Changing Beliefs
Handout 5

Nilanjan Das
University College London

---

1 The Problem of Self-Locating Information

- *Self-Locating Information*. Not all information one has is objective, impersonal information about the world. Some information one has is about oneself, about where one is, or what time it is. Call the latter kind of information *self-locating*.

- *A Reflection Principle*.

  > *Popularity*. I know who I am, but I have little reason to think that I am popular. But I am rationally certain of two things. First, my friends are polite: if I ask them if I am popular, they'll say "Yes." Second, I take what my friends say at face value: if my friends say that I am popular, I'll be quite confident that I am. So, I gather all my friends around, and ask them if I am popular. Predictably, they say "Yes." So, I will raise my confidence that I am popular.

  This way of increasing my confidence seems irrational. This seems to support:

  > WEAK REFLECTION. Suppose an agent at t_1 is rationally certain that, at a later time t_2, they will rationally—without any evidence loss, or failure of factivity, or introspection—assign a credence greater than than r to a (non-self-locating) proposition P. Then, they cannot at t_1 rationally assign a credence of r to P.

- *The Claim*. In certain cases where an agent receives purely self-locating evidence without losing any non-self-locating evidence, there is no satisfactory updating rule that satisfies WEAK REFLECTION.

2 Self-Locating Content

- *The Traditional Doctrine of Propositions*.

- *Two-Place Relation*. An attitude relation (believing, desiring, etc.) is a two-place relation between an agent and a content.[1]
- *Frege's Constraint*. Contents are assigned to attitudes in a manner that accommodates Frege cases.[2]
- *Absoluteness*. The contents of attitudes are absolute, i.e. contents do not vary in truth-value across individuals or times.

- *Perry's Examples*.[3]
  - *The Messy Shopper*. Perry once followed a trail of sugar along a supermarket floor, looking for the shopper with the torn sack to tell him that he was making a mess. With each trip around the store, the trail became thicker, but there was no sign of the messy shopper. Finally, Perry realized that he was the shopper with the torn sack that he was trying to catch. Having realized this, Perry of course stopped following the trail and turned the torn sack upright.
  - *The Stanford Amnesiac*. The amnesiac Rudolph Lingens is lost in the Stanford library. Lingens's amnesia is severe, and he has forgotten who he is. After reading a biography of Rudolph Lingens, he has a belief he could express by saying, Rudolph Lingens has been to Paris. But at the same time, he also has a belief that he could express by saying, I have never been to Paris.

- *Perry's Argument*.
  - In these cases, both Perry and Lingens don't know something about themselves. Perry doesn't know *de se* that he is making the mess. Likewise, Lingens doesn't know *de se* that he has been to Paris.
  - Whatever contents they don't know cannot be captured by any content that is absolute. For example, the content expressed by, "I am making a mess," may be true of Perry but not of you or me.

  So, the traditional doctrine of propositions—insofar as it incorporates *Absoluteness*—is false.[4]

- *Lewis' Argument*. Suppose mad Heimson believes *de se* that he is Hume.

[1] An attitude (type) like the belief that snow is white is a mental state that consists in having a certain relational property, i.e. the property of standing in the relation of believing to the content that snow is white.

[2] So if a rational agent could have a belief he could express by saying, Hesperus is bright without having a belief he could express by saying, Phosphorus is bright, these two beliefs have different contents.

[3] cf. Perry (1979), 21-22

[4] It's not immediately obvious that these aren't just Frege cases.

> Pushing my cart down the aisle I was looking for Clark Kent to tell him he was making a mess. I kept passing by Superman, but couldn't find Clark Kent. Finally, I realized, Superman was Clark Kent. I believed at the outset that Clark Kent was making a mess... But I didn't believe that Superman was making a mess. That seems to be something that I came to believe. And when I came to believe that, I stopped looking around and I told Superman to clean up after himself. My change in beliefs seems to explain my change in behavior. (Cappelen and Dever 2013, 33)

> The second problem arises when we ask why Heimson is wrong. He believes he is Hume. Hume believed that too. Hume was right. If Hume believed he was Hume by believing a proposition, that proposition was true. Heimson believes just what Hume did. But Hume and Heimson are worldmates. Any proposition true for Hume is likewise true for Heimson. So Heimson, like Hume, believes he is Hume by believing a true proposition. So he's right. But he's not right. He's wrong, because he believes he's Hume and he isn't.
>
> There are two ways out. (1) Heimson does not, after all, believe what Hume did. Or (2) Heimson does believe what Hume did, but Heimson believes falsely what Hume believed truly. (Lewis 1979, 142)

Lewis then offers an argument against (1).[5] Suppose Heimson is an exact replica of Hume's. Then, if Heimson doesn't have the same belief as Hume, then beliefs can't be intrinsic properties of an agent. But beliefs are intrinsic properties of an agent. So, Heimson can't have the same belief as Hume.

For now, let's just that these arguments are sound, and explore their consequences.

3 The Initial Problem for Bayesianism

In what follows, we shall adopt Lewis' way of representing self-locating contents.[6]

- *Centred Propositions*. I'll represent contents of mental states as *centred propositions*, i.e., sets of *centred worlds*.

  > *Centred Worlds*. A centred world is just a triple $\langle w, i, t\rangle$ where w is a (metaphysically or epistemically) possible world, i is a subject who exists in w, and t is a time at which i exists in w.

  A centred proposition P is *non-self-locating* iff, for any centred world $\langle w, i, t\rangle$ in P, any other centred world $\langle w, i^*, t^*\rangle$ that shares the same world-coordinate is also in P.

[5] Ninan (2016): "For consider Oscar, here on Earth, where H2O fills the oceans and falls from the sky, and Twin Oscar, on far-away Twin Earth, where XYZ fills the oceans and falls from the sky (Putnam 1973). We can suppose that Oscar's head and Twin Oscar's head are in perfect match' in every way that is at all relevant to what they believe. Suppose that Oscar has a belief he could express in English by saying, Water contains hydrogen, and that Twin Oscar has a belief could express in Twin English by saying, Water contains hydrogen. Given Lewis's internalist premise, it would seem to follow that the content of Oscar's belief is identical to the content of Twin Oscar's belief. And yet Oscar's belief is true, while Twin Oscar's belief is false (assuming XYZ contains no hydrogen and that "hydrogen" in Twin English refers to hydrogen). Thus the content they believe cannot be an absolute proposition."

[6] Examples:

- The self-locating information that I am Nilanjan corresponds to the set of centred worlds $\langle w, i, t\rangle$ where i is Nilanjan.
- The self-locating information that it is now noon corresponds to the set of centred worlds $\langle w, i, t\rangle$ where t is noon on some day.
- The *non*-self-locating information that Nilanjan is asleep at noon on 27/02/2019 correspsonds to the set of centred worlds $\langle w, i, t\rangle$ where w is a world in which Nilanjan is asleep at noon on 27/02/2019.

- *A Revised Conception of Evidence*. Suppose we accept:

  THE REVISED PROPOSITIONALIST CONCEPTION OF EVIDENCE. An agent's total evidence at any time is a centred proposition.

- *Purely Self-Locating Evidence*. Between any two times t_1 and t_2, an agent receives purely self-locating evidence (without losing any non-self-locating evidence) just in case:
  - At t_1, the agent's total evidence is E_1, and at t_2, their total evidence is E_2.
  - For any non-self-locating centred proposition P, E_1 entails P iff E_2 entails P.
  - E_1 and E_2 are distinct.

- *An Example*. Suppose I am looking at a clock. At present, my evidence entails that it's now noon, and I am rationally certain that it's now noon.
  - Here, my total evidence is a centred proposition E_1, such that, for any centred world $\langle w, i, t\rangle$ in E, t is noon on some day.
  - Then, I learn that it's 12.01 pm without acquiring non-self-locating evidence. So, my evidence no longer entails that it's noon. Rather, my total evidence is a distinct centred proposition E_2, such that, for any centred world $\langle w, i, t^*\rangle$ in E_2, t is 12.01 pm on some day.

  Intuitively, I should now be certain that it's 12.01 pm. Can Bayesians explain this change?

- *The Answer*. They cannot.
  - For any Bayesian, if an agent's prior credence function at t_1 is p and the strongest evidence they gain (without losing any evidence) between t_1 and t_2 is E, then their posterior credence in any proposition H should be $p(H|E) = \frac{p(H \cap E)}{p(E)}$ (provided $p(E) > 0$).
  - In this case, the agent is losing their earlier evidence that it's now noon, and the strongest evidence that they are gaining is that it's 12.01 pm. But their prior credence function assigned a credence of 0 to the latter centred proposition.

4 Two Updating Rules

Can we formulate updating rules that handle these simple cases? There are two salient options.

- *Extending the Formal Framework*. Let a self-locating frame $\langle S, E, \pi\rangle$ be a triple such that:
  - S is the set of all centred worlds $\langle w, i, t\rangle$ that are epistemically possible for an agent independently of any empirical evidence. For simplicity, we'll assume that it's finite.

- E is an evidence function that maps any centred world $\langle w, i, t\rangle$ to a centred proposition (i.e., a subset of S), which reflects i's total evidence in w at t.
- π is a regular ur-prior that is epsitemically rational for the agent to use.

- *Two Constraints on the Prior*.
  - THE CENTRED PRINCIPAL PRINCIPLE. For any centred world s in S and any centred proposition H, $\pi(H|[Chance_{now}(H) = r] \cap E(s)) = r$, provided that $E(s)$ doesn't contain any inadmissible evidence about H.
  - CENTRED INDIFFERENCE. For any centred world s in S, if two centred worlds s_1 and s_2 are compatible with $E(s)$, then $\pi(\{s_1\}|E(s)) = \pi(\{s_2\}|E(s))$.

- *A Few More Useful Notions*.
  - *Updating Plans*. Let an updating plan U be a function that maps any s in S to a probability function p.
  - *Non-Self-Locating Evidence*. For any centred world s in S, the non-self-locating evidence at s is $NSE(s) = \{\langle w, i, t\rangle \in S : (\exists w)(\langle w, i^*, t^*\rangle \in E(s))\}$.
  - *Non-Self-Locating Partition*. Let W be the finest partition of non-self-locating propositions that are subsets S.

- *The Central Question*. Should acquiring purely self-locating evidence without losing any non-self-locating evidence affect our doxastic attitudes towards non-self-locating propositions?
  - If you think the answer is "Yes," you'll like:

    UR-PRIOR CONDITIONALIZATION. Suppose an agent can be represented by a self-locating frame $\langle S, E, \pi\rangle$. Then, an updating plan U is epistemically rational for the agent to (plan to) comply with iff, for any centred world s in S,

    $$U(s) = \pi(.|E(s)).$$
  - If you think the answer is "No," you'll like a slightly more complicated rule.

    COMPARTMENTALIZED UR-PRIOR CONDITIONALIZATION.This rule is defended by Meacham (2008). Suppose an agent can be represented by a self-locating frame $\langle S, E, \pi\rangle$. Then, an updating plan U is epistemically rational for the agent to (plan to) comply with iff, for any centred world s in S,

    $$U(s) = \sum_{H \in W} \pi(.|H \cap E(s))\pi(H|NSE(s))).$$

- *Observations*.
  - Under some circumstances, e.g., when the agent's evidence is factive, introspective, and tells the agent who they are and what time it is, the predictions of the two rules will coincide if CENTRED INDIFFERENCE is true.
  - COMPARTMENTALIZED UR-PRIOR CONDITIONALIZATION coincides with Bayesian conditionalization for non-self-locating propositions.

– Under some circumstances, e.g., when the agent's evidence is factive, introspective, and tells the agent who they are and what time it is, UR-PRIOR CONDITIONALIZATION coincides with Bayesian conditionalization for non-self-locating propositions if CENTRED INDIFFERENCE is true.

- *A Failure of Reflection*.[7]

  > *Flashes*. Some scientists are going to put me into a state of dreamless sleep on Sunday night. No matter what happens, they are going to wake me up twice—once on Monday and once on Tuesday—erasing the memory of the previous awakening on the second occasion. On Monday, I will see a green flash. On Monday night, the scientists will flip a coin. If the coin lands heads, then I will see a red flash on Tuesdayg, but otherwise a green flash. On Sunday, I am given all this information, and I retain it throughout the process.

  Suppose I wake up on Monday and see a red flash. Should my credence that the outcome of the coin flip is heads (*Heads*) should decrease?

  – If you like COMPARTMENTALIZED UR-PRIOR CONDITIONALIZATION, you will say "No."
    * Suppose I don't have any inadmissible information about the outcome of the coin flip on Sunday. So, by the CENTRED PRINCIPAL PRINCIPLE, I will assign a credence of 0.5 to *Heads* on Sunday. But, at every stage on Monday, I'll only receive purely self-locating evidence.
    * COMPARTMENTALIZED UR-PRIOR CONDITIONALIZATION says that my doxastic attitudes towards non-self-locating propositions should remain unchanged when I only receive purely self-locating evidence without losing any non-self-locating evidence. So, my credence in *Heads* should remain fixed at 0.5.
  – If you like UR-PRIOR CONDITIONALIZATION, you will say "Yes."
    * For simplicity, assume that my initial evidence on Monday after waking up is $E_{mon} = \{\langle h, i, t_{mon}\rangle, \langle h, i, t_{tue}\rangle, \langle t, i, t_{mon}\rangle, \langle h, i, t_{tue}\rangle\}$. Initially, since my evidence doesn't contain any inadmissible information, by the PRINCIPAL PRINCIPLE, my credence in *Heads* should be $\pi(Heads|E_{mon}) = 0.5$.
    * By CENTRED INDIFFERENCE, I assign a credence of 0.25 to each centred world in E. So, if I were to then learn that I am not being woken up for the second time in a world where the coin lands heads, my total evidence would be $E^*_{mon} = E_{mon} \sim \{\langle h, i, t_{tue}\rangle$. So, my credence in *Heads* should be $\pi(Heads|E^*_{mon}) = \frac{1}{3}$.

  So, if I am rationally certain that you will update according to UR-PRIOR CONDITIONALIZATION, WEAK REFLECTION will fail.

[7] This is similar to the case discussed by Dorr (2002).

5 Sleeping Beauty

- *Another Example*.

  > *Sleeping Beauty*.[8] Some scientists are going to put me into a state of dreamless sleep on Sunday night. During the two days that my sleep will last, they will briefly wake me up either once or twice, depending on the toss of a fair coin. If the coin lands heads, they will wake me on Monday but not on Tuesday; if the coin lands tails, they will wake me on both Monday and Tuesday. After each waking, they will put me to back to sleep and erase my memory of that waking. Before I am put to sleep, I learn all this information, and retain it throughout the process.

  Suppose it's Monday morning and I've just woken up. How confident should I be that the coin landed heads? The *Halfer* says: it's $\frac{1}{2}$. The *Thirder* says: it's $\frac{1}{3}$.

- *The Thirder's Reasoning*. When I wake up on Monday, my total evidence is $E_{mon} = \{\langle h, i, t_{mon}\rangle, \langle t, i, t_{mon}\rangle, \langle t, i, t_{tue}\rangle\}$. So, if UR-PRIOR CONDITIONALIZATION is true, then my credence in *Heads* should be:

  $$\pi(Heads|E_{mon}) = \pi(Heads \cap Mon|E_{mon}) + \pi(Heads \cap \sim Mon|E_{mon}) = \pi(Heads|E_{mon} \cap Mon)\pi(Mon|E_{mon}).$$

  Suppose we assume that:

  - ASSUMPTION 1. $\pi(Heads|E_{mon} \cap Mon) = 0.5$ (by CENTRED PRINCIPAL PRINCIPLE).
  - ASSUMPTION 2. $\pi(Mon|E_{mon} \cap \sim Heads) = 0.5$. (by CENTRED INDIFFERENCE)

  Then, $\pi(Heads|E_{mon}) = \frac{1}{3}$.

- *The Halfer's Reasoning*. There are two ways of blocking this reasoning.

  - The Single Halfer accepts UR-PRIOR CONDITIONALIZATION, but rejects ASSUMPTION 1. Rather, they claim that $\pi(Heads|E_{mon} \cap Mon) = \frac{2}{3}$.
  - The Double Halfer could reject UR-PRIOR CONDITIONALIZATION.

6 Problem Cases

- *Self-Manipulation*.

  > *Sleeping Beauty Redux*. I know who I am, and want to find out whether I am popular. On Sunday, I have a credence of 0.5 that I'm popular. But I want to be confident on Monday that I'm popular. I ask my friend to put me into a state of dreamless sleep on Sunday evening, and then find out whether I am popular. During the two days that my sleep will

[8] See Elga (2000).

last, she will briefly wake me either once or twice, depending on the outcome of her investigation. If I am not popular, she will wake me on Monday but not on Tuesday; if I am popular, she will wake me up on both Monday and Tuesday. After each waking, my friend will put me to back to sleep with a drug that makes me forget that waking. Before I am put to sleep, I learn all this, and retain this information throughout the process.

- *Chancy Awakening*. The thirder also cannot handle a generalized version of *Sleeping Beauty* (discussed by Roger White 2006).

  > *Generalized Sleeping Beauty*. Some scientists are going to put me into a state of dreamless sleep on Sunday night. During the two days that my sleep will last, they will activate either once or twice a random waking device that has an adjustable chance $c \in (0, 1]$ of waking me when activated on any occasion. If the coin lands heads, they will activate the device on Monday but not on Tuesday; if the coin lands tails, they will activate the device on both Monday and Tuesday. After each waking, they will put me to back to sleep with a drug that makes me forget that waking. Before I am put to sleep, I learn all this information as well as the value of c and retain it throughout the process.

- *A Dutch Book*. Suppose the Single Halfer is right. Then, the following principle fails.

  > THE PRINCIPLE OF EXPLOITABILITY. An epistemically and instrumentally rational agent—who acts only when rationally certain of who they are and what the time is—cannot be predictably exploitable unless they undergo evidence loss or lack perfect access to their own evidence or has false evidence.

  Suppose my credence in *Heads* goes from being 0.5 to being $\frac{2}{3}$ on Monday (after I am told it's Monday).

  - Suppose I am offered a bet that pays 1 if the coin lands tails and nothing otherwise at the maximum price I consider fair. Then, I will buy it by paying 0.5.
  - Suppose, on Monday after I have been told that it's Monday, the bookie buys this bet back at the minimum price I consider fair. Then, I will buy it by taking the payment of $\frac{1}{3}$.

  As a result, I will undergo net loss.

- *Symmetry of Relevance*. The Double Halfers cannot preserve the following principle in *Flashes* and *Sleeping Beauty*.

  > SYMMETRY OF RELEVANCE. Suppose an agent's total evidence E is such that it is rational for her to assign non-zero credence to both P and Q

relative to E. If learning only P makes it rational for the agent to increase her credence in Q, then learning only Q should make it rational for the agent to increase her credence in P.

- *Violations of the Principal Principle.* The Double Halfers cannot be preserve the CENTRED PRINCIPAL PRINCIPLE in the following case due to Titelbaum (2012).

  > Suppose the experimenters enjoy flipping the coin so much that they decide to flip it once more on Tuesday night. The coin remains fair, the Tuesday flip has no impact on anything having to do with Beauty (it's just an idle coin flip), and the Tuesday flip will be performed whether Beauty awakens that day or not. Assume also that Beauty is informed on Sunday night that the extra Tuesday flip will occur. When Beauty awakens Monday morning (uncertain what day it is), how confident should she be in the proposition that *today*'s coin flip comes up heads?

  The Double Halfer must say that the answer is 0.625, which seems wrong.

7 Appendix

- *Some Definitions*.
  - Let a self-locating frame $\langle S, E, \pi \rangle$ be partitional iff, for any s in S, $s \in E(s)$, and, if $s^* \in E(s)$, $E(s) = E(s^*)$.
  - Let a self-locating frame $\langle S, E, \pi \rangle$ be self-aware iff, if $s^* \in E(s)$, then s and s^* have the same individual and time coordinates.
  - For any self-locating frame $\langle S, E, \pi \rangle$, let an updating rule U be Bayesian iff, for any s, s^* in S, if $NSE(s) \subseteq NSE(s^*)$ and $U(s) = p$ and $U(s^*) = p^*$, then, for any non-self-locating proposition H, $p^*(H) = p(H|NSE(s^*))$.
- *Two Claims*.
  - PROPOSITION 1. For any partitional and self-aware self-locating frame $\langle S, E, \pi \rangle$, if π satisfies CENTRED INDIFFERENCE, then an updating plan U satisfies UR-PRIOR CONDITIONALIZATION iff it satisfies COMPARTMENTALIZED UR-PRIOR CONDITIONALIZATION.
  - PROPOSITION 2. For any partitional self-locating frame $\langle S, E, \pi \rangle$,
    - If U satisfies COMPARTMENTALIZED UR-PRIOR CONDITIONALIZATION, then it is Bayesian.
    - If the frame is self-aware, π satisfies CENTRED INDIFFERENCE, and U satisfies UR-PRIOR CONDITIONALIZATION, then U is Bayesian.

References

Dorr, Cian. 2002. Sleeping Beauty: In Defence of Elga. *Analysis*, **62**(4), 292–296.

Elga, Adam. 2000. Self-Locating Belief and the Sleeping Beauty Problem. *Analysis*, **60**(2), 143–147.

Lewis, David. 1979. Attitudes de Dicto and de Se. *Philosophical Review*, **88**(4), 513–543.

Meacham, Christopher J.G. 2008. Sleeping Beauty and the Dynamics of de Se Beliefs. *Philosophical Studies*, **138**(2), 245–269.

Ninan, Dilip. 2016. What is the Problem of De Se Attitudes? *In:* Torre, Stephan, & Garcia-Carpintero, Manuel (eds), *About Oneself: De Se Thought and Communication*. Oxford University Press.

Perry, John. 1979. The Problem of the Essential Indexical. *Noûs*, **13**(1), 3–21.

Titelbaum, MichaelG. 2012. An Embarrassment for Double-Halfers. *Thought: A Journal of Philosophy*, **1**(2), 146–151.

White, Roger. 2006. The Generalized Sleeping Beauty Problem: A Challenge for Thirders. *Analysis*, **66**(2), 114–119.