Alexandru Marcoci

Monty Hall saves Dr. Evil: on Elga's restricted principle of indifference

Article (Published version) (Refereed)

Erkenn
https://doi.org/10.1007/s10670-018-0018-4

ORIGINAL RESEARCH

Monty Hall Saves Dr. Evil: On Elga's Restricted Principle of Indifference

Alexandru Marcoci[1]

Received: 16 May 2017 / Accepted: 2 May 2018

Abstract In this paper I show that Elga's argument for a restricted principle of indifference for self-locating belief relies on the kind of mistaken reasoning that recommends the 'staying' strategy in the Monty Hall problem.

1 Elga's Restricted Principle of Indifference

Elga (2004) argues for a restricted principle of indifference for self-locating belief.

> *Indifference*: A rational agent ought to assign equal credence to worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable.[1]

This principle is restricted in two ways. Firstly, it only applies to centred worlds rather than being a full-blown principle of indifference. In this sense it differs from the kind of indifference principle usually discussed in philosophy of probability. Secondly, it only applies to centred worlds that agree on all uncentred propositions. In other words, this principle wouldn't apply in the following case: suppose W is the actual world centred on you and W' is a Matrix-like world in which one of the people connected to machines has the exact same subjective experiences you have in W. This principle does not recommend assigning equal credence to W and W'

[1] This is an elaboration of Elga's (2004: 387) INDIFFERENCE. The way I formulate this principle here: (1) brings to the fore that it is imposing a constraint on the credal state of any rational agent; and (2) elucidates the scope of the principle. See Weatherson (2005: 614) for a detailed discussion of (2).

✉ Alexandru Marcoci
a.marcoci@lse.ac.uk

[1] Department of Government, London School of Economics and Political Science, Houghton Street, London WC2A2AE, UK

Published online: 06 June 2018

because although you and your counterpart have indistinguishable experiences, you inhabit different worlds. Elga labels a principle of indifference that would apply in such a situation the ABSURD-CLAIM-THAT-I-DON'T-ENDORSE (387). To illustrate his restricted principle of indifference, Elga introduces the story of Dr. Evil:

> *Dr. Evil*: "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, (…) 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'" (383)

Elga argues that upon receiving this message, Dr. Evil should assign equal credence to being Dr. Evil and to being Dup. In this paper I rationally reconstruct Elga's argument and show that it relies on the kind of mistaken reasoning that recommends the 'staying' strategy in the *Monty Hall* problem.

2 Elga's Argument for *Indifference*

Consider the following variation of *Dr. Evil*:

> *Comatose Dr. Evil*: Just like *Dr. Evil*, only that the scientists tell Dr. Evil that when he was asleep they've put Dup to sleep, too, and that they flipped a coin with bias .9 towards Tails. Then they made sure only one of them woke up: if the coin landed Heads, it was Dr. Evil (and Dup is in a coma). If the coin landed Tails, it was Dup (and Dr. Evil is in a coma).[2]

If the coin lands Heads, Dr. Evil is reading the message from PDF. If the coin lands Tails, Dr. Evil is in a coma on the Moon and Dup is reading the message back in the skepticism lab. In *Comatose Dr. Evil*, Elga argues Dr. Evil ought to align his credence that he is Dr. Evil to the bias of the coin. In other words, upon reading the message, Dr. Evil's degree of belief in being Dr. Evil ought to be .1.[3]

[2] This is a variation of Coma in Elga (2004: 390–391). Elga in fact moves away from *Dr. Evil* and develops his entire argument for INDIFFERENCE based on a completely analogous set of scenarios involving Al and his Duplicate. Nevertheless, there is no need to do that, and I will present his reasoning as it applies to *Dr. Evil*.

[3] Those familiar with Elga's (2000) discussion of the Sleeping Beauty problem may find surprising what he says about Dr. Evil's degrees of belief in *Comatose Dr. Evil*. Such a view goes against the thirder answer to the Sleeping Beauty problem. Titelbaum (2012) has already noticed this tension: "it was Elga himself who originally argued for the 1/3 answer to the Sleeping Beauty Problem, an answer that is incompatible with the Relevance-Limiting Thesis's position on the irrelevance of centered evidence to

Consider another variation of *Dr. Evil*:

Coin Toss Dr. Evil: Just like *Dr. Evil*, only that the scientists tell Dr. Evil that while they were duplicating him they flipped a coin with bias .9 towards Tails. But they assure him the coin toss had no impact on the duplication process.

In *Coin Toss Dr. Evil*, upon receiving the message from PDF, Dr. Evil should assign probability .1 to the coin having landed Heads (H, and T for Tails). Secondly, since the coin toss is independent from the duplication process, he should assign the same probability conditional on him being Dr. Evil (E, and D for Dup). That is:

$$P(H) = .1 \tag{1}$$

$$P(H|E) = .1 \tag{2}$$

Suppose in *Coin Toss Dr. Evil*, PDF were to send Dr. Evil a second message saying that if the coin landed Heads then Dup fell in a coma and Dr. Evil is now reading the message and if the coin landed Tails, Dup is reading the message and Dr. Evil is in a coma on the Moon, that is HE ∨ TD. Elga argues that in such case, Dr. Evil's credal state in *Coin Toss Dr. Evil* upon reading the second message should align with his credal state in *Comatose Dr. Evil* upon reading the message of that scenario.[4] In other words,

$$P(H|HE \vee TD) = .1 \tag{3}$$

(1)–(3) are enough to derive Dr. Evil's degree of belief in being himself in *Coin Toss Dr. Evil* after being told about the coin toss but before receiving the information that HE ∨ TD:

- From (2) and (3): P(H|HE ∨ TD) = P(H|E)
- By def. of cond. prob.: P(HE)/(P(HE) + P(TD)) = P(HE)/(P(HE) + P(TE))
- By simplification: P(TD) = P(TE)
- By independence of the duplication and the coin toss: P(T)P(D) = P(T)P(E)
- By simplification: P(D) = P(E).

Therefore, Dr. Evil should assign equal credence to being himself and to being Dup in *Coin Toss Dr. Evil*, after being told about the duplication, but before being told that HE ∨ TD. Since the coin toss in *Coin Toss Dr. Evil* has no causal impact on the duplication process, Dr. Evil's credal state after being told about the duplication and the coin toss (but before receiving the second message) is the same as his credal state in *Dr. Evil* upon simply being told he had been duplicated. It is true that in *Coin Toss Dr. Evil* the scientists tell Dr. Evil more than in *Dr. Evil*, but that

Footnote 3 continued
uncentered propositions. A thirder about Sleeping Beauty can't just assume that [Dr. Evil] should assign a degree of belief of 0.10 to heads when he awakens in [*Comatose Dr. Evil*]!" (353).

[4] "So when [Evil] wakes up in the [*Comatose Dr. Evil*] case, he has just the evidence about the coin toss as he would have if he had been awakened in [*Coin Toss Dr. Evil*] and then been told [HE ∨ TD]." (Elga 2004: 391).

additional information has no bearing on whether he is Dup or Dr. Evil. Therefore, in *Dr Evil* he should divide his credence equally between being Dr. Evil and being Dup upon receiving the message about duplication from the scientists.

Finally, *Dr. Evil* is taken by Elga to be a prototypical example of a rational agent contemplating worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable. Consequently the move from 'a rational agent' to Dr. Evil is done *without loss of generality*. That means that whatever rational requirements bind Dr. Evil's credal state, they ought to bind, on pain of irrationality, any agent. In particular, if Dr. Evil is rationally required to assign equal credence to the centred worlds he is contemplating, so should any agent. Since the above argument establishes, according to Elga, that Dr. Evil should indeed be indifferent between the world centred on himself and the one centred on Dup, so should any other rational agent, and *Indifference* follows.

3 The Monty Hall Problem

At the same time PDF is trying to thwart Dr. Evil's plans, on some TV set Monty Hall attempts to trick a contestant into making the losing choice in a game show:

> *Monty Hall*: Monty presents a game contestant with three doors. Behind two of these doors there is a goat. One of the doors, however, hides a brand new car. The contestant is asked to pick a door. Monty then opens one of the other two doors such that he doesn't give the prize away. Afterwards he asks the contestant which door she wants to open - the one she initially chose, or the other remaining closed door.

Suppose the door behind which the car is hidden is chosen at random. Suppose further that the contestant first picks Door 1. Monty hopes the contestant will reason in the following way: 'initially, there was a 1/3 chance the car was behind Door 1. Now that Monty opened one door hiding a goat, there are only two possible locations the car could be in, i.e. behind Door 1 or behind the door Monty left unopened. Therefore the probability the car is behind the door I selected increased to ½ and the probability the car is behind the other unopened door is also ½. So there is no reason for me to switch.'[5]

Here is a probabilistic model vindicating the above informal reasoning. Let Car 1 stand for the car being behind Door 1, Car 2 for the car being behind Door 2 and Car 3 for the car being behind Door 3. The sample space assumed above is {Car 1, Car 2, Car 3}, while the information Monty gives away when opening Door X that hides a goat is taken to be $\sim$ Car X. Finally, P(Car 1) = P(Car 2) = P(Car 3) = 1/3. Assume Monty opens Door 2. Then, upon receiving the information from Monty, the contestant will update her beliefs in the following way:

[5] This is a standard assumption made in this puzzle, but notice it relies on a type of conservativism: a rational agent should not revise her strategy, unless she has a positive reason to do so.

$$\begin{aligned} \mathrm{P}(\text{Car 1} \mid \sim \text{Car 2}) &= \mathrm{P}(\text{Car 1} \,\&\, \sim \text{Car 2})/\mathrm{P}(\sim \text{Car 2}) \\ &= \mathrm{P}(\text{Car 1})/(\mathrm{P}(\text{Car 1}) + \mathrm{P}(\text{Car 3})) \\ &= (1/3)/(2/3) \\ &= 1/2 \end{aligned}$$

Therefore the contestant should 'stay'. As it is well known, however, this reasoning is incorrect.

Bovens and Ferreira (2010: 474–476), following Speed's (1985) discussion of Shafer (1985), explain the mistake in terms of the fact that when we are informed of some proposition "we do not only learn the proposition in question, but also that we have learned the proposition as one of the many propositions that we might have learned." (474) The difference between updating on some proposition rather than updating on learning that proposition is nicely highlighted in Halpern (2004: 128–9):

> If I think my wife is much more clever than I, then I might be convinced that I will never learn of her infidelity should she be unfaithful. So, my conditional probability for Y, 'I will learn that my wife is cheating on me', given X, 'She will cheat on me', is very low. Yet, the probability of Y if I actually learn X is clearly 1.

Applying this insight to the *Monty Hall* problem Bovens and Ferreira explain the contestant's mistaken reasoning by the fact that she updated only on the content of the information Monty gave her when he opened Door 2 and revealed a goat. If she instead were to consider how the information Monty can pass on to her is constrained she would notice that the probability Monty would open a particular door is not the same irrespective of the state of world. This is easy to see: assume the car is behind Door 3, then the goats are behind Doors 1 and 2. Monty cannot open the former, as this is the one the contestant chose at the beginning of the round. Therefore Monty is forced to open Door 2. An analogous reasoning applies if the car is behind Door 2. But if the car is behind Door 1, then Monty can open either Door 2 or Door 3. So the probability with which he would open Door 2, say, in this case can be lower than 1. This asymmetry in how Monty can communicate with the contestant is made clear by considering the protocol under which information can accrue to the contestant. A conditional probability table can be used to specify a protocol:

Protocol 1 for *Monty Hall*	Car 1	Car 2	Car 3
"Goat 2"	½	0	1
"Goat 3"	½	1	0

In this table, each row corresponds to a possible item of information the contestant could receive. Each cell corresponds to the probability with which the contestant will learn that item of information at each possible world. This table can be used to construct a sophisticated event space in which we take into consideration every piece of information the agent could receive. Such a space would contain four atomic events with non-zero probability: {(Car 1, "Goat 2"), (Car 1, "Goat 3"), (Car 2, "Goat 3"), (Car 3, "Goat 2")}. Note the differences between this sophisticated model and the model above vindicating the mistaken reasoning recommending 'staying'. Firstly, the model contains 4 events with positive probability. Secondly, in this model we express the fact that Door 2 hides a goat and that Door 3 hides a goat as disjoint propositions which are not reducible to $\sim$ Car 2 and $\sim$ Car 3, respectively. In other words, carefully accounting for the process by which information accrues to the contestant turns the set of propositions on which she can conditionalise into a partition. This is implicit in Shafer's formal model of protocols as trees (Shafer 1985: Appendix 1) and is discussed at length in Grünwald (2013). The latter also formulates a rule of thumb: "briefly, for general spaces Y, if the set of events X on which you can condition is not a partition of Y, then conditioning on any of these events is unsafe." (Grünwald 2013: 243).

We can now calculate again how the contestant should change her degrees of belief upon Monty opening Door 2, say, and revealing a goat.

$$\begin{aligned}
&\mathrm{P}(\text{Car 1}|\,\text{``Goat 2''}) \\
&= \frac{\mathrm{P}(\text{``Goat2''}|\text{Car 1})\mathrm{P}(\text{Car 1})}{\mathrm{P}(\text{``Goat2''}|\text{Car 1})\mathrm{P}(\text{Car 1}) + \mathrm{P}(\text{``Goat2''}|\text{Car 2})\mathrm{P}(\text{Car 2}) + \mathrm{P}(\text{``Goat2''}|\text{Car 3})\mathrm{P}(\text{Car 3})} \\
&= \frac{(1/2 * 1/3)}{(1/2 * 1/3) + (0 * 1/3) + (1 * 1/3)} \\
&= 1/3
\end{aligned}$$

Therefore, taking into account the asymmetry of the way in which information may accrue to her, the contestant learns something new about where the car may be. Is this the only protocol that would make sense in *Monty Hall*? Although the puzzle is quite detailed with respect to how information is being delivered to the contestant, the scenario does not say Monty flips a fair coin in order to choose which door to open when the car is behind Door 1. Another protocol compatible with the story would be:

Protocol 2 for *Monty Hall*	Car 1	Car 2	Car 3
"Goat 2"	¾	0	1
"Goat 3"	¼	1	0

This represents a situation in which Monty would have a preference for opening Door 2 when the car is behind Door 1 and the contestant chooses Door 1 at the

beginning of the game. Imagine him flipping a coin with bias ¾ towards opening Door 2 when the car is behind Door 1. Then upon learning that Door 2 hides a goat, the contestant's credence in the car being behind Door 1 should go from 1/3 to 3/7.

So what is the correct answer then: 'staying' or 'switching'? The first probabilistic model suggests the contestant's rational posterior credence should be ½ and hence she should 'stay'. The second probabilistic model suggests her credence should be 1/3 (or 3/7 depending on the bias of the coin Monty flips) and hence she should 'switch'. What this shows is that the solution *Monty Hall* problem is sensitive to the structure of the interaction between Monty and the contestant is, despite the puzzle being silent on some of the details.

Are we rationally required, though, to take protocols into account? Bovens and Ferreira (2010: 480) and Shafer (1985: 264) claim it is implicit in the Principle of Total Evidence that an agent's probability model should give probabilities for all the different ways her learning may turn out. The motivation for this goes back to the observation that when you receive some information Y, you don't only learn the propositional content of Y, but also that you have received Y instead of Y'. So, insofar as Y represents your evidence, so does the fact that you learned Y instead of Y'. Therefore if you take the Principle of Total Evidence as a requirement for forming rational beliefs, then protocols should be taken into account.

4 The Protocol of *Coin Toss Dr. Evil*

In his formal analysis of *Coin Toss Dr. Evil* (see Sect. 2), Elga assumes there are four possible states of the world: either the person reading the message from PDF is Dr. Evil or he is Dup; and either the coin landed Heads or it landed Tails, viz. {HE, TE, HD, TD}. Furthermore, he assumes the information PDF could send Dr. Evil, i.e. that either he is himself and the coin came up heads or that he is Dup and the coin came up Tails, is reducible to a disjunction of two possible states of the world (HE ∨ TD). Finally, Elga calculates the probability of H given HE ∨ TD. However, this is an analogous strategy to the one that leads to the 'staying' solution in the *Monty Hall* problem, as shown in the previous section. The moral of *Monty Hall* is that the Principle of Total Evidence requires that in cases in which information X is conveyed to an agent the appropriate formal model of her learning X should account for her learning not only that X is the case but also that she has learned X "as one of the many propositions that [she] might have learned." (Bovens and Ferreira 2010: 474) In other words, if we were to focus on the conditional probability of H given HE ∨ TD we would indeed get Elga's conclusion; but then we wouldn't be correctly modelling the fact that Dr. Evil *learns* HE ∨ TD. If we want to model the latter, then we need to formally account for the protocol with which information accrues to Dr. Evil.

So what is the protocol underlying *Coin Toss Dr. E*vil? We know PDF could send a second message to Dr. Evil saying "Heads and Dr. Evil or Tails and Dup". Following the discussion above, this should be modelled as a new proposition, "HE ∨ TD". However, we don't know anything else about what other information

the scientists could include in that second message. Consider the following protocol, where a and b are arbitrary parameters:

Protocol 1 for *Coin Toss Dr. Evil*	HE	TE	HD	TD
"HE ∨ TD"	a	0	0	b
…	…	…	…	…

Given this protocol, the probability of the coin having landed Heads given the scientists' message (P') is:

$$
\begin{aligned}
&\mathrm{P'(H)} = \mathrm{P(H|``HE \vee TD'')} \\
&= \frac{\mathrm{P(``HE \vee TD''|H)P(H)}}{\mathrm{P(``HE \vee TD''|HE)P(HE) + P(``HE \vee TD''|TE)P(TE) + P(``HE \vee TD''|HD)P(HD) + P(``HE \vee TD''|TD)P(TD)}} \\
&= \frac{\mathrm{P(``HE \vee TD''|HE \vee HD)P(HE \vee HD)}}{\mathrm{aP(HE) + bP(TD)}} \\
&= \frac{\mathrm{P((``HE \vee TD''\&HE) \vee (``HE \vee TD''\&HD))}}{\mathrm{aP(HE) + bP(TD)}} \\
&= \frac{\mathrm{P(``HE \vee TD''\&HE) + P(''HE \vee TD''\&HD)}}{\mathrm{aP(HE) + bP(TD)}} \\
&= \frac{\mathrm{P(``HE \vee TD''|HE)P(HE)}}{\mathrm{aP(HE) + bP(TD)}} \\
&= \frac{\mathrm{aP(H)P(E)}}{\mathrm{aP(H)P(E) + bP(T)P(D)}} \\
&= \frac{\mathrm{aP(E)}}{\mathrm{aP(E) + 9bP(D)}}
\end{aligned}
$$

Elga claims that P'(H) should be equal to the probability of Heads, that is 1/10. Solving the equation

$$\frac{\mathrm{aP(E)}}{\mathrm{aP(E) + 9bP(D)}} = \frac{1}{10}$$

we obtain that

$$\mathrm{aP(E)} = \mathrm{bP(D)}.$$

Therefore (assuming there are no extreme values) the probability of being Dr. Evil is equal to the probability of being Dup if and only if a = b. In other words, the agent should consider it equally likely to be told HE ∨ TD in a Heads world in which he is Dr. Evil as in a Tails world in which he is Dup. This is by no means certain. One could easily conceive of the following protocol underlying *Coin Toss Dr. Evil*:

Protocol 2 for *Coin Toss Dr. Evil*	HE	TE	HD	TD
"HE ∨ TD"	½	0	0	1
"HD ∨ TE"	0	½	½	0

Protocol 2 for *Coin Toss Dr. Evil*	HE	TE	HD	TD
"HE ∨ TE ∨HD"	½	½	½	0

In this case, PDF can send three messages to Dr. Evil and they have different likelihoods based on the possible world that obtains. The relevant asymmetry is that the scientists will definitely announce HE ∨ TD whenever TD obtains but will only announce it with probability ½ when HE obtains. In this case, even if one were to accept Elga's claim that P'(H) = 1/10, then P(E) = 2P(D), which means that Dr. Evil would consider it twice more likely to be himself rather than Dup.

To sum up, if we apply the Principle of Total Evidence in *Coin Toss Dr. Evil* and model the information Dr. Evil receives explicitly, Elga's conclusion only follows if a = b. Nevertheless, the assumption that a = b is not incompatible with Elga's *Coin Toss Dr. Evil* (the scenario underdetermines the different messages PDF could send to Dr. Evil). So *prima facie* it may seem that Elga's argument simply requires an additional innocuous assumption about the protocol underlying *Coin Toss Dr. Evil* for the conclusion that P(E) = P(D) to go through.

5 Against Elga's Argument

The fact that the conditional probability of the scientists' announcement in HE has to be equal to the conditional probability in TD spells trouble for Elga's argument for *Indifference*. Recall Elga's argumentative strategy:

> Claim A: Dr. Evil's credal state after receiving the message from PDF in *Comatose Dr. Evil* is identical to his credal state in *Coin Toss Dr. Evil* after being told he has been duplicated and learning "HE ∨ TD".
>
> Claim B: Therefore Dr. Evil should assign equal credences to being himself and being Dup upon being told he has been duplicated in *Coin Toss Dr. Evil* (and before receiving the second message).[6]
>
> Claim C: But upon learning he has been duplicated in *Coin Toss Dr. Evil* (and before receiving the second message), his credal state is identical to his credal state in *Dr. Evil* (modulo the irrelevant difference that he now knows a coin independent of his duplication has been flipped).
>
> Claim D: Therefore, in *Dr. Evil*, he should assign equal credences to being himself and being Dup.
>
> Claim E: Given *Dr. Evil* is a prototypical scenario for the restricted principle of indifference for self-locating beliefs, *Indifference* holds.

Consider Claim A. The argument in the previous section establishes that Claim A only holds if a particular restriction is placed on the protocol under which information is passed to Dr. Evil/Dup by the scientists. Not all possible learning

[6] This follows from (1) to (3) *above*.

scenarios will support Claim A. So if Claim A were to hold, then the scientists should follow a protocol compatible with a = b, and both Dr. Evil and Dup should be aware of this protocol.

One could reply to this that there is nothing preventing us from tweaking *Coin Toss Dr. Evil* in order to account for this protocol. Assume we come up with a story that makes the receipt of the second message from PDF equally likely in HE as in TD. Let's call this new scenario *Coin Toss Dr Evil*$^+$. Claim A only holds if we replace *Coin Toss Dr. Evil* with *Coin Toss Dr. Evil*$^+$.

Now, consider Claim C. If Claim C were to hold, then the same knowledge of the protocol which is now embedded into *Coin Toss Dr. Evil*$^+$ should obtain in *Dr. Evil*. Dr. Evil should be aware that the scientists could flip a fair coin independently of the duplication process, and if they do flip it, they could announce that the coin came up heads to Dr. Evil or that the coin came up tails to Dup. Finally, Dr. Evil should also be aware that it is as likely for them to announce this if he indeed is Dr. Evil and the coin came up heads as it is if he is in fact Dup and the coin came up tails. So, for Claim C to hold *Dr. Evil* has to be replaced with *Dr. Evil*$^+$.

However, Dr. Evil's credal state in *Dr. Evil*$^+$ contains this protocol and hence his credal state is no longer a prototypical credal state of an agent faced with worlds that agree on all uncentred propositions and are centred on agents whose experiences are indistinguishable. Hence, Dr. Evil can no longer serve as the instantiation of an arbitrary rational agent as the move from *Indifference* to *Dr. Evil*$^+$ cannot be done without loss of generality. To wit, Dr. Evil assigns a credence of ½ to being Dr. Evil not in a prototypical case of *Indifference*, but in a case in which information accrues to him according to a particular protocol. In consequence, the step back from *Dr. Evil* to *Indifference* is no longer warranted.

Therefore, either Elga's argument fails at the very outset when credences from *Comatose Dr. Evil* are imported to *Coin Toss Dr. Evil*, or at the last step when Dr. Evil's credences cannot be attributed to an arbitrary rational agent dealing with worlds agreeing on all uncentred propositions and centred on agents whose experiences are indistinguishable.

Before concluding, here is another way of making the same argument as in the above pages. Suppose Elga's argument is correct and hence:

> In [*Coin Toss Dr. Evil*], the coin toss is irrelevant to whether and how the duplication occurs. So [Evil]'s state of opinion (when he awakens) as to whether he is [Evil] or the duplicate ought to be the same in [*Coin Toss Dr. Evil*] as it is in [*Dr. Evil*]. (Elga 2004: 388)

Consider now a variation of *Coin Toss Dr. Evil* in which it is made clear that Protocol 2 underwrites the informational exchange between PDF and Dr. Evil and the latter knows this. In such a scenario the toss of the coin would also be "irrelevant to whether and how the duplication occurs". Therefore, by Elga's reasoning, Dr. Evil's credence function in *Dr. Evil* ought to match his credence function in this modified scenario, too. But as we saw above, with Protocol 2 in place, P(E) = 2P(D). Consequently in *Dr. Evil*, Dr. Evil ought to believe both that the probability of being himself is equal to that of being Dup and equal to 1/2, and

that it is twice the probability of being Dup. This would make Dr. Evil probabilistically incoherent.

6 Conclusion

In this paper I show that Elga's argument for *Indifference* fails. This failure is interesting for two reasons. Firstly, the restricted principle of indifference is part of both the Halfer (e.g. Elga 2000; Dorr 2002) and Thirder (e.g. Lewis 2001) answers to the Sleeping Beauty problem[7] as well as part and parcel of several arguments in the literature on self-location (e.g. Leitgeb and Bradley 2006; Ross 2010). Secondly, the mistake in Elga's argument is in itself interesting, as it illustrates the need for specifying a precise sample space when applying conditionalization. In this respect, the paper shows that *Monty Hall* still has important lessons to teach us.

Acknowledgements I would like to thank Luc Bovens, Richard Bradley, Christian List, Graham Oddie, James Nguyen, Silvia Milano and an anonymous referee for their very generous comments on earlier versions of this paper. Thanks also to audiences at the Bristol-LSE Graduate Conference in Formal Epistemology and the Second Munich Graduate Workshop in Mathematical Philosophy.

References

Bovens, L., & Ferreira, J.-L. (2010). Monty Hall drives a wedge between Judy Benjamin and the Sleeping Beauty: A reply to Bovens. *Analysis, 70*(3), 473–481.

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

Elga, A. (2000). Self-locating belief and the Sleeping Beauty Problem. *Analysis, 60*(266), 143–147.

Elga, A. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research, 69*(2), 383–396.

Grünwald, P. (2013). Safe probability: Restricted conditioning and extended marginalization. In L. C. van der Gaag (Ed.), *Symbolic and quantitative approaches to reasoning with uncertainty: 12th European conference, ECSQARU 2013, Utrecht, The Netherlands, July 8–10, 2013* (pp. 242–253). Berlin: Springer.

Halpern, J. (2004). Sleeping Beauty reconsidered: Conditioning and reflection in asynchronous systems. In *Oxford studies in epistemology* (Vol. 1, pp. 111–142). Oxford University Press

Leitgeb, H., & Bradley, D. (2006). When betting odds and credences comeapart: more worries for Dutch book arguments. *Analysis, 66,* 119–127.

Lewis, D. (2001). Sleeping Beauty: Reply to Elga. *Analysis, 61*(3), 171–176.

Meacham, C. (2008). Sleeping Beauty and the dynamics of de se beliefs. *Philosophical Studies, 138*(2), 245–269.

Ross, J. (2010). Sleeping Beauty, countable additivity, and rational dilemmas. *The Philosophical Review, 119,* 411–447.

[7] See Meacham (2008) for a detailed discussion of this.

Shafer, G. (1985). Conditional probability. *International Statistical Review/Revue Internationale de Statistique, 53*(3), 261–275.

Speed, T. P. (1985). Discussion of paper by G. *International Statistical Review/Revue Internationale de Statistique, 53*(3), 276.

Titelbaum, M. G. (2012). *Quitting certainties: A Bayesian framework modeling degrees of belief*. Oxford: OUP.

Weatherson, B. (2005). Should we respond to evil with indifference? *Philosophy and Phenomenological Research, 70*(3), 613–635.