A has module M while B doesn't. Then A's module might correctly deduce that they will output the same decision, but only if A disregards the difference between them, i.e.module M. So A can 'know' they will output the same decision, but if it acts on that knowledge, it makes it incorrect. If A and B both had module M, then they could both act on the knowledge and it would remain correct.

ADT

Given the above definition, anthropic decision theory (ADT) can be simply stated as:

> **Anthropic Decision Theory** (ADT): An agent should first find all the decisions linked with their own. Then they should maximise expected utility, acting as if they simultaneously controlled the outcomes of all linked decisions, and using the objective (non-anthropic) probabilities of the various worlds.

ADT is similar to SSA in that it makes use of reference classes. However, SSA needs to have the reference class information established separately before it can calculate probabilities, and different reference classes give very different results. In contrast, the reference class for ADT is part of the definition. It is not the class of identical or similar agents; instead, it is the class of linked decisions which (by definition) is the class of decisions that the agent can prove are linked. Hence the whole procedure is perfectly deterministic, and known for a given agent.

It can be seen that ADT obeys all the axioms in the Sleeping Beauty problems, so must reach the same conclusions as there.

Linking non-identical agents

Now, module M is enough when the agents/algorithms are strictly identical, but fails when they differ slightly. For instance, imagine a variant of the selfless Sleeping Beauty problem where the two agents aren't exactly identical in tails world. The first agent has the same utility as before, while the second agent has some personal displeasure in engaging in trade -- if she buys the coupon, she will suffer a single -£0.05 penalty for doing so.

Then if the coupon is priced at £0.60, something quite interesting happens. If the agents do not believe they are linked, they will refuse the offer: their expected returns are 0.5(-0.6 + (1-0.6)) = -0.1 and -0.1-0.05=-0.15 respectively. If however they believe their decisions are linked, they will calculate the expected return from buying the coupon as 0.5 (-0.60 + 2(1-0.60)) = 0.1 and 0.1-0.05 = 0.05 respectively. Since these are positive, they will buy the coupon: meaning their assumption that they were linked was actually correct!

If the coupon is priced at £0.66, things change. If the two agents assume their decisions are linked, then they will calculate their expected return from buying the coupon as 0.5(-0.66 + 2(1-0.66))= 0.01 and 0.01-0.05=-0.04 respectively. The first agent will buy, and the second will not -- they were wrong to assume they were linked

A more general module that gives this kind of behaviour is:

> Module N: Let H be the hypothesis that the decision of A ('myself') and those of algorithm B are linked. I will then compute what each of us will decide if we were both to accept H. If our ultimate decisions are indeed the same, and if the other agent also has a module N, then I will accept H.

The module N gives correct behaviour. It only triggers if the agents can prove that accepting H will ensure that H is true -- and then N makes them accept H, hence making H true.

For the coupon priced at £0.60, it will correctly tell them they are linked, and they will both buy it. For the coupon priced at £0.66, it will not trigger, and both will refuse to buy it -- though they reach the same decision, they will not have done so if they had assumed they were linked. For a coupon priced above £2/3, module N will correctly tell them are linked again, and they will both refuse to buy it.

Anthropic Decision Theory VI: Applying ADT to common anthropic problems

A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this and previous posts 1 2 3 4 5 6.

Having presented ADT previously, I'll round off this mini-sequence by showing how it behaves with common anthropic problems, such as the Presumptuous Philosopher, Adam and Eve problem, and the Doomsday argument.

The Presumptuous Philosopher

The Presumptuous Philosopher was introduced by Nick Bostrom as a way of pointing out the absurdities in SIA. In the setup, the universe either has a trillion observers, or a trillion trillion trillion observers, and physics is indifferent as to which one is correct. Some physicists are preparing to do an experiment to determine the correct universe, until a presumptuous philosopher runs up to them, claiming that his SIA probability makes the larger one nearly certainly the correct one. In fact, he will accept bets at a trillion trillion to one odds that he is in the larger universe, repeatedly defying even strong experimental evidence with his SIA probability correction.

What does ADT have to say about this problem? Implicitly, when the problem is discussed, the philosopher is understood to be selfish towards any putative other copies of himself (similarly, Sleeping Beauty is often implicitly assumed to be selfless, which may explain the diverge of intuitions that people have on the two problems). Are there necessarily other similar copies? Well, in order to use SIA, the philosopher must believe that there is nothing blocking the creation of presumptuous philosophers in the larger universe; for if there was, the odds would shift away from the larger universe (in the extreme case when only one presumptuous philosopher is allowed in any universe, SIA finds them equi-probable). So the expected number of presumptuous philosophers in the larger universe is a trillion trillion times greater than the expected number in the small universe.

Now if the philosopher is indeed selfish towards his copies, then ADT reduces to SSA-type behaviour: the philosopher will correctly deduce that in the larger universe, the other trillion trillion philosophers or so will have their decision linked with his. However, he doesn’t care about them: any benefit that accrue to them are not of his concern, and so if he correctly guesses that he resides in the larger universe, he will accrue a single benefit. Hence there will be no paradox: he will bet at 1:1 odds of residing in either the larger or the smaller universe.

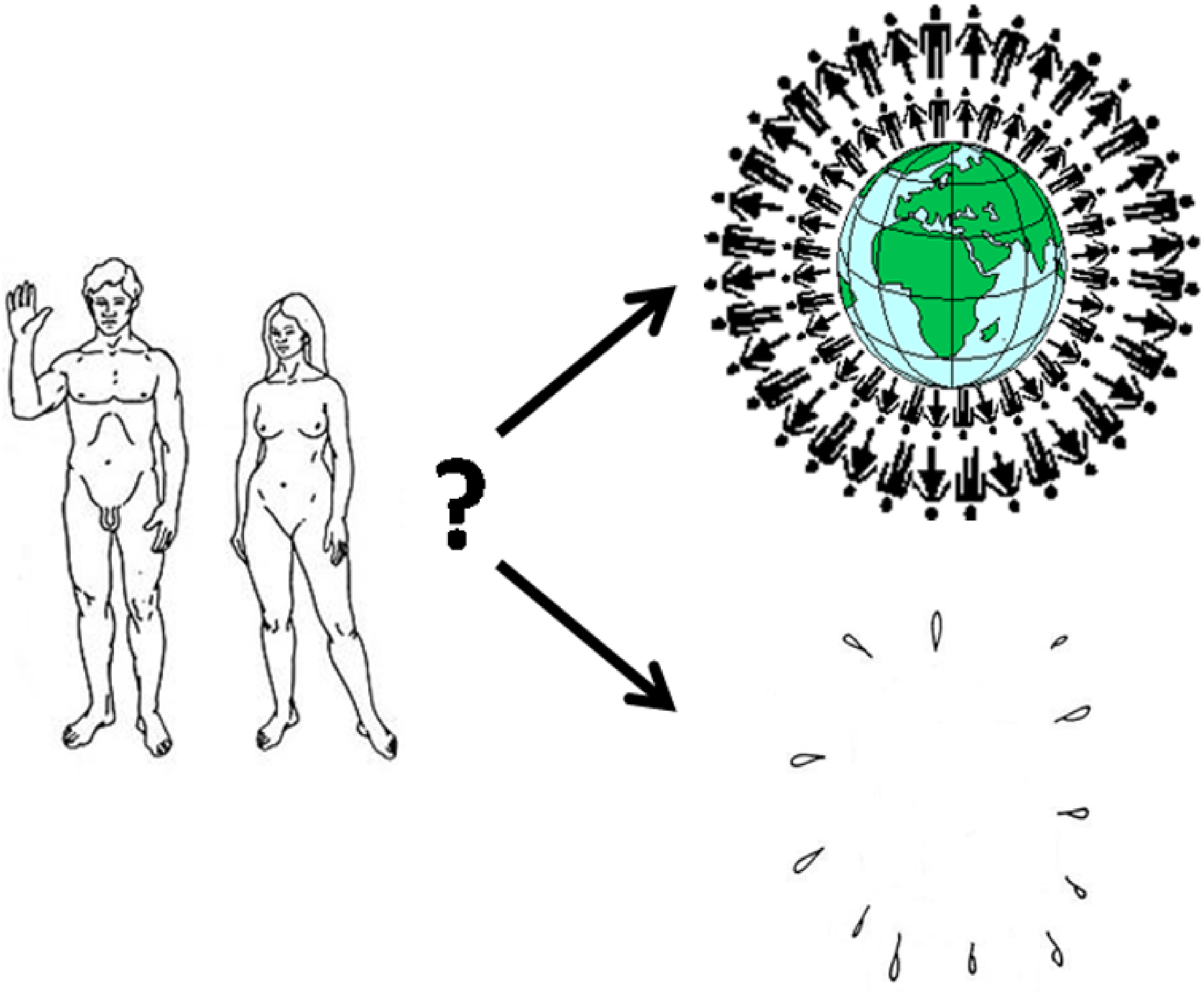
If the philosopher is an altruistic total utilitarian, on the other hand, he will accept bets at odds of a trillion trillion to one of residing in the larger universe. But this no longer counter-intuitive (or at least, no more counter-intuitive than maximising expect utility with very small probabilities): the other presumptuous philosophers will make the same bet, so in the larger universe, their total profit and loss will be multiplied by a trillion trillion. And since the philosopher is altruistic, the impact on his own utility is multiplied by a trillion trillion in the large universe, making his bets rational.

At this point, it might be fair to ask what would happen if some of the philosophers were altruistic while others were selfish. How would the two interact; would the selfless philosopher be incorrectly believing his own decision was somehow ‘benefiting’ the selfish ones? Not at all. The decisions of the selfless and selfish philosophers are not linked: they both use ADT, but because they have very different utilities, they cannot prove that their decisions are linked. Which is fortunate, because they aren’t.

Adam and Eve

The presumptuous philosopher thought experiment was designed to show up problems with SIA reasoning. Another thought experiment by the same author as designed to show problems with SSA reasoning.

In this thought experiment, Adam and Eve, the first two humans, have the opportunity to breed or not to breed. If they breed, they will produce trillions of trillions of descendants. Under SSA odds, the probability of being Adam or Eve in a universe with trillions of trillion humans is tiny, while the corresponding probability in a universe with just two observers is one. Therefore Adam and Eve should conclude that whatever they do, it is tremendously unlikely that they will succeed in breeding. Nick Bostrom them proceeds to draw many amusing consequences from this ‘future affecting’ paradox, such as the couple forming the firm intention of having sex (and hence risking pregnancy) if an edible animal doesn’t wander through the entrance of their cave in the next few minutes. There seems to be something very wrong with the reasoning, but it is a quite natural consequence of SSA.



What does ADT have to say on the problem? This depends on whether the decisions of Adam and Eve and their (potential) descendants are linked. There is prima facia no reason for this to be the case; in any case, how could potential future descendants make the decision as to whether Adam and Eve should breed? One way of imagining this is if each human is born fully rational, aware of the possible world they is born into, but in ignorance as to their identity and position in that world. Call this the ignorant rational baby stage. They can then make a decision as to what they would do, conditional upon discovering their identity. They may then decide or not to stick to their decision. Hence we can distinguish several scenarios:

- These agents have no ‘ignorant rational baby stage’, and do not take it into account.
- These agents have an ‘ignorant rational baby stage’, but do not allow precommitments.
- These agents have an ‘ignorant rational baby stage’, but do allow precommitments.
- These agents have no ‘ignorant rational baby stage’, but require themselves to follow the hypothetical precommitments they would have made had they had such a stage.

To make this into a decision problem, assume all agents are selfish, and know they will be confronted by a choice between a coupon C_1 that pays out £1 to Adam and Eve if they have no descendants and C_2 that pays out £1 to Adam and Eve if they have (trillions of trillions of) descendants. Assume Adam and Eve will have sex exactly once, and the (objective) chance of them having a successful pregnancy is 50%. Now each agent must decide on the relative values of the two coupons.

Obviously in situation 1, the decisions of Adam and Eve and their descendants are not linked, and ADT means that Adam and Eve will value C_1 compared with C_2 according to their objective estimates of having descendants, i.e. they will value then equally. There is no SSA-like paradox here. Their eventual descendants will also value the coupons as equally worthless, as they will never derive any value from them.

Now assume there is a 'ignorant rational baby stage'. During this stage, the decisions of all agents are linked, as they have the same information, the same (selfish) preferences, and they all know this. Each rational baby can then reason:

> "If I am in the world with no descendants, then I am Adam or Eve, and the C_1 coupon is worth £1 to me (and C_2 is worthless). If, on the other hand, I am in the world with trillions of trillions of descendants, there is only two chances in $2+10^{24}$ of me being Adam or Eve, so I value the C_2 coupon at £$2/(2+10^{24})$ (and C_1 is worthless). These worlds are equiprobable. So I would value C_1 as being $1+0.5 \times 10^{24}$ times more valuable than C_2."

So the rational babies in situations 2 and 3 would take C_1 as much more valuable than C_2, if the deal was proposed to them immediately. Since there are no precommitments in situation 2, once the rational babies discover who they actually are, they would revert to situation 1 and take them as equally valuable. If precommitments are allowed, then the rational babies would further reason:

> "The strategy 'upon discovering I am Adam or Eve, take C_1' nets me an expected £1/2, while the strategy 'upon discovering I am Adam or Eve, take C_2' nets me an expected £$1/(2+10^{24})$, because it is very unlikely that I would actually discover that I am Adam and Eve. Hence the strategy 'upon discovering I am Adam or Eve, accept trades between C_1 and C_2 at $2:(2+10^{24})$ ratios' is neutral in expected utility, and so I will now precommit to accepting any trade at any ratios slightly better than this."

So in situation 3, even after discovering that they are Adam or Eve, they will continue to accept deals at ratios that would seem to imply that they believe in the SSA odds, i.e. that they are nearly certain to not have descendants. But it is a lot less of a paradox now; it simply arises because there was a time when they were uncertain as to what their actual position was, and the effects of this uncertainty were 'locked in' by their precommitment.

Situation 4 is very interesting. By construction, it reproduces the seemingly paradoxical behaviour, but here there was never a rational baby stage where the behaviour made sense. Why would any agent follow such a behaviour? Well, mainly because it allows trade between agents who might not otherwise be able to agree on a 'fair' distribution of goods. If all agents agree to the division that they would have wanted had they been ignorant of their identity (a 'rawlsian veil of ignorance' situation), then they can trade between each other without threats or bargaining in these simple cases.

If the agents are simply altruistic average utilitarians, then Adam and Eve would accept SSA odds in all four situations; things that benefit them specifically are weighted more highly in a universe with few people. So the varying behaviour above is a feature of selfishness, not of SSA-type behaviour, and it seems precommitments become very important in the selfish case. This certainly merits further study. Temporal consistency is a virtue, but does it extend to situations like this, where the agent makes binding decisions before knowing their identity? Certainly *if* the agent had to make a decision immediately, and if there were anyone around to profit from temporal inconsistencies, the agent should remain consistent, which means following precommitments. However this is not entirely obvious that it should still be the case if there were no-one to exploit the inconsistency.

This is not, incidentally, a problem only of ADT - SIA has similar problem under the creation of ‘irrelevant’ selfless agents who don’t yet know who they are, while SSA has problems under the creation of agents who don’t yet know what reference class they are in.

The Doomsday argument

Closely related to the Adam and Eve paradox, though discovered first, is the Doomsday argument. Based on SSA’s preference for ‘smaller’ universes, it implies that there is a high probability of the human race becoming extinct within a few generations - at the very least, a much higher probability than objective factors would imply.

Under SIA, the argument goes away, so it would seem that ADT must behave oddly: depending on the selfishness and selflessness of the agents, they would give different probabilities to the extinction of the human race. This is not the case, however. Recall that under ADT, decisions matter, not probabilities. And agents that are selfish or average utilitarians would not be directly concerned with the extinction of the human race, so would not act in bets to prevent this.

This is not a specious point - there are ways of constructing the doomsday argument in ADT, but they all rely on odd agents who are selfish with respect to their own generation but selfless with respect to the future survival of the human race. This lacks the potency of the original formulation: having somewhat odd agents behaving in a somewhat odd fashion is not very surprising. For the moment, until a better version is produced, we should simply say that the doomsday argument is not present in ADT.

Sleeping Anti-Beauty

Sleeping Anti-Beauty is a thought experiment similar to the Sleeping Beauty experiment, but with one important caveat: the two copies in the tails world hate each other. This works best if the two copies are duplicates, rather than the same person at different times. One could imagine, for instance, that a week after the experiment, all copies of Sleeping Beauty are awakened and made to fight to the death - maybe they live in a civilization that prohibits more than one copy of an agent from existing. The single copy in the heads world will be left unmolested, as she has nobody to fight.

That means that the Sleeping Beauties in the tail world are in a zero sum game; any gain for one is a loss for the other, and vice-versa. Actually, we need a few more assumptions for this to be true: the Sleeping Beauties have to be entirely selfish apart from their rivalry, and they do not get offered any goods for immediate consumption. Given all these assumptions, what does ADT have to say about their decisions?

As usual, all existent copies of Sleeping Beauty are offered a coupon that pays out £1 if the coin fell tails, and asked how much she would be willing to give for that. ADT reasoning proceeds for each agent as follows:

> “In the heads world, if I pay £x, all that happens is that I lose £x. In the tails, world, if I pay £x, I gain £(1-x). However my other hated copy will make the same decision, and also gain £(1-x). This causes me the same amount of loss as the gain of £(1-x) does, so I gain nothing at all in the tails world, whatever £x is. So I value the coupon precisely at zero: I would not pay any amount to get it.”

In this, and other similar decisions, Sleeping Beauty would act as if she had an absolute certainty of being in the heads world, offering infinity to one odds of this being the case, as she cannot realise any gains - or losses! - in the tails world.

It should be noted that selfish Sleeping Beauty can be correctly modelled by seeing it as a 50-50 mix of selfless Sleeping Beauty and Sleeping Anti-Beauty.

Anthropic Decision Theory V: Linking and ADT

A near-final version of my Anthropic Decision Theory paper is available on the arXiv. Since anthropics problems have been discussed quite a bit on this list, I'll be presenting its arguments and results in this, subsequent, and previous posts 1 2 3 4 5 6.

Now that we've seen what the 'correct' decision is for various Sleeping Beauty Problems, let's see a decision theory that reaches the same conclusions.

Linked decisions

Identical copies of Sleeping Beauty will make the same decision when faced with same situations (technically true until quantum and chaotic effects cause a divergence between them, but most decision processes will not be sensitive to random noise like this). Similarly, Sleeping Beauty and the random man on the street will make the same decision when confronted with a twenty pound note: they will pick it up. However, while we could say that the first situation is linked, the second is coincidental: were Sleeping Beauty to refrain from picking up the note, the man on the street would not so refrain, while her copy would.

The above statement brings up subtle issues of causality and counterfactuals, a deep philosophical debate. To sidestep it entirely, let us recast the problem in programming terms, seeing the agent's decision process as a deterministic algorithm. If agent α is an agent that follows an automated decision algorithm A, then if A knows its own source code (by quining for instance), it might have a line saying something like:

> Module M: If B is another algorithm, belonging to agent β, identical with A ('yourself'), assume A and B will have identical outputs on identical inputs, and base your decision on this.

This could lead, for example, to α and β cooperating in a symmetric Prisoner's Dilemma. And there is no problem with A believing the above assumption, as it is entirely true: identical deterministic algorithms on the same input do produce the same outputs. With this in mind, we give an informal definition of a linked decision as:

> **Linked decisions**: Agent α's decisions are linked with agent β's, if both can prove they will both make the same decision, even after taking into account the fact they know they are linked.

An example of agents that are *not* linked would be two agents α and β, running identical algorithms A and B on identical data, except that A has module M while B doesn't. Then A's module might correctly deduce that they will output the same decision, but only if A disregards the difference between them, i.e.module M. So A can 'know' they will output the same decision, but if it acts on that knowledge, it makes it incorrect. If A and B both had module M, then they could both act on the knowledge and it would remain correct.

ADT

Given the above definition, anthropic decision theory (ADT) can be simply stated as:

> **Anthropic Decision Theory** (ADT): An agent should first find all the decisions linked with their own. Then they should maximise expected utility, acting as if they simultaneously controlled the outcomes of all linked decisions, and using the objective (non-anthropic) probabilities of the various worlds.

ADT is similar to SSA in that it makes use of reference classes. However, SSA needs to have the reference class information established separately before it can calculate probabilities, and different reference classes give very different results. In contrast, the reference class for ADT is part of the definition. It is not the class of identical or similar agents; instead, it is the class of linked decisions which (by definition) is the class of decisions that the agent can prove are linked. Hence the whole procedure is perfectly deterministic, and known for a given agent.

It can be seen that ADT obeys all the axioms in the Sleeping Beauty problems, so must reach the same conclusions as there.

Linking non-identical agents

Now, module M is enough when the agents/algorithms are strictly identical, but fails when they differ slightly. For instance, imagine a variant of the selfless Sleeping Beauty problem where the two agents aren't exactly identical in tails world. The first agent has the same utility as before, while the second agent has some personal displeasure in engaging in trade -- if she buys the coupon, she will suffer a single -£0.05 penalty for doing so.

Then if the coupon is priced at £0.60, something quite interesting happens. If the agents do not believe they are linked, they will refuse the offer: their expected returns are 0.5(-0.6 + (1-0.6)) = -0.1 and -0.1-0.05=-0.15 respectively. If however they believe their decisions are linked, they will calculate the expected return from buying the coupon as 0.5 (-0.60 + 2(1-0.60)) = 0.1 and 0.1-0.05 = 0.05 respectively. Since these are positive, they will buy the coupon: meaning their assumption that they were linked was actually correct!

If the coupon is priced at £0.66, things change. If the two agents assume their decisions are linked, then they will calculate their expected return from buying the coupon as 0.5(-0.66 + 2(1-0.66))= 0.01 and 0.01-0.05=-0.04 respectively. The first agent will buy, and the second will not -- they were wrong to assume they were linked

A more general module that gives this kind of behaviour is:

> Module N: Let H be the hypothesis that the decision of A ('myself') and those of algorithm B are linked. I will then compute what each of us will decide if we were both to accept H. If our ultimate decisions are indeed the same, and if the other agent also has a module N, then I will accept H.

The module N gives correct behaviour. It only triggers if the agents can prove that accepting H will ensure that H is true -- and then N makes them accept H, hence making H true.

For the coupon priced at £0.60, it will correctly tell them they are linked, and they will both buy it. For the coupon priced at £0.66, it will not trigger, and both will refuse to buy it -- though they reach the same decision, they will not have done so if they had assumed they were linked. For a coupon priced above £2/3, module N will correctly tell them are linked again, and they will both refuse to buy it.

"Solving" selfishness for UDT

With many thanks to Beluga and lackofcheese.

When trying to decide between SIA and SSA, two anthropic probability theories, I concluded that the question of anthropic probability is badly posed and that it depends entirely on the values of the agents. When debating the issue of personal identity, I concluded that the question of personal identity is badly posed and depends entirely on the values of the agents. When the issue of selfishness in UDT came up recently, I concluded that the question of selfishness is...

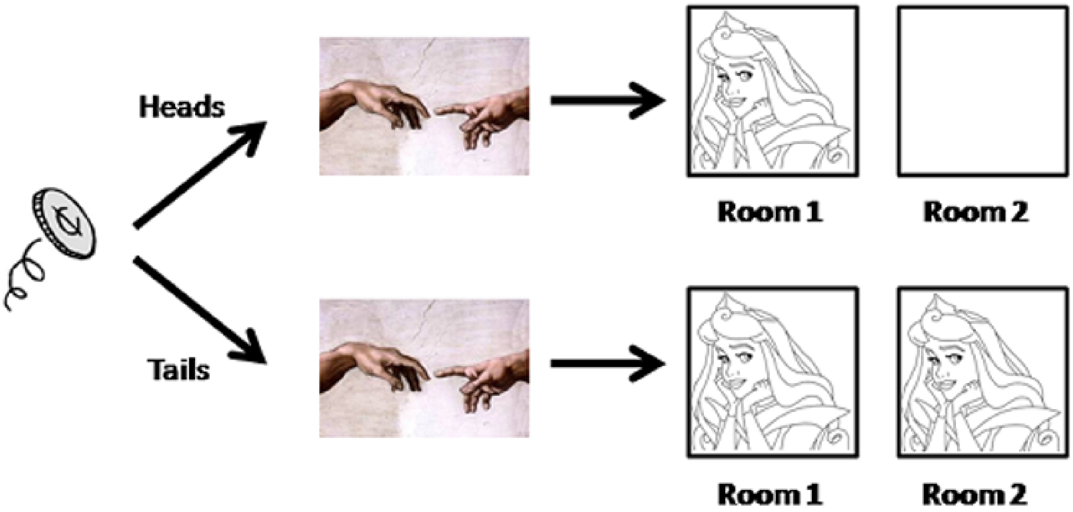
But let's not get ahead of ourselves.

A selfish scenario

Using Anthropic Decision Theory, I demonstrated that selfish agents using UDT should reason in the same way that average utilitarians did - essentially behaving 'as if' SSA were true and going for even odds of heads and tail ("halfer") in the Sleeping Beauty problem.

Then Beluga posted an argument involving gnomes, that seemed to show that selfish UDT agents should reason as total utilitarians did - essentially behaving 'as if' SIA were true and going for 2:1 odds of heads and tail ("thirder") in the Sleeping Beauty problem. After a bit of back and forth, lackofcheese then refined the argument. I noticed the refined argument was solid, and incidentally made the gnomes unnecessary.

How does the argument go? Briefly, a coin is flipped and an incubator machine creates either one person (on heads) or two people (on tails), each in separate rooms.



Without knowing what the coin flip was or how many people there were in the universe, every new person is presented with a coupon that pays £1 if the coin came out tails. The question is - assuming utility is linear in money - what amount £x should the created person(s) pay for this coupon?

The argument from Beluga/lackofcheese can be phrased like this. Let's name the people in the tails world, calling them Jack and Roger (yes, they like dressing like princesses - what of it?). Each of them reasons something like this:

> "There are four possible worlds here. In the tails world, I, Jack/Roger, could exist in Room 1 or in Room 2. And in the heads world, it could be either me existing in Room 1, or the other person existing in Room 1 (in which case I don't exist). I'm completely indifferent to what happens in worlds where I don't exist (sue me, I'm selfish). So if I buy the coupon for £x, I expect to make utility: **0.25(0) + 0.25(-x) + 0.5(1-x)=0.5-0.75x**. Therefore I will buy the coupon for **x<£2/3**."

That seems a rather solid argument (at least, if you allow counterfactuals into worlds where you don't exist, which you probably should). So it seems I was wrong and that selfish agents will indeed go for the SIA-like "thirder" position.

Not so fast...

Another selfish scenario

The above argument reminded me of one I made a long time ago, when I "proved" that SIA was true. I subsequently discarded that argument, after looking more carefully into the motivations of the agents. So let's do that now.

Above, I was using a subtle intuition pump by using the separate names Jack and Roger. That gave connotations of "I, Jack, don't care about worlds in which I, Jack, don't exist..." But in the original formulation of the Sleeping Beauty/incubator problem, the agents were strictly identical! There is no Jack versus Roger issues - at most, these are labels, like 1 and 2.

It therefore seems possible that the selfish agent could reason:

> "There are three possible worlds here. In the tails world, I either exist in Room 1 or Room 2. And in the heads world, either I exist in Room 1, or an identical copy of me exists in Room 1, and is the only copy of me in that world. *I fail to see any actual difference between those two scenarios.* So if I buy the coupon for £x, I expect to make utility: **0.5(-x) + 0.5(1-x)=0.5-x**. Therefore I will buy the coupon for **x<£1/2**."

The selfish agent seems on rather solid ground here in their heads world reasoning. After all, would we treat someone else differently if we were told "That's not actually your friend; instead it's a perfect copy of your friend, while the original never existed"?

Notice that even if we do allow for the Jack/Roger distinction, it seems reasonable for the agent to say "If I don't exist, I value the person that most closely resembles me." After all, we all change from moment to moment, and we value our future selves. This idea is akin to Nozick's "closest continuer" concept.

Each selfish person is selfish in their own unique way

So what is really going on here? Let's call the first selfish agent a thirder-selfish agent, and the second a halfer-selfish agent. Note that both types of agents have perfectly consistent utility functions defined in all possible actual and counterfactual universes (after giving the thirder-selfish agent some arbitrary constant C, which we may as well set to zero, in worlds where they don't exist). Compare the two versions of Jack's utility:

	"Jack in Room 1"	"Roger in Room 1"
Heads: buy coupon	-x/-x	**0/-x**
Heads: reject coupon	0/0	0/0
Tails: buy coupon	1-x/1-x	1-x/1-x
Tails: reject coupon	0/0	0/0

The utilities are given as thirder-selfish utility/halfer-selfish utility. The situation where there is a divergence is indicated in bold - that one difference is key to their different decisions.

At this point, people could be tempted to argue as to which type of agent is *genuinely* the selfish agent... But I can finally say:

- The question of selfishness is badly posed and depends entirely on the values of the agents.

What do I mean by that? Well, here is a selfish utility function: "I expect all future copies of Stuart Armstrong to form a single continuous line through time, changing only slowly, and I value the happiness (or preference satisfaction) of all these future copies. I don't value future copies of other people."

That seems pretty standard selfishness. But this is not a utility function; it's a partial description of a class of utility functions, defined only in one set of universes (the set where there's a single future timeline for me, without any "weird" copying going on). Both the thirder-selfish utility function and the halfer-selfish one agree in such single timeline universes. They are therefore both extensions of the same partial selfish utility to more general situations.

Arguing which is "correct" is pointless. Both will possess all the features of selfishness we've used in everyday scenarios to define the term. We've enlarged the domain of possible scenarios beyond the usual set, so our concepts, forged in the usual set, can extend in multiple ways.

You could see the halfer-selfish values as a version of the "Psychological Approach" to personal identity: it values the utility of the being closest to itself in any world. A halfer-selfish agent would cheerfully step into a teleporter where they are scanned, copied onto a distant location, then the original is destroyed. The thirder-selfish agent might not. Because the thirder-selfish agent is actually underspecified: the most extreme version would be one that does not value any future copies of themselves. They would indeed "jump off a cliff knowing smugly that a different person would experience the consequence of hitting the ground." Most versions of the thirder-selfish agent that people have in mind are less extreme than that, but defining (either) agent requires quite a bit of work, not simply a single word: "selfish".

So it's no wonder that UDT has difficulty with selfish agents: the concept is not well defined. Selfish agent is like "featherless biped" - a partial definition that purports to be the whole of the truth.

Personal identity and values

Different view of personal identity can be seen as isomorphic with a particular selfish utility function. The isomorphism is simply done by caring about the utility of another agent if and only if they share the same personal identity.

For instance, the psychological approach to personal identity posits that "You are that future being that in some sense inherits its mental features—beliefs, memories, preferences, the capacity for rational thought, that sort of thing—from you; and you are that past being whose mental features you have inherited in this way." Thus a psychological selfish utility function would value the preferences of a being that was connected to the agent in this way.

The somatic approach posits that "our identity through time consists in some brute physical relation. You are that past or future being that has your body, or that is the same biological organism as you are, or the like." Again, this can be used to code up a utility function.

Those two approaches (psychological and somatic) are actually broad categories of approaches, all of which would have a slightly different "selfish" utility function. The non-branching view, for instance, posits that if there is only one future copy of you, that is you, but if there are two, there is no you (you're effectively dead if you duplicate). This seems mildly ridiculous, but it still expresses very clear preferences over possible worlds that can be captured in a utility function.

Some variants allow for *partial* personal identity. For instance, discounting could be represented by a utility function that puts less weight on copies more distant in the future. If you allow "almost identical copies", then these could be represented by a utility function that gives partial credit for similarity along some scale (this would tend to give a decision somewhere in between the thirder and halfer situations presented above).

Many of the "paradoxes of identity" dissolve entirely when one uses values instead of identity. Consider the intransitivity problem for some versions of psychological identity:

> First, suppose a young student is fined for overdue library books. Later, as a middle-aged lawyer, she remembers paying the fine. Later still, in her dotage, she remembers her law career, but has entirely forgotten not only paying the fine but

> everything else she did in her youth. [...] the young student is the middle-aged lawyer, the lawyer is the old woman, but the old woman is not the young student.

In terms of values, this problem is non-existent: the young student values herself, the lawyer and the old woman (as does the lawyer) but the old woman only values herself and the lawyer. That value system is inelegant, perhaps, but it's not ridiculous (and "valuing past copies" might be decision-relevant in certain counterfactual situations).

Similarly, consider the question as to whether it is right to punish someone for the law-breaking of a past copy of themselves. Are they the same person? What if, due to an accident or high technology, the present copy has no memory of law-breaking or of being the past person? Using identity gets this hopelessly muddled, but from a consequentialist deterrence perspective, the answer is simple. The past copy presumably valued their future copy staying out of jail. Therefore, from the deterrence perspective, we should punish the current copy to deter such actions. In courts today, we might allow amnesia to be a valid excuse, simply because amnesia is so hard and dangerous to produce deliberately. But this may change in the future: if it becomes easy to rewire your own memory, then deterrent punishment will need to move beyond the classical notions of identity and punish people we would currently consider blameless.

Evolution and identity

Why are we convinced that there is such a thing as selfishness and personal identity? Well, let us note that it is in the interest of evolution that we believe in it. The "interests" of the genes are to be passed on, and so they benefit if the carrier of the gene in the present values the survival of the (same) carrier of the gene in the future. The gene does not "want" the carrier to jump off a cliff, because whatever the issues of personal identity, it'll be the same gene in the body that gets squashed at the end. Similarly, future copies of yourself are the copies that you have the most control over, through your current actions. So genes have exceptionally strong interests in making you value "your" future copies. Even your twin is not as valuable as you: genetically your're equivalent, but your current decisions have less impact over them than over your future self. Thus is selfishness created.

It seems that evolution has resulted in human copies with physical continuity, influence over (future) and memories of (past), and in very strong cross-time caring between copies. These are unique to a single time line of copies, so no wonder people have seen them as "defining" personal identity. And "The person tomorrow is me" is probably more compact than saying that you care about the person tomorrow, and listing the features connecting you. In the future, the first two components may become malleable, leaving only caring (a value) as the remnants of personal identity.

This idea allows us to do something we generally can't, and directly compare the "quality" of value systems - at least from the evolutionary point of view, according to the value system's own criteria.

Here is an example of an inferior selfish decision theory: agents using CDT, and valuing all future versions of themselves, but not any other copies. Why is this inferior? Because if the agent is duplicated, they want those duplicates to cooperate and value each other equally, because that gives the current agent the best possible expected utility. But if each copy has the same utility as the agent started with, then CDT guarantees rivalry, probably to the detriment of every agent. In effect, the agent wants its future self to have different selfish/indexical values from the ones it has, in order to preserve the same overall values.

This problem can be avoided by using UDT, CDT with precommitments, or a selfish utility function that values all copies equally. Those three are more "evolutionarily stable". So is, for instance, a selfish utility function with an exponential discount rate - but not one with any other discount rate. This is an interesting feature of this approach: the set of evolutionary stable selfish decision theories is smaller than the set of selfish decision theories. Thus there are many circumstances where different selfish utilities will give the same decisions under the same decision theory, or where the different decision theories/utilities will self-modify to make identical decisions.

One would like to make an argument about Rawlsian veils of ignorance and UDT-like initial pre-commitments leading to general altruism or something... But that's another argument, for another time. Note that this kind of argument cannot be used against the most ridiculous selfish utility function of all: "me at every moment is a different person I don't value at all". Someone with *that* utility function will quickly die, but, according to its own utility, it doesn't see this as a problem.

To my mind, the interesting thing here is that while there are many "non-indexical" utility functions that are stable under self-modification, this is not the case for most selfish and indexical ones.