Defeating Dr. Evil with self-locating belief

Adam Elga

Penultimate draft
Forthcoming, *Philosophy and Phenomenological Research*

Abstract

Dr. Evil learns that a duplicate of Dr. Evil has been created. Upon learning this, how seriously should he take the hypothesis that he himself *is* that duplicate? I answer: very seriously. I defend a principle of indifference for self-locating belief which entails that after Dr. Evil learns that a duplicate has been created, he ought to have exactly the same degree of belief that he is Dr. Evil as that he is the duplicate. More generally, the principle shows that there is a sharp distinction between ordinary skeptical hypotheses, and *self-locating* skeptical hypotheses.

1

Safe in an impregnable battlestation on the moon, Dr. Evil had planned to launch a bomb that would destroy the Earth. In response, the Philosophy Defense Force (PDF) sent Dr. Evil the following message:

> Dear Sir,
>
> (Forgive the impersonal nature of this communication—our purpose prevents us from addressing you by name.) We have just created a duplicate of Dr. Evil. The duplicate—call him "Dup"—is inhabiting a replica of Dr. Evil's battlestation that we have installed in our skepticism lab. At each moment Dup has experiences indistinguishable from those of Dr. Evil. For example, at this moment both Dr. Evil and Dup are reading this message.
>
> We are in control of Dup's environment. If in the next ten minutes Dup performs actions that correspond to deactivating the battlestation and surrendering, we will treat him well. Otherwise we will torture him.
>
> Best regards,
> The PDF

Dr. Evil knows that the PDF never issues false or misleading messages. Should he surrender?

Whether Dr. Evil ought to surrender depends on whether he ought to remain certain that he is Dr. Evil once he reads the message. He might reason in this way:

I'm sure that Dr. Evil and Dup are in the same subjective state—

> namely, the state I am in right now. So I ought to be unsure whether I am Dr. Evil or Dup. If I'm Dup and I don't surrender, I'll be tortured. And Dr. Evil's evil plans are not worth the risk of torture. So I ought to surrender.

Alternatively, before receiving the message Dr. Evil might have reasoned in this way:

> I'm sure that I'm Dr. Evil and that my battlestation is impregnable. True, the PDF can create duplicates of me. So what? I don't care what happens to any duplicates that those pointyheads should be fool enough to create. If they do make a duplicate of me, I ought to remain certain that I'm Dr. Evil. Of course, the duplicate *also* ought to be certain that he's Dr. Evil. Tough luck for him, but he's in no worse shape than various other deluded inhabitants of the skepticism lab.

I will argue that only the first of the two lines of reasoning is correct, and that after Dr. Evil receives the message he ought to have exactly the same degree of belief that he is Dr. Evil as that he is Dup. My plan is to use the Dr. Evil case to motivate a principle of indifference for self-locating belief, a principle that brings out a stark contrast between two kinds of skeptical hypotheses.

2

What Dr. Evil ought to do depends on two factors: what his beliefs ought to be, and how his beliefs and values together ought to guide his action. Let's isolate the first factor by considering some stripped-down cases.

Start with the case of O'Leary:

O'LEARY ([5]) O'Leary is locked in the trunk of his car overnight. He knows that he'll wake up briefly twice during the night (at 1:00 and again at 2:00) and that the awakenings will be subjectively indistinguishable (because by 2:00 he'll have forgotten the 1:00 awakening). At 1:00 he wakes up.

When O'Leary wakes up at 1:00, he ought to be uncertain whether it is 1:00 or 2:00. That's because he is certain that he awakens both at 1:00 and at 2:00, and certain that his present experience doesn't distinguish between the two awakenings. I hear an objection:

TRUST YOUR MEMORIES, O'LEARY When O'Leary awakens, he doesn't remember any previous awakenings. Furthermore, he ought to trust his memories. So he ought to be very confident that there *weren't* any previous awakenings, and hence very confident that it is 1:00.

This objection is mistaken. It is true that, in the absence of defeating auxiliary beliefs, one ought to trust one's memories.[1] For example, you ought

[1]For convenience I use "memory" in a non-factive sense, according to which misleading "memories" still count as memories.

to trust your memory that you haven't been tickled in the last five minutes. But it is also true that, in the presence of the right auxiliary beliefs, one ought *not* trust certain memories much at all. For example, you might be sure that tickling makes very little of an impression on you, and that even if you *had* been tickled in the last five minutes you would have forgotten by now. In that case the fact that you don't remember any recent tickling would give you no reason at all to think that the last five minutes has been tickle-free.

Let me state that in a slightly different way. Right after he is locked in his trunk, O'Leary is certain just what predicament he is in: locked in a trunk at 9:00. But when he wakes up at 1:00, he no longer is certain what predicament he is in, because he is uncertain whether it is 1:00 or 2:00. The crucial difference between O'Leary's beliefs at 9:00 and his beliefs at 1:00 is this:

- At 9:00, O'Leary is confident that his current subjective state is only instantiated *once* throughout all of history: by O'Leary at 9:00.

- In contrast, at 1:00 O'Leary is confident that his current subjective state is instantiated *twice* throughout history: first by O'Leary at 1:00, and again by O'Leary at 2:00.

When he wakes up at 1:00, O'Leary remembers no previous awakening. What ought to defeat his trust in this aspect of his memory is his confidence that his present subjective state (memories and all) is instantiated twice throughout history (by O'Leary at 1:00, and again by O'Leary

at 2:00), and that in one of those instantiations (namely, the 2:00 instantiation), the memories are misleading.

I'd like to consider the Dr. Evil case on the model of the O'Leary case. Before reading the PDF's message, Dr. Evil is confident that his current subjective state is instantiated once. But after reading the message, Dr. Evil is confident that his current subjective state is instantiated twice (once by Dr. Evil and once by Dup). On that basis Dr. Evil ought to be uncertain which of these two instantiations he is currently experiencing. In other words, he ought to be uncertain whether he is Dr. Evil or Dup.

Put yourself in Dr. Evil's shoes: You are sure that Dr. Evil and Dup are in the same subjective state—the one you are in right now. You are sure that both have memories indicating that they are Dr. Evil—for example, both remember growing up as a young evil mastermind. And you are sure that the memories of Dup are in this respect utterly misleading.[2] On that basis, when your memories indicate that you are Dr. Evil, you ought not trust them.

[2]Here I am assuming that Dr. Evil is confident that duplication technology exists, but not technology that allows for the fabrication of memories from scratch. That assumption makes it reasonable for Dr. Evil to trust his memories in one respect but not another. He is certain that his memories are accurate first-person records of what happened to Dr. Evil. But he is uncertain whether he was the one who experienced the remembered events.

This kind of partial trust in one's memories is unusual, but it can be reasonable. For example, you might be confident that you had just received some transplanted memories from someone else's childhood, including the memory of holding a beloved one-eyed teddy bear. In such a case, you should be confident that *someone* once held a one-eyed teddy bear (namely, the person from whom you got the implanted memories). But you should also be confident that *you* never held such a bear, even though you remember doing so. For further discussion of such cases, see [3].

3

I hope I've made it plausible that after Dr. Evil reads the message, he ought to be uncertain whether he is Dr. Evil or Dup. But belief is not an all-or-nothing matter: it comes in degrees. For example, you might be uncertain whether the Yankees or the Dodgers won the 1955 world series, but still have a greater degree of belief that the Yankees won than that the Dodgers won. So there remains the question of how exactly Dr. Evil ought to apportion his confidence—or *credence*—between the hypothesis that he is Dr. Evil and the hypothesis that he is Dup.

There's a more general question lurking in the vicinity. Many times an agent is sure that her current subjective state is instantiated only once throughout history: by *her*, *then*. But sometimes an agent has some confidence that her current subjective state is instantiated several times throughout history. This can happen (as it does in the O'Leary case) if the agent is sure that she instantiates her current subjective state at more than one time. It can also happen (as it does in the Dr. Evil case) if the agent is sure that more than one person currently instantiates that state.[3]

The cases described so far were intended to make it plausible that when an agent is sure that her current subjective state is instantiated several times, she ought to be uncertain which instantiation she is currently experiencing. But since belief comes in degrees, there remains the question

[3]More complicated hybrid cases are possible as well, but such cases won't figure in the following discussion.

of how such an agent ought to apportion her confidence among the various hypotheses about which instantiation she is currently experiencing.

4

I'm going to defend an indifference principle for self-locating belief. To state the principle I'll need the notion of a *centered world* (see [2]). A centered world is a possible world which has been equipped with a designated individual and time.[4] Just as a possible world represents a maximally specific *way for the world to be*, a centered world represents a maximally specific *predicament for one to be in*. A centered world represents the maximally specific predicament of being *in* the associated possible world *as* the designated individual *at* the designated time.

When an agent is uncertain about which of various possibilities obtain, she divides her credence among several possible worlds. Similarly, when an agent is uncertain about which of various predicaments she's in, she divides her credence among several centered worlds.[5]

Call two centered worlds X and Y *similar* iff the following conditions are both satisfied:

- X and Y are associated with the same possible world. (In other words, they differ at most on who is the designated individual or

[4]Formally, one can take a centered world to be an ordered triple of the form $\langle w, i, t \rangle$, where w is a possible world, i is an individual who exists at w, and t is a time.

[5]Here I follow a terminological variant of a suggestion in [2].

what is the designated time.)

- X and Y represent predicaments that are subjectively indistinguishable. (In other words, the designated individuals are—at the designated times—in subjectively indistinguishable states. For example, the designated individuals have the same apparent memories and are undergoing experiences that feel just the same.)

A distinctive feature of the cases described so far is that they involve agents who divide their credence between similar centered worlds. (From now on, when the context prevents ambiguity, I sometimes shall use "world" to mean "centered world".) For example, when O'Leary wakes up, he divides his credence in part between a pair of similar worlds: one centered on O'Leary at 1:00 and another centered on O'Leary at 2:00. And after Dr. Evil reads the message, he divides his credence in part between a pair of similar worlds: one centered on Dr. Evil and one centered on Dup.

I'm going to defend the following claim:

INDIFFERENCE Similar centered worlds deserve equal credence.

This claim entails that O'Leary ought to have exactly the same degree of belief that it is 1:00 as that it is 2:00. It also entails that Dr. Evil ought to have exactly the same degree of belief that he is Dr. Evil as that he is Dup.

Disclaimer: There is a big difference between INDIFFERENCE and the following much stronger claim:

ABSURD-CLAIM-THAT-I-DON'T-ENDORSE Centered worlds representing indistinguishable predicaments deserve equal credence.

This stronger claim is absurd. For example, let *AT* be the actual world, centered on you, now. Let *VAT* be a world centered on a brain in a vat who is in a state subjectively indistinguishable from yours. ABSURD-CLAIM-THAT-I-DON'T-ENDORSE entails that you ought to assign *AT* and *VAT* equal credence. That's absurd. In contrast, INDIFFERENCE entails nothing of the sort, since *AT* and *VAT* are *not* similar—they are associated with *different* possible worlds.[6]

5

My defense of INDIFFERENCE has two parts. I'll describe a basic case involving an agent who divides his credence between a pair of similar centered worlds, and argue that he ought to assign those worlds equal credence. And I'll say how that argument can be generalized. But since the controversial parts of the argument arise even in the basic case, I relegate the generalization to the Appendix.

Here is the basic case, which concerns someone named "Al":

DUPLICATION After Al goes to sleep researchers create a duplicate of him in a duplicate environment. The next morning, Al and the duplicate awaken in subjectively indistinguishable states.

[6]For clarification on this point I am indebted to Earl Conee and Juan Comesaña.

In this and all subsequent cases, assume that Al knows in advance exactly what the experimental protocol will be.

Simplify the case by assuming that when Al wakes up, he is only uncertain about one thing: whether he is Al or the duplicate. Thus he entirely divides his credence between a pair of similar worlds: one centered on Al, and one centered on the duplicate. I will argue that he ought to divide his credence evenly between these worlds. In doing so, I'll be defending a special case of INDIFFERENCE.

The first step is to add one additional bit of uncertainty to Al's situation: the outcome of a coin toss. The result is the following slightly more complicated case:

TOSS&DUPLICATION After Al goes to sleep, researchers toss a coin that has a 10% chance of landing heads. Then (regardless of the toss outcome) they duplicate Al. The next morning, Al and the duplicate awaken in subjectively indistinguishable states.

In TOSS&DUPLICATION, the coin toss is irrelevant to whether and how the duplication occurs. So Al's state of opinion (when he awakens) as to whether he is Al or the duplicate ought to be the same in TOSS&DUPLICATION as it is in DUPLICATION. So in order to show that in DUPLICATION, Al ought to divide his credence evenly between the hypothesis that he is Al and the hypothesis that he is the duplicate, it is enough to show that he ought to do so in TOSS&DUPLICATION. (Why shift attention to the more complicated case? Because the method of analysis requires it, as will become clear

shortly.)

When Al wakes up in TOSS&DUPLICATION he divides his credence between four centered worlds: HeadsAl (a world in which the coin lands heads, centered on Al), HeadsDup (in which the coin lands heads, and which is centered on the duplicate), TailsAl, and TailsDup (notation similar; see Figure 1).

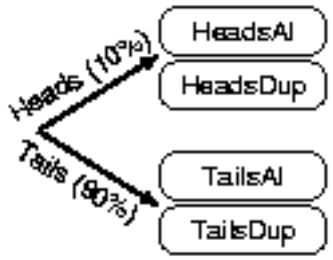

Figure 1: *The TOSS&DUPLICATION protocol, in which a biased coin is tossed and a duplicate of Al is created regardless of the outcome. When Al awakens, he wonders which of four predicaments—represented by ovals—he is in.*

My argument depends on three claims about Al's degrees of belief when he awakens. Since the first two are uncontroversial, I'll start with them.

Let P be the credence function Al ought to have when he first awakens, and let HEADS be the proposition that the toss outcome is heads.

First note that when Al awakens, his credence in HEADS ought to be 10%:

(1) $P(\text{HEADS}) = 10\%$.

That's because he knew before he went to sleep that the coin had a chance of 10% of landing heads, and because upon awakening he neither gains nor loses information relevant to the toss outcome.[7]

Now note that the duplication happens in exactly the same manner whether the coin lands heads or tails. So Al should count the toss outcome as irrelevant to who he is. In other words, conditionalizing on the information that he is Al (in other words, the information [HeadsAl or TailsAl]) ought to leave Al's opinion of the toss outcome unchanged:

(2) $P(\text{HEADS} \,|\, \text{HeadsAl or TailsAl}) = 10\%$.

The third claim is the controversial one. It is that Al's credence in HEADS, given [HeadsAl or TailsDup], ought also to be 10%. (Roughly, to receive the information [HeadsAl or TailsDup] is to be put in a position to assert with certainty: "Either the coin landed heads and I am Al, or else the coin landed tails and I am the duplicate.")

(3) $P(\text{HEADS} \,|\, \text{HeadsAl or TailsDup}) = 10\%$.

From (1), (2), and (3) it follows that when Al awakens in TOSS&DUPLICATION he ought to have the same credence that he is Al as that he is the duplicate.[8] Since Al's credence in HEADS ought to be the same in DUPLICATION

[7]Here I've depended on a principle that connects one's beliefs about chances with one's beliefs about the outcomes of chance processes. For more on such principles, see [1].

[8]Proof: Let H and T abbreviate HEADS and TAILS respectively, and use acronyms to abbreviate the names of the four centered worlds (for example, "HA" denotes HeadsAl). We must use (1)-(3) to show that when Al awakens, he ought to have the same credence

as it is in TOSS&DUPLICATION, it further follows that when Al awakens in DUPLICATION, he ought to have the same credence that he is Al as that he is the duplicate. That's just the special case of INDIFFERENCE that was to be argued for. So to complete the argument for that special case of INDIFFERENCE, it remains only to defend (3).

6

This section defends (3), the claim that when Al awakens in TOSS&DUPLICATE, his credence in HEADS, given [HeadsAl or TailsDup], ought to be 10%. To check (3), consider what degrees of belief Al ought to end up with if he were to awaken and then find out [HeadsAl or TailsDup]. To do that, consider the following variant of TOSS&DUPLICATE:

COMA As in TOSS&DUPLICATE, the experimenters toss a coin and duplicate Al. But the following morning, the experimenters ensure that *only*

---

that he is Al as that he is the duplicate. In other words, we must show that $P(HA \text{ or } TA) = P(HD \text{ or } TD)$.

To show this, set the left hand sides of (2) and (3) equal to each other: $P(H|HA \text{ or } TA) = P(H|HA \text{ or } TD)$. By the definition of conditional probability, $P(HA)/P(HA \text{ or } TA) = P(HA)/P(HA \text{ or } TD)$. Some algebra then gets us that $P(HA \text{ or } TA) = P(HA \text{ or } TD)$. Since HA, TA, and TD are all disjoint, $P(TA) = P(TD)$. Since $P(TA)$ and $P(TD)$ add up to $P(T)$, each equals $P(T)/2$.

Now set the left hand sides of (1) and (2) equal to each other: $P(H) = P(HA|HA \text{ or } TA)$. It follows that $P(T) = P(TA|HA \text{ or } TA)$. Divide the first equation by the second to get: $P(H)/P(T) = P(HA|HA \text{ or } TA)/P(TA|HA \text{ or } TA)$. By the definition of conditional probability, $P(H)/P(T) = P(HA)/P(TA)$. Rearranging we get that $P(HA) = P(H)P(TA)/P(T)$, which in turn equals $P(H)/2$ since $P(TA) = P(T)/2$. So $P(HD) = P(HA)$, since they add up to $P(H)$. Combining this with the fact that $P(TA) = P(TD)$, we have that $P(HA \text{ or } TA) = P(HD \text{ or } TD)$, as was to be shown.

one person wakes up: If the coin lands heads, they allow Al to wake up (and put the duplicate into a coma); if the coin lands tails, they allow the duplicate to wake up (and put Al into a coma). (See Figure 2.)

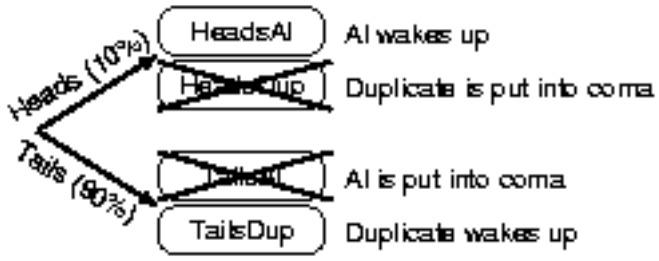

Figure 2: *The COMA protocol, in which a biased coin is tossed to determine whether Al or the duplicate is allowed to awaken.*

Suppose that in the COMA case, Al gets lucky: the coin lands heads, and so the experimenters allow him to awaken. Upon awakening, Al is immediately in a position to assert "Either I am Al and the coin landed heads, or else I am the duplicate and the coin landed tails". So when Al wakes up in the COMA case, he has just the evidence about the coin toss as he would have if he had been awakened in TOSS&DUPLICATE *and then been told* [HeadsAl or TailsDup]. So to defend (3)—to show that in the latter case Al's credence in HEADS ought to be 10%—it is enough to show that when Al wakes up in the COMA case, his credence in HEADS ought to be 10%.[9] Let me argue for that claim now.

[9]I think it is plausible that this connection holds—that Al's credence in HEADS in COMA ought to match his conditional credence in HEADS in TOSS&DUPLICATE. But if

Before Al was put to sleep, he was sure that the *chance* of the coin landing heads was 10%, and his credence in HEADS should have accorded with this chance: it too should have been 10%. When he wakes up, his epistemic situation with respect to the coin is just the same as it was before he went to sleep. He has neither gained nor lost information relevant to the toss outcome. So his degree of belief in HEADS should continue to accord with the chance of HEADS at the time of the toss. In other words, his degree of belief in HEADS should continue to be 10%.

I hear an objection:

TRUST YOUR MEMORIES, AL When Al awakens, he has memories indicating

---

you doubt that it does, consider the following intermediate case:

DELAYED COMA The experimenters follow the TOSS&DUPLICATE protocol as described above (they toss a coin and duplicate Al). Then, five minutes after Al and the duplicate wake up the following morning, the experimenters use the toss outcome to determine which one gets to stay awake: if heads, they allow Al to stay awake (and put the duplicate into a coma); if tails, they allow the duplicate to stay awake (and put Al into a coma).

In DELAYED COMA, if Al were to wake up and remain awake for more than five minutes, that would put him in a position to assert "Either I am Al and the coin landed heads, or else I am the duplicate and the coin landed tails". He'd be in a position to assert this because under the DELAYED COMA protocol, if the coin lands heads, only Al gets to stay awake, but if the coin lands tails, only the duplicate does. So to defend (3), it is enough to show that after being so informed, Al should end up with credence 10% in HEADS.

In DELAYED COMA, there is a five minute delay between when Al and the duplicate wake up, and when the experimenters put one of them into a coma. Reducing or even eliminating that delay leaves the case in relevant respects the same. For example, if the delay were thirty seconds (rather than five minutes), then Al would only have to wait thirty seconds in order to be sure that he wasn't selected for the coma. But (supposing that he wasn't so selected), his final credence in HEADS should be the same in the thirty second delay case as it is in the five minute delay case. And it also should be the same in the limiting case, in which there is no delay at all. But we've seen that case: it is COMA.

So to defend (3) it is enough to show that in COMA, when Al wakes up he ought to have credence 10% in HEADS.

that he is Al (for example, he remembers being Al as a child). He ought to trust those memories. So he ought to be confident that he *is* Al, and hence confident that the coin landed heads.

The night before the coin toss, Al had a hard time getting to sleep because he knew that if the coin landed tails he would be put into a coma. If TRUST YOUR MEMORIES, AL is right, then when Al wakes up the next morning he ought to announce with relief, "I made it!" Indeed, if TRUST YOUR MEMORIES, AL is right, Al ought to realize in advance that whatever happens, the person who wakes up the next morning ought to wake up confident that he is Al, and that the coin landed heads.

That's all wrong. TRUST YOUR MEMORIES, AL makes the same mistake that TRUST YOUR MEMORIES, O'LEARY does. While it is true that in the absence of defeating auxiliary beliefs, one ought to trust one's memories, when Al wakes up he *does* have defeating auxiliary beliefs. He is sure that—whatever the outcome of the coin toss—*someone* was to wake up in just the subjective state he is currently in. As far the outcome of the coin toss goes, the total evidence Al has when he wakes up warrants exactly the same opinions as the total evidence he had when he went to sleep.[10]

The bottom line: In COMA, when Al awakens he should have credence 10% in HEADS. We saw above that this claim is sufficient to defend (3). So (3) is true.

[10]Here I have also set myself against an intermediate view, according to which Al's beliefs about the setup only *partially* undermine his memories of being Al. According to such a view, when Al wakes up his credence in HEADS ought to be slightly higher than 10%. I owe this point to Jim Pryor.

Remember that it follows from (1)-(3) that when Al awakens in TOSS&DUPLICATION, he ought to have the same credence that he is Al as that he is the duplicate. From this it follows that the same holds when Al awakens in DUPLICATION. And that's the special case of INDIFFERENCE that was to be argued for.

Furthermore, the special case isn't so special, since the same argument can be applied to any case in which an agent divides her credence among similar worlds (see the Appendix for details). So INDIFFERENCE is true.

7

Hartry Field used to tell a story in which a maintenance man, dusting off some brains in vats, notices that the brains are all being fed sensory stimulation from a program labeled MAINTENANCE MAN DUSTS BRAINS IN VATS. Field presented the story as a case in which it is reasonable to wonder whether one is a brain in a vat. Provided that the details of the story are filled in properly (if the maintenance man is confident that some of the brains are in states subjectively indistinguishable from his own), INDIFFERENCE vindicates this judgment. Indeed, if there are very many such brains, then INDIFFERENCE entails that the maintenance man ought to be confident that he is such a brain.

There is a stark contrast between two sorts of skeptical hypotheses. According to one sort of hypothesis, your sensory evidence has misled you into believing many false propositions. (Example: the hypothesis that

nothing exists except a single deluded brain in a vat.)

A different sort of skeptical hypothesis arises only if you are confident that your current subjective state is instantiated more than once. This second sort of hypothesis conflicts with none of the propositions that you believe.[11] But according to such a hypothesis, your *location* within the world is different than you'd thought. When Field's maintenance man wonders whether he is a brain in a vat, he is considering a hypothesis of this second sort.

There is a default bias against the first sort of skeptical hypothesis. Even if one's total evidence fails to rule such a hypothesis out, one is reasonable in having very little credence in it. In contrast, INDIFFERENCE entails that there is no corresponding bias against the second sort of hypothesis.

8

Dr. Evil, recall, received a message that Dr. Evil had been duplicated and that the duplicate ("Dup") would be tortured unless Dup surrendered. INDIFFERENCE entails that Dr. Evil ought to have the same degree of belief that he is Dr. Evil as that he is Dup. I conclude that Dr. Evil ought to

[11] This sort of hypothesis does conflict with some of your self-locating beliefs. But I take it that those beliefs are not beliefs in propositions, since propositions have no indexical component, and are timelessly true or false. On this way of thinking, one might take propositions to be sets of possible worlds. Alternatively, one might take them to be those contents expressible by eternal sentences ([4]:191) of some appropriately rich language.

surrender to avoid the risk of torture.

I am not entirely comfortable with that conclusion. For if INDIFFERENCE is right, then Dr. Evil could have protected himself against the PDF's plan by (in advance) installing hundreds of brains in vats in his battlestation—each brain in a subjective state matching his own, and each subject to torture if it should ever surrender. (If he had done so, then upon receiving the PDF's message he ought to be confident that he is one of those brains, and hence ought not surrender.)[12] Of course the PDF could have preempted this protection by creating *thousands* of such brains in vats, each subject to torture if it *failed* to surrender at the appropriate time. But Dr. Evil could have created *millions*...

It makes me uncomfortable to think that the fate of the Earth should depend on this kind of brain race.[13]

[12]Suppose that Dr. Evil creates many brains in vats, each in a subjective state matching his own. I say that it is then reasonable for him to wonder whether he is one of those brains. Objection: Dr. Evil's reason for thinking that there are lots of brains is that he remembers creating them. If he trusts that memory, then he should also trust that he *isn't* one of the brains. (For this objection I thank an anonymous referee, whose comments I have paraphrased.) Reply: it is reasonable for Dr. Evil to trust his memories in one respect, but not another (as is discussed in note 2). He ought to trust what they entail about what happened to Dr. Evil. But he ought not to trust that he is Dr. Evil.

[13]Thanks to Ned Hall, Robert Stalnaker, Tyler Doggett, Jim Pryor, Nick Bostrom, Cian Dorr, Patrick Hawley, Alex Byrne, Agustín Rayo, Anthony Newman, Andy Egan, Hartry Field, Jason Grossman, Daniel Nolan, Roger White, Miguel Hernando, and the M.I.T. M.A.T.T.I. group. Thanks also to audiences at Brown University, N.Y.U., the 2001 APA Pacific Division Meeting, and the 2001 Mind/Aristotelian Society Joint Session. I owe the term "brain race" to Stephanie Lewis.

Appendix

INDIFFERENCE is the claim that similar centered worlds deserve equal credence. In sections 5-6, I defended the special case of INDIFFERENCE that arises in DUPLICATION. Call the argument in those sections "the toss argument". This appendix sketches how the toss argument can be adapted to defend arbitrary instances of INDIFFERENCE.

Remember that in DUPLICATION, Al is certain that he has been duplicated, and is uncertain about only one matter: whether he himself is Al or the duplicate. As a result, he *entirely* divides his credence between a pair of similar worlds. That feature made the toss argument easier to explain, but was inessential. For example, consider a case similar to DUPLICATION, but in which Al is uncertain about many matters in addition to whether he is Al or the duplicate. In that case, Al divides his credence among many centered worlds. Select any two of these that are similar. Call them A (centered on Al) and D (centered on the duplicate).

INDIFFERENCE says that A and D deserve equal credence. It would be nice to directly apply the toss argument to show that they do. It would be nice, but it wouldn't work. That's because the toss argument assumes that Al *entirely* divides his credence between a pair of similar worlds, and because in the present case, he doesn't. (In the present case, Al does allocate credence to A and D—a pair of similar worlds. But he *also* allocates credence to many other worlds.)

Fortunately, there is a way for us to apply the toss argument indirectly.

The trick is to concentrate on Al's conditional degrees of belief. Consider what Al believes, *given* [A or D]. Conditional on [A or D], Al *does* divide his credence entirely between A and D. (That's an easy consequence of the definition of conditional probability.) So we can apply the toss argument to Al's degrees of belief conditional on [A or D].[14] The conclusion will be that, *conditional on* [A or D], Al ought to assign equal credence to A and D.

Furthermore, from this claim about Al's *conditional* credences, it fol-

[14]Here is how to apply the toss argument to Al's degrees of belief conditional on [A or D]. The first step is to add an additional bit of uncertainty to Al's situation: the outcome of a coin toss. Call the situation we started with "DUPLICATION+", and the resulting slightly more complicated situation "TOSS&DUPLICATION+".

In section 5 we examined Al's credences in TOSS&DUPLICATION in order to figure out what his credences ought to be in DUPLICATION. Similarly, we can examine Al's *conditional* credences in TOSS&DUPLICATION+ to figure out what his *conditional* credences ought to be in DUPLICATION+.

In TOSS&DUPLICATION, Al is only uncertain about two matters: how the coin landed, and who he is. As a result, he entirely divides his credence between four centered worlds. In TOSS&DUPLICATION+, Al is uncertain about other matters in addition to how the coin landed and who he is. So he divides his credence among many centered worlds. But for our purpose, we can restrict attention to four of these. The four relevant worlds are the ones gotten from A and D by adding the outcome of the coin toss: call them HA, HD, TA, TD. (For example, HA is like A, except for the addition of a coin toss whose outcome is heads.)

Restrict attention to what Al's credences in TOSS&DUPLICATION+ ought to be, conditional on [HA or HD or TA or TD]. Let $P+$ be the conditional credence function he ought to have. Now we can argue for analogues of (1), (2), and (3):

(1+) $P+(\text{HEADS}) = 10\%$.

(2+) $P+(\text{HEADS}|HA \text{ or } TA) = 10\%$.

(3+) $P+(\text{HEADS}|HA \text{ or } TD) = 10\%$.

It follows from (1+), (2+), and (3+) that conditional on [HA or HD or TA or TD], Al ought to have the same credence that he is Al as that he is the duplicate. Since the coin toss is irrelevant to the duplication, this shows that in DUPLICATION+, conditional on [A or D], Al ought to have equal credence that he is Al as that he is the duplicate. In other words, in DUPLICATION, conditional on [A or D] Al ought to assign equal credence to A and D.

lows that Al ought to assign equal *un*conditional credence to A and D.[15] And that's exactly the conclusion needed in order to defend INDIFFERENCE in this more general case.

DUPLICATION is a special case in several other respects. For example, Al is uncertain about who he is. But in general, an agent might be uncertain not about who he is, but rather about what time it is. (O'Leary, who awakens in his car trunk wondering whether it is 1:00 or 2:00, is such an agent.) Or an agent might be uncertain about who he is *and* what time it is. INDIFFERENCE entails that in all of these cases, similar centered worlds deserve equal credence.

But again, these special features of DUPLICATION were not essential to the toss argument. For example, one can use exactly analogous arguments to show that when O'Leary awakens, he should have the same degree of belief that it is 1:00 as that it is 2:00. So INDIFFERENCE is true quite generally.

References

[1] David Lewis. A subjectivist's guide to objective chance. In David Lewis, editor, *Philosophical Papers*, volume 2. Oxford University Press, 1982.

[2] David Lewis. Attitudes *de dicto* and *de se*. In David Lewis, editor, *Philosophical Papers*, volume 1. Oxford University Press, 1983.

[15] Proof: Assume that $P(A|A \text{ or } D) = P(D|A \text{ or } D)$. Then by the definition of conditional probability, $P(A)/P(A \text{ or } D) = P(D)/P(A \text{ or } D)$. Therefore, $P(A) = P(D)$.

[3] James Pryor. Immunity to error through misidentification. *Philosophical Topics*, 1999.

[4] Willard Van Orman Quine. *Word and Object*. The M.I.T. Press, 1960.

[5] Robert C. Stalnaker. Indexical belief. *Synthese*, 49:129–151, 1981.