Erkenntnis (2024) 89:355–366
https://doi.org/10.1007/s10670-022-00536-8

ORIGINAL RESEARCH

A Bayesian View on the Dr. Evil Scenario

Feraz Azhar[1,2] · Alan H. Guth[3] · Mohammad Hossein Namjoo[4]

Received: 1 September 2021 / Accepted: 13 March 2022 / Published online: 21 June 2022

Abstract

In *Defeating Dr. Evil with Self-Locating Belief*, Adam Elga proposes and defends a principle of indifference for self-locating beliefs: if an individual is confident that his world contains more than one individual who is in a state subjectively indistinguishable from his own, then he should assign equal credences to the hypotheses that he is any one of these individuals. Through a sequence of thought experiments, Elga in effect claims that he can derive the credence function that should apply in such situations, thus justifying his principle of indifference. Here we argue, using a Bayesian approach, that Elga's reasoning is circular: in analyzing the third of his thought experiments, he uses an assertion that is justifiable only if one assumes, from the start, the principle of indifference that he is attempting to justify. We agree with Elga that the assumption of equal credences is a very reasonable principle, in the absence of any reason to assign unequal credences, but we do not agree that the equality of credences can be so derived.

1 Introduction

Self-locating beliefs—namely, those beliefs that situate an agent at a location or a time—are commonplace. You acquire a self-locating belief whenever you come to learn of the time by glancing at your watch or if you learn of your location by observing

✉ Feraz Azhar
fazhar@nd.edu

Alan H. Guth
guth@ctp.mit.edu

Mohammad Hossein Namjoo
mh.namjoo@ipm.ir

[1] Department of Philosophy, University of Notre Dame, Notre Dame, IN 46556, USA

[2] Black Hole Initiative, Harvard University, Cambridge, MA 02138, USA

[3] Department of Physics, Laboratory for Nuclear Science, and Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[4] School of Astronomy, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

a street sign. The converse situation can also arise, in which you may be initially certain about some self-locating belief (such as the time) but then you gain a new piece of evidence (that your watch is broken), rendering you uncertain about the self-locating belief [Bradley (2007)]. More generally, the question arises as to how one should distribute one's credence over self-locating beliefs about which one is uncertain.

In *Defeating Dr. Evil with Self-Locating Belief*, Elga (2004) defends a response to this question. The response amounts to a version of the principle of indifference—where, under specified circumstances, one should spread one's credence equally over the various hypotheses. More specifically, the version of the principle of indifference that Elga defends is as follows:

> INDIFFERENCE: Similar centered worlds deserve equal credence (Elga, 2004, p. 387).

A centered world is a possible world with a designated individual and a designated time. It can be specified as a triple $(w; i; t)$, where w is a possible world, i is some individual in w, and t is a time. Similar centered worlds satisfy two conditions: (i) they agree on the first argument of the triple (but not necessarily on the latter two); (ii) the individuals that exist in each centered world are in subjectively indistinguishable states.[1] Note that while principles of indifference have been extensively discussed, the formulation in Elga (2004) has been particularly influential.[2]

In justifying INDIFFERENCE, Elga considers a sequence of three (related) thought experiments: DUPLICATION, TOSS&DUPLICATION, and COMA. We claim that Elga's analysis of the COMA scenario is flawed, relying on an assertion which can be justified only by assuming the truth of the proposition that he is trying to demonstrate. We are not disputing the claim that INDIFFERENCE is reasonable—indeed, in an upcoming paper we will make use of such a principle in assessing certain cosmological theories. Our remarks here target only the justification that Elga provides for INDIFFERENCE. For us, INDIFFERENCE is a principle that one can reasonably adopt, based on some sort

[1] INDIFFERENCE can be contrasted with the claim that two triples in which the possible worlds themselves are different (but where the individuals are in subjectively indistinguishable states) should receive equal credence—a claim that Elga holds to be absurd, since two possible worlds can have very different levels of plausibility.

[2] For example, Bradley devotes an entire chapter of his Ph.D. thesis [Bradley (2007)] and an entire section of a paper [Bradley (2011)] to defending Elga's formulation of indifference against criticisms by Weatherson (2005). Birch (2013)—in his refutation of Bostrom's (2003) argument that we may very well be living in a computer simulation—explores Elga's formulation of indifference as a possible justification of Bostrom's assumptions. Wilson (2017) describes the applicability of the principle of indifference in Everettian quantum mechanics, employing Elga's formulation throughout. Sebens and Carroll (2018) view their work as applying Elga's principle of indifference to Everettian quantum mechanics to show that it leads to the Born rule for probabilities of the outcomes of measurements. And the *proof* that Elga provides for his formulation of the principle of indifference has also been persuasive. Carroll and Sebens (2015, p. 4) declare that

> Elga …has given convincing arguments in favor of indifference in the case of identical classical observers. Crucially, this result is not simply postulated as the simplest approach to the problem, but rather derived from seemingly innocuous principles of rational reasoning.

of appeal to the absence of any reason to assign unequal credences, but not based on any determinative calculation of the sort presented by Elga.[3]

Our plan for this paper is as follows. In Sect. 2 we summarize Elga's thought experiments, and describe our claim that the analysis of COMA is flawed. Sect. 3 describes a Bayesian calculation that shows in detail how we believe the COMA scenario should be analyzed. We summarize our argument in Sect. 4. In an appendix, we reexamine step-by-step a crucial footnote from Elga's paper, showing how the conclusions found there are modified by our analysis of COMA.

2 Reanalysis of Elga's Thought Experiments

Elga's first thought experiment describes two similar centered worlds in which a person named "Al" is duplicated. The experiment is described as follows.

> 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 (Elga, 2004, p. 388).

The issue at stake is how Al should distribute his credence between the hypothesis that he is Al and the hypothesis that he is the duplicate. In order to justify why Al should distribute his credence *evenly* between the two hypotheses (which is indeed what would be implied by INDIFFERENCE), Elga introduces two further experiments: TOSS&DUPLICATION and COMA (the latter will be described shortly).

> 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 (Elga, 2004, p. 388).

In our discussion we will generalize the chance of the coin landing heads from the specified value of 10% to an arbitrary $P_0(H)$, assuming only that $P_0(H)$ is not equal to 0 or 1. Elga claims (and we agree) that

> ...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 (Elga, 2004, pp. 388–389).

To show that Al ought to divide his credence evenly between the hypothesis that he is Al and that he is the duplicate (in TOSS&DUPLICATION), Elga states three claims, which we (re)describe here. Following Elga (2004) and Weatherson (2005), we will use the following abbreviations:

[3] Since we agree that INDIFFERENCE is a reasonable principle, we also agree with Elga's conclusion that Dr. Evil ought to surrender, assuming that the fear of torture outweighs the thrill of his evil plans!

$$
\begin{aligned}
H &: \text{the coin lands ‘heads’;}\\
T &: \text{the coin lands ‘tails’;}\\
A &: \text{I am Al;}\\
D &: \text{I am Dup (Al’s duplicate).}
\end{aligned}
$$

The credence function that Al ought to have immediately upon awakening will be denoted by $P(\cdot)$. Elga’s three claims can be stated as follows:

(C1) Al’s credence in H ought to be equal to the chance of the coin landing heads:

$$P(H) = P_0(H). \tag{1}$$

(C2) Al’s credence in H, given ((H and A) or (T and A)), ought to be equal to his credence in H:

$$P(H|HA \text{ or } TA) = P_0(H). \tag{2}$$

(C3) Al’s credence in H, given ((H and A) or (T and D)), ought to be equal to his credence in H:

$$P(H|HA \text{ or } TD) = P_0(H). \tag{3}$$

Claim (C3) (which Elga indeed deems to be controversial) is established by considering another thought experiment, viz. Coma.

> Coma: As in Toss&Duplication, the experimenters toss a coin and duplicate Al. But the following morning, the experimenters ensure that *only 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) (Elga, 2004, pp. 390–391).

Elga then claims (a claim with which we agree) that one can determine the value of the left-hand side of Eq. (3) by considering what Al’s credence in H should be in Coma. That is, when Al awakens in Coma, were he to indeed awaken, his situation would be exactly as it would have been in Toss&Duplication, but then updated by the new information (HA or TD). Thus, Al’s credence function in Coma, if he awakens, ought to be given by

$$P_{\text{Coma}}(\cdot) = P(\cdot|HA \text{ or } TD). \tag{4}$$

In particular,

$$P_{\text{Coma}}(H) = P(H|HA \text{ or } TD). \tag{5}$$

Al’s credence in H in Coma is ascertained, by Elga, using the following argument (on p. 392)[4]:

[4] Here, we have modified the exact quote to accord with our notation, changing “HEADS” to H and “10%” to $P_0(H)$.

> Before Al was put to sleep, he was sure that the *chance* of the coin landing heads was $P_0(H)$, and his credence in H should have accorded with this chance: it too should have been $P_0(H)$. 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 H should continue to accord with the chance of H at the time of the toss. In other words, his degree of belief in H should continue to be $P_0(H)$.

In short, Elga's claim is that in Coma, if Al awakens, he ought to continue to set his credence in the coin landing heads to the known chance of heads, $P_0(H)$.

Given claims (C1), (C2), and (C3), Elga (correctly) concludes that $P(A) = P(D) = 1/2$ in Toss&Duplication and thus (by the reasoning described above) in Duplication. We, however, take issue with the justification provided for claim (C3).

From our point of view, Elga appears to be using circular reasoning in (C3), where he claims that Al's credence in H should remain equal to $P_0(H)$ when he awakens in Coma. In particular, it is not the case that when Al wakes up "his epistemic situation with respect to the coin toss is just the same as it was before he went to sleep. He has neither gained nor lost information relevant to the toss outcome." We contend that Al has both gained and lost information relevant to the toss outcome:

(i) He has lost information as regards his identity: before Al went to sleep he was sure that he was Al; after he awakens, the possibility arises that he is Dup. No amount of introspection (or an inspection of his external environment) can reveal to him whether he is Al or Dup. (This is what we take to be the meaning of the assumption that Al and Dup would be in "subjectively indistinguishable" states.)

(ii) He has gained the information that he is now in a predicament in which his identity (again, about which he is now unsure) is perfectly correlated with the toss outcome.

Al's uncertainty about his identity [as described in (i)] is directly relevant to his belief about the toss outcome [as described in (ii)].

In particular, when Al awakens in Toss&Duplication, he has no way of knowing if he is Al or Dup. If he adopts Indifference, he will conclude that he should give equal credence to each possibility. However, since Elga is trying to demonstrate Indifference, circularity can be avoided only if we allow Al to adopt an initial credence that is not equal to 1/2. Hence we set $P(A)$, Al's credence that he is Al upon awakening in Toss&Duplication, equal to some arbitrary 'prior' credence Pr_A, which Al is free to choose. [5]

In Coma, as discussed just above Eq. (4), Al (if and when he awakens) begins with the same credences as in Toss&Duplication, which are then updated by the new information (HA or TD). The role of Pr_A is important, because if Pr_A is not

[5] For some background on 'prior' credence functions, see for example Meacham (2016), Dorr and Arntzenius (2017), and Isaacs, Hawthorne, and Russell (forthcoming). In contrast to some ways of understanding such a prior credence function, in the present setting $P(\cdot)$ is not to be understood as the credence function of an agent who has no evidence whatsoever. Al, for example, does know (among other things) that he is in a scenario in which he has been duplicated.

equal to 1/2, then clearly Al's credence in H *is* affected by the new information. For example, if Pr_A is nearly one, then his credence in H should obviously increase on awakening in COMA, since if he is Al, then the coin must have landed heads. This issue is well-described by a standard Bayesian analysis, which we give in the next section. We will see that Al's credence in H upon awakening in COMA should remain $P_0(H)$ if and only if Al assumes that $Pr_A = 1/2$. This means that Elga's conclusion is valid if and only if one assumes INDIFFERENCE from the start. [6]

3 A Bayesian Approach

In this section we apply Bayes' theorem to determine how Al should update his credence in H, using the new information, (HA or TD), that he acquires on awakening in the COMA scenario. To determine how Al should update his credence in H, we will compute the right-hand side of Eq. (5), recalling that $P(\cdot)$ is the credence function in TOSS&DUPLICATION.

Our derivation will make use of the fact that in TOSS&DUPLICATION the H/T choice is independent of the A/D choice, which follows as a consequence of claim (C2) (with which we do not take issue). This independence follows from the same words that Elga uses to justify claim (C2): "So Al should count the toss outcome as irrelevant to who he is" (Elga, 2004, p. 389). Formally, the independence of H and A can be derived from claim (C2) by noting that (H or T) is true, so $P(H|HA \text{ or } TA) = P(H|A)$, and therefore (C2) implies that $P(H|A) = P_0(H)$. This is of course a way of stating that A and H are independent. Using Bayes' theorem, this statement is equivalent to $P(A|H) = Pr_A$.

Recalling that $P(H) = P_0(H)$ is the initially specified chance that the coin landed heads, Bayes' theorem can be used to rewrite the right-hand side of Eq. (5) as follows:

$$P(H|HA \text{ or } TD) = \frac{P(HA \text{ or } TD|H)}{P(HA \text{ or } TD)} P_0(H) \tag{6a}$$

$$= \frac{P(A|H)}{P(HA) + P(TD)} P_0(H) \tag{6b}$$

$$= \frac{Pr_A}{P_0(H)Pr_A + P_0(T)Pr_D} P_0(H) \tag{6c}$$

$$\equiv F\, P_0(H), \tag{6d}$$

[6] Note that Weatherson (2005) has also objected to Elga's claim (C3), but makes no mention of circularity or anything similar. His primary objection (although he raises others as well) relies on the view that one must distinguish between *risky* propositions, for which "we have good reason to assign a particular probability," and propositions which are *uncertain*, for which we "aren't really in a position to assign anything like a precise numerical probability." Weatherson argues that Al's question about his identity falls in the category of uncertainty. The result of the coin toss was risky before Al went to sleep, but when he awakens in the COMA scenario, the result of the coin toss becomes correlated with Al's identity, which changes it from risky to uncertain. Thus Weatherson questions whether Al can assign any credence, when he awakens, to the coin having landed heads, and suggests that maybe assigning a range of credences would be more appropriate.

where $Pr_D \equiv 1 - Pr_A$ is Al's prior credence in TOSS&DUPLICATION that he is Dup, and

$$F \equiv \frac{Pr_A}{P_0(H)Pr_A + P_0(T)Pr_D}. \tag{7}$$

In Eq. (6c) we have used the independence of H and A (and that of T and D). As we claimed at the end of the previous section, Al's credence in H after awakening in COMA, which is equal to $P(H|HA \text{ or } TD)$, remains equal to $P_0(H)$ only if $F = 1$. By rewriting F as

$$F = \frac{1}{1 + \dfrac{P_0(T)}{Pr_A}\big[Pr_D - Pr_A\big]}, \tag{8}$$

one can easily see that $F = 1$ only if the prior credences Pr_A and Pr_D (which must sum to 1) are each taken to be 1/2.[7]

Thus, taking into account the implications of Bayes' theorem, claim (C3) should be modified to a new claim, which we will denote by (C3′):

(C3′) Al's credence in H if he awakens in COMA, $P_{\text{COMA}}(H)$, which is equal to $P(H|HA \text{ or } TD)$, ought to be equal to the product of F [given in Eq. (7)] and his credence in heads $P(H) = P_0(H)$ as he awakens, before he updates his credences with the new information (HA or TD):

$$P_{\text{COMA}}(H) = P(H|HA \text{ or } TD) = F\,P_0(H). \tag{9}$$

Note that (C3′) agrees with (C3) only if Pr_A is assumed to be 1/2.

In his footnote 8, Elga uses claims (C1), (C2), and (C3) to show that $P(A) = P(D)$, which completes his demonstration that INDIFFERENCE is true for this situation. (In an appendix, he generalizes the argument to defend arbitrary instances of INDIFFERENCE.) Since we have argued that (C3) should be replaced by (C3′), the conclusion will of course be altered. In deriving Eq. (9), we calculated $P(H|HA \text{ or } TD)$, and hence F [see Eqs. (6) and (7)], assuming that Pr_A and $P_0(H)$ were given. We could imagine, however, that we were given the values of F and $P_0(H)$, and were asked to infer the value of $P(A)$, Al's credence in being Al when he awakens in TOSS&DUPLICATION. Equation (7) would still hold [replacing Pr_A and Pr_D by generic quantities $P(A)$ and $P(D)$], so we could then find $P(A)$ directly from Eq. (7), recalling that $P(D) = 1 - P(A)$ and $P_0(T) = 1 - P_0(H)$. The result would be

$$P_{\text{inferred}}(A) = \frac{F P_0(T)}{1 - F + 2F P_0(T)}, \tag{10}$$

where we use the special notation $P_{\text{inferred}}(A)$ to denote the value of $P(A)$ that is inferred from assumptions (C1), (C2), and (C3′). This calculation is a version of

[7] Recall that we have assumed that $P_0(H)$ is not equal to zero or one. If we had allowed $P_0(H) = 0$, then $F P_0(H)$ would equal $P_0(H)$ for any finite value of F. If we had allowed $P_0(H) = 1$ (and hence $P_0(T) = 0$), then Eq. (7) shows that F would equal 1 for any nonzero value of Pr_A.

Elga's derivation in footnote 8, which is shorter than Elga's argument and allows an arbitrary value for F. We can see immediately that if we assume that $F = 1$, as in claim (C3), we recover Elga's result that $P_{\text{inferred}}(A) = 1/2$. However, if we use the version of the claim based on Bayesian reasoning, namely (C3′), we find instead [by substituting Eq. (7) into Eq. (10)] the trivial conclusion that

$$P_{\text{inferred}}(A) = Pr_A. \tag{11}$$

That is, Al's credence in being Al upon awakening in TOSS&DUPLICATION, as inferred by considering the experiment in COMA, is exactly equal to whatever prior credence Pr_A that he assumed. Al will conclude that $P_{\text{inferred}}(A) = 1/2$ only if he assumed a prior credence Pr_A equal to 1/2.

The argument above is an abbreviated version of Elga's footnote 8, modified to use (C3′) instead of (C3). In an appendix we show in detail that if the steps of Elga's footnote 8 are followed exactly, but where F is introduced as in Eq. (9), we retrieve exactly the result of Eq. (10). Thus we see that once the circularity in Elga's formulation is removed, the demonstration of the truth of INDIFFERENCE disappears. INDIFFERENCE can still be adopted as a reasonable principle, but Elga's derivation of it is flawed.

4 Conclusion

Elga (2004) discusses a specific version of the principle of indifference, which applies to a situation where a possible world includes two or more individuals who, at some specified time for each of them, are in subjectively indistinguishable states. He illustrates this situation with an example called DUPLICATION, in which someone named Al is duplicated while he sleeps, along with his environment, so that Al and his duplicate awaken in subjectively indistinguishable states. Through a sequence of three thought experiments, Elga argues that when Al awakens, he ought to assign equal credence to being the duplicate or to being Al.

In this paper we have argued that, while it is perfectly reasonable for Al to *assume* that he is equally likely to be the duplicate, he is not compelled to make this assumption. The reasoning that Elga used, we believe, is circular. Specifically, we differ in the analysis of the third thought experiment, called COMA, in which a coin with a 10% chance of landing heads is tossed while Al sleeps, and then the duplication takes place as before. If the coin lands heads, only Al is allowed to wake up, with the duplicate remaining in a coma; but if the coin lands tails, only the duplicate is allowed to wake up. Our disagreement centers on the credence that Al should have, when (and if) he awakens, in the coin having landed heads. Elga's conclusions are based on the claim that Al's credence in heads when he awakens should remain 10%. We argue, however, that this claim is true if and only if Al *assumes* that he is equally likely to be Al or the duplicate, which is exactly the conclusion that Elga is trying to demonstrate. If Al does not make this assumption, then he might, for example, assume that he is much more likely to be Al than the duplicate. In that case, his credence in heads should be increased upon learning that he has woken up. We carried out a Bayesian analysis of this thought experiment, and showed that Al may assume any prior credence in his being Al, and no inconsistencies arise.

As long as Al has no reason to believe that he is more likely to be either Al or the duplicate, then we agree that it is reasonable to quantify this absence of evidence by adopting the default proposition that he is equally likely to be either. Elga's argument, however, did not rely on adopting a default proposition. Instead, Elga claimed to show directly from the descriptions of the thought experiments that Al could *deduce* that he should have equal credence in being Al or Dup. At the end of his argument, Elga proclaimed "So, INDIFFERENCE is true." Thus, Elga was arguing that INDIFFERENCE is more than a reasonable proposition, but is instead a logically compelling conclusion. We maintain, however, that this argument is flawed.

In an email exchange with Adam Elga, he pointed out that circularity could be avoided by accepting (C3) as an "undefended premise in the argument." To ensure the absence of circularity, it is important that (C3) has "independent appeal—credibility that does not derive from an antecedent commitment to INDIFFERENCE." We completely agree that if one does not provide support for (C3), but instead accepts it as a premise, then there is no circularity. For Elga, (C3) has appeal that is independent of INDIFFERENCE. For us, however, (C3) has no such appeal; but there is no cause for debate, since the status of Elga's argument turns on whether (C3) has such appeal, and we and Elga agree that there is no reason why intelligent folks should necessarily agree about the appeal of an undefended premise. Nevertheless, from our point of view, Elga's original argument remains circular.

Acknowledgements We thank Adam Elga for a very interesting and helpful email exchange and for his permission to include the summary of this exchange that appears in the final paragraph of Sect. 4. We also thank Robert Audi, Jeremy Butterfield, Kris McDaniel, and Nicholas Teh for helpful advice. F. A. acknowledges support from the Black Hole Initiative at Harvard University, which is funded through a grant from the John Templeton Foundation and the Gordon and Betty Moore Foundation. A. H. G.'s work was supported in part by the U.S. Department of Energy under Contract No. DE-SC0012567. The views presented in this paper do not necessarily reflect those of any person or funding agency mentioned above.

Appendix A: Generalizing Elga's Footnote 8

In Sect. 3, we defined F to be the Bayesian update factor with which Al's credence in H is multiplied when he acquires the new information (HA or TD), that is, when he learns that either he is Al and the coin landed heads, or else he is the duplicate (Dup) and the coin landed tails. In Eq. (10), we inverted the Bayesian update formula to determine $P_{\text{inferred}}(A)$, Al's credence in being Al, in terms of F and $P_0(T)$. [Recall that we are using the special notation $P_{\text{inferred}}(A)$ and $P_{\text{inferred}}(D)$ for Al's credence in being Al, or in being Dup, when expressed as a function of F and $P_0(T) \equiv 1 - P_0(H)$.] This formula shows that if one assumes that $F = 1$, then one concludes that $P_{\text{inferred}}(A) = 1/2$. Elga assumed that $F = 1$, without justification in our opinion, and concluded that $P_{\text{inferred}}(A) = 1/2$.

Elga's demonstration that the claims (C1), (C2), and (C3) imply that $P(A) = P(D) = 1/2$ is given in his footnote 8, which does not use Bayes' theorem. Since Elga's derivation is rather different from our derivation of Eq. (10), a reader could suspect that Elga's derivation might uncover more information than our Eq. (10). Here we show that this is not the case: if Elga's derivation is generalized to allow F to be arbitrary (rather than assuming that $F = 1$), the final result is the same as Eq. (10). To make this clear, we will go through the equations of Elga's footnote 8 step by step, but allowing for an arbitrary value of F. At each step, Elga's equation can be obtained by setting $F = 1$. We will indent the remainder of this paragraph to indicate that we are following Elga—for the most part using his language, to facilitate the comparison.

> Elga begins by setting the left-hand sides of Eqs. (2) and (3) equal to each other. Using Eq. (9) instead of Eq. (3), this gives
>
> $$P(H|HA \text{ or } TA) = \frac{1}{F} P(H|HA \text{ or } TD). \tag{A1}$$
>
> Rewriting Eq. (A1) using the definition of conditional probability, we obtain
>
> $$\frac{P(HA)}{P(HA \text{ or } TA)} = \frac{P(HA)}{F P(HA \text{ or } TD)}. \tag{A2}$$
>
> Some algebra then gets us that
>
> $$P(HA \text{ or } TA) = F P(HA \text{ or } TD). \tag{A3}$$
>
> Since HA, TA, and TD are all disjoint,
>
> $$P(TA) = F P(TD) + (F - 1) P(HA). \tag{A4}$$
>
> Since $P(TA)$ and $P(TD)$ add up to $P_0(T)$,
>
> $$P(TA) = \frac{F}{1+F} P_0(T) - \frac{1-F}{1+F} P(HA), \tag{A5}$$
> $$P(TD) = \frac{1}{1+F} P_0(T) + \frac{1-F}{1+F} P(HA). \tag{A6}$$
>
> Now set the left-hand sides of Eqs. (1) and (2) (in the main text) equal to each other:
>
> $$P_0(H) = P(HA|HA \text{ or } TA). \tag{A7}$$

It follows that

$$P_0(T) = P(TA|HA \text{ or } TA). \tag{A8}$$

Dividing the first equation [Eq. (A7)] by the second equation [Eq. (A8)], we obtain

$$\frac{P_0(H)}{P_0(T)} = \frac{P(HA|HA \text{ or } TA)}{P(TA|HA \text{ or } TA)}. \tag{A9}$$

Using the definition of conditional probability, we thus obtain

$$\frac{P_0(H)}{P_0(T)} = \frac{P(HA)}{P(TA)}. \tag{A10}$$

Rearranging, we get that

$$P(HA) = \frac{P_0(H)P(TA)}{P_0(T)}, \tag{A11}$$

which in turn can be written as

$$P(HA) = \frac{FP_0(T)}{1 - F + 2FP_0(T)} P_0(H), \tag{A12}$$

where $P(TA)$ was replaced using Eq. (A5). So, since $P(HD)$ and $P(HA)$ add up to $P_0(H)$ [and $P_0(H) + P_0(T) = 1$],

$$P(HA) = \frac{FP_0(T)}{1 - FP_0(H)} P(HD). \tag{A13}$$

Combining this with the fact that

$$P(TA) = \frac{FP_0(T)}{1 - FP_0(H)} P(TD), \tag{A14}$$

[as a result of Eqs. (A4) and (A10)], we have that

$$P(HA \text{ or } TA) = \frac{FP_0(T)}{1 - FP_0(H)} P(HD \text{ or } TD). \tag{A15}$$

Elga's footnote 8 ends with the equation corresponding to Eq. (A15), but the argument can be spelled out by noting that since (H or T) is true, Eq. (A15) can be rewritten as

$$P_{\text{inferred}}(A) = \frac{FP_0(T)}{1 - FP_0(H)} P_{\text{inferred}}(D). \tag{A16}$$

In this form we can see immediately that if we were to assume that $F = 1$, we would recover Elga's result, that is, $P_{\text{inferred}}(A) = P_{\text{inferred}}(D)$. More generally, however, since $P_{\text{inferred}}(D) + P_{\text{inferred}}(A) = 1$, Eq. (A16) implies that

$$P_{\text{inferred}}(A) = \frac{F P_0(T)}{1 - F + 2F P_0(T)}, \tag{A17}$$

in agreement with Eq. (10).

References

Birch, J. (2013). On the 'Simulation Argument' and Selective Scepticism. *Erkenntnis, 78*, 95–107. https://doi.org/10.1007/s10670-012-9400-9

Bostrom, N. (2003). Are We Living in a Computer Simulation? *Philosophical Quarterly, 53*, 243–255. https://doi.org/10.1111/1467-9213.00309

Bradley, D. (2007). Bayesianism and Self-Locating Beliefs *or* Tom Bayes meets John Perry. Doctoral Dissertation (Stanford University, CA, USA).

Bradley, D. (2011). Confirmation in a Branching World: The Everett Interpretation and Sleeping Beauty. *British Journal for the Philosophy of Science, 62*, 323–342. https://doi.org/10.1093/bjps/axq013

Carroll, S. M. & Sebens, C. T. (2015). Many Worlds, the Born Rule, and Self-Locating Uncertainty. Updated version of a paper in: *Quantum Theory: A Two-Time Success Story. Yakir Aharonov Festschrift.* D. C. Struppa and J. M. Tollaksen (eds.) Milan: Springer, (2014), pp. 157–169. https://doi.org/10.1007/978-88-470-5217-8

Dorr, C. & Arntzenius, F. (2017). Self-Locating Priors and Cosmological Measures. In: *The Philosophy of Cosmology.* K. Chamcham, J. Silk, J. D. Barrow, and S. Saunders (eds.) Cambridge: Cambridge University Press, pp. 396–428. https://doi.org/10.1017/9781316535783

Elga, A. (2004). Defeating Dr. Evil with Self-Locating Belief. *Philosophy and Phenomenological Research, LXIX*(2), 383–396. https://doi.org/10.1111/j.1933-1592.2004.tb00400.x

Isaacs, Y., Hawthorne, J., & Russell, J. S. (forthcoming). Multiple Universes and Self-Locating Evidence. *Philosophical Review*. PhilArchive. https://philarchive.org/rec/ISAMUA-2

Meacham, C. J. G. (2016). Ur-Priors, Conditionalization, and Ur-Prior Conditionalization. *Ergo, 3*(17), 444–492. https://doi.org/10.3998/ergo.12405314.0003.017

Sebens, C. T., & Carroll, S. M. (2018). Self-locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics. *British Journal for the Philosophy of Science, 69*, 25–74. https://doi.org/10.1093/bjps/axw004

Weatherson, B. (2005). Should we Respond to Evil with Indifference? *Philosophy and Phenomenological Research LXX*(3), 613–635. https://doi.org/10.1111/j.1933-1592.2005.tb00417.x

Wilson, A. (2017). The Quantum Doomsday Argument. *British Journal for the Philosophy of Science, 68*, 597–615. https://doi.org/10.1093/bjps/axv035