PLEASE DO NOT SHARE:

An Explanation of the Monty Hall Problem

John Wilcox

A Description of the Monty Hall problem

The Monty Hall problem is one of the most well-known psychological brainteasers. As one cognitive scientist remarks, it is "the most expressive example of *cognitive illusions* or *mental tunnels* in which even the finest and best-trained minds get trapped" (Piattelli-Palmarini, 1994, pp. 161). There are different versions of the problem, but a fairly standard one is as follows:

> Suppose you are on a gameshow where a prize is randomly placed behind one of three doors: door A, door B and door C. Behind the other two doors are goats. You do not know which of the three doors conceals the prize. You are asked to select a door, although that door remains closed for the time being. Monty Hall, the game show host, knows where the prize is, and he will then open one of the other doors you did not chose to show that it concealed a goat. If the door you first selected conceals the prize, he will open one of the other two doors at random. If the door you first selected does not conceal the prize, and one of the other two doors does, then he will open the one unselected door that does not conceal the prize.
>
> So suppose you play the game, selecting a door, and then Monty Hall opens one of the other doors. For example, suppose you select door A, and then Monty Hall opens door C to show you a goat.
>
> Here is the key question: should you switch doors from your initial door (door A) and instead opt to open the other unopened door (door B)?

The puzzle arises because, according to academic consensus, you should switch doors: it is more probable that door B conceals the prize.

Explanation of the Correct Response and the Law of Likelihood

The reason you should switch doors is explained by standard probability theory, the mathematics of which is uncontroversial. To explain the mathematics clearly, though, we first need to clarify some probabilistic concepts and notation.

Let $P(X)$ stand for the initial probability of some door X concealing the prize. For example, $P(B)$ is the initial probability that door B conceals the prize, and it is 1/3, just like the probabilities that the other doors conceal the prize. So we have three hypotheses: A, that door A conceals the prize; B, that door B conceals the prize; and C, that door C conceals the prize. Each hypothesis has a probability of 1/3, so $P(A) = P(B) = P(C) = 1/3$.

Now, we are ultimately interested in the probability of these three hypotheses *after we receive the evidence that door C was opened and concealed a goat*. So let the symbol '*c*' represent the proposition that door C was opened by Monty Hall and that it concealed only a goat. In this case, we are interested in $P(B \mid c)$—that is, the probability that door *B* conceals the prize given that door C was opened as such.

According to standard probability theory, this probability can be calculated with Bayes' theorem (which we will soon see below). Bayes' theorem states that this probability is calculable via a ratio involving two kinds of probabilities. One of these is a *prior probability*, the probability for a hypothesis prior to receipt of some evidence—such as the evidence that door C is opened. In this case, the prior probabilities are $P(A) = P(B) = P(C) = 1/3$.

The other kind of probability is called a *likelihood*, and such probabilities are the focal subject of this paper. A likelihood has acquired a specific and technical meaning in the statistical and philosophical literature: it refers to the probability of some evidence given some hypothesis (Hacking, 2016; Hawthorne, 2018; Bandyopadhyay, 2011). This usage of the term departs from convention: the common public—and many psychologists—talk of a "likelihood" as though it is interchangeable with any kind of "probability", but this is not the case in the probabilistic literature. In this sense, a likelihood is not to be confused with the probability of a hypothesis given some evidence (a confusion known as the *confusion of the inverse*). To illustrate the difference between a probability and its inverse, consider the probability of being a human given that one is a doctor. The probability of being a human given that one is a doctor is 100% (since all doctors are humans). But the probability of being a doctor given that one is a human is much lower (since many humans are not doctors). In this sense, a likelihood is the probability of the *evidence given that the hypothesis is true*, and *not* the probability that the hypothesis is true given that evidence.

Let us now apply these concepts to the case of the Monty Hall problem.

Recall that we are interested in $P(B \mid c)$—the probability that door *B* conceals the prize given that door C was opened and concealed a goat. According to Bayes' theorem, this probability is specified by the equation whose right-hand terms are prior probabilities and likelihoods:

(1)

$$P(B|c) = \frac{2}{3} = \frac{P(c|B)P(B)}{P(c|A)P(A) + P(c|B)P(B) + P(c|C)P(C)}$$
$$= \frac{1 \times \frac{1}{3}}{0.5 \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}}$$

Of course, the reader does not need to know everything about this theorem for the purposes of this paper, but they do need to note three points.

The first point is that, according to Bayes' theorem, the probability that door B conceals the prize is $\frac{2}{3}$, twice as probable as door A concealing the prize.

The second point, and one which the public is often unaware of, is that Bayes's theorem delivers the *correct* verdict that $P(B|c) = \frac{2}{3}$. There are many arguments for why this is the case (Rosenhouse, 2009), but perhaps the most straight-forward argument is that, when running numerous computer simulations of the Monty Hall problem as described above, door B will contain the prize approximately $\frac{2}{3}$ of the time—not $\frac{1}{2}$ of the time, as many think. In appendix X, we provide some code so that the reader can themselves run simulations of the Monty Hall problem.

In any case, the academic consensus is that Bayes' theorem delivers the right verdict, but the question arises as to how it does this, and this is the third point which the reader needs to note: Bayes' theorem delivers the verdict that it does largely because of the *likelihoods*. Notice that, in equation (1), the values of the prior probabilities of the hypotheses are all the same at $\frac{1}{3}$. The hypotheses differ *only* in that they make the evidence more or less probable relative to each other. The hypothesis that door C conceals the prize gives a probability of 0% that door C would be opened. Consequently, Bayes' theorem gives the probability that door C conceals the prize a probability of 0% conditional on door C being opened. The hypothesis that door A conceals the prize gives a probability of 50% (or 0.5 in decimal notation) that door C would be opened. In contrast, the hypothesis that door B conceals the prize gives a probability of 100% (or 1 in decimal notation) that door C would be opened, twice that of the likelihood given by door A. As a result, the probability that door B conceals the prize given that door C is opened ultimately becomes twice the probability of door A: $\frac{2}{3}$ compared to $\frac{1}{3}$.

This, then, illustrates a general and uncontroversial theorem about probabilistic reasoning: the *law of likelihood*. There are various statements of the law (Hacking, 2016, chap. 5; Hawthorne, 2018), but the following is a relatively simple version for our purposes. Let h_1 and h_2 stand for two distinct and mutually exclusive hypothesis, meaning that they cannot be simultaneously true. Let e stand for some evidence. Furthermore, let $0 < P(h_1) < 1$ and $0 < P(h_2) < 1$, meaning that neither hypotheses have a probability of 0% or 100%. Then, the law of likelihood specifies that:

(2)

$$\text{If } P(e|h_1) > P(e|h_2), \text{ then } \frac{P(h_1|e)}{P(h_2|e)} > \frac{P(h_1)}{P(h_2)}$$

Proof of the theorem can be found in appendix D.

Informally put, what this means is that if one hypothesis h_1 makes the evidence e more probable than another hypothesis h_2—or, in other words, if the likelihood of the evidence given

h_1 is greater than the likelihood of the evidence given h_2—then the evidence *raises* the probability of h_1 relative to h_2 and, by implication, it *lowers* the probability of h_2 relative to h_1.[1]

To apply this to our example in the Monty Hall problem, it is more probable that door C would be opened if door B concealed the prize than if door A concealed the prize, so $P(c|B) > P(c|A)$. Consequently, the evidence *raises* the probability of door B concealing the prize relative to door A concealing the prize, since the probability for door B goes from $\frac{1}{3}$ to $\frac{2}{3}$ while the probability for door A remains at $\frac{1}{3}$. Of course, this does not lower the probability of door A concealing the prize in absolute terms: it was $\frac{1}{3}$ both before and after door C was opened. Instead, it lowers the probability of door A concealing the prize *relative* to door the probability of door B concealing the prize: both probabilities were equal, but now one probability is twice the size of the other.

It is important to dispel two misconceptions about the law of likelihood. The first is that it is a specifically *Bayesian* phenomenon—that only Bayesians endorse the principle, or that frequentists and other non-Bayesians do not adopt the principle. This is incorrect: Bayes' theorem is universally accepted in probability theory (Weirich, 2011, 234). What differentiates Bayesians and non-Bayesians is not how Bayes' theorem relates one probability to other probabilities. Rather, it is about such matters as to how probability statements are to be interpreted, how probabilities relate to degrees of belief and how Bayes' theorem is to be used. So the points that are made in this paper do not concern only Bayesian reasoning; they concern *all* probabilistic reasoning. The second misconception is that this principle only applies in toy cases, like the case of Monty Hall, but not in real world problems. Again, this is incorrect. It applies to any case where it is meaningful to talk about the probability of the evidence given competing hypotheses, and such cases arguably arise in many parts of life: in medical diagnosis, in scientific contexts and in others.

Regardless, the main point is that Bayes' theorem delivers the correct verdict that it does because of facts about likelihoods, and this theorem entails the law of likelihood.

The Psychology of the Monty Hall Problem: Causes of Incorrect Responses

The preceding discussion gives an account of the correct way to reason about the Monty Hall problem, but we now turn to consider how people actually do reason about the Monty Hall problem.

As alluded to earlier, the correct response to the problem is to switch doors, but studies have generally reported that most participants give incorrect responses and prefer to stay with their initial selection. For example, Burns and Wieth (2004) surveyed thirteen studies of the participants

[1] Note that some philosophers, such as Hawthorne (2018), speak of the "likelihood of the evidence", while others, such as Nola (2013), speak of the "likelihood of the hypothesis" to refer to the same thing. I have followed Hawthorne as I think this is a less confusing way of speaking about likelihoods.

responses to the standard Monty Hall problem. They found that switch rates ranged from 9% to 23%, with a mean of 14.5% (*SD* =4.5). As they note, the consistency of low switch rates is also remarkable given that the studies varied with respect to problem wording, methods of presenting the problem and the cultures and languages of the participants. Furthermore, even educated participants have faltered on the problem. Marilyn vos Savant initially popularized the problem by writing on it in the *Parade* magazine in 1990, and she found that 65% of respondents writing with university addresses disagreed with the correct solution to the problem (vos Savant, 1997). Additionally, Schechter (1998, 108-109) relates how even Paul Erdõs, a renowned mathematician of the 20th century, initially disagreed with the correct solution to the problem. Apparently, he only changed his mind after a computer simulation of the problem convinced him it was correct.

So what then are the causes of such responses? Two comprehensive literature reviews have recently been published (Saenen et al., 2018; Tubau et al., 2015), and two prominent causes are discussed by Tubau et al. (2015): 1) emotional-based choice biases and 2) cognitive limitations in understanding and representing probabilities. The first cause relates to how participants are averse to switching from their first choice. Some suggest this aversion this arises from an illusion that their first choice can favorably influence the outcome (Granberg & Dorr, 1998) or alternatively from a desire to reduce regret, a regret which is higher if one loses after switching than losing after sticking with their initial choice (Petrocelli et al., 2011). The second cause relates to the frequent observation that participants commonly believe that the probabilities of the two outcomes are equal given the observation that two options remain (De Neys & Verschueren, 2006; Tubau et al., 2003).

Why Some Popular Solutions Do Not Work

A range of interventions have been tested to determine whether they improve reasoning in the Monty Hall, and these are reviewed comprehensively by Saenen et al. (2018).

Of course, there are many putative "solutions" to the Monty Hall problem—that is, attempts to 1) give the correct answers and 2) explain why that answer is the correct one. So far, the only one which delivers the uncontroversially correct answer is the solution with Bayes' theorem. But other putative solutions are also popular. While it is impossible to review all of them here, we argue that there are theoretical reasons to think that two particular approaches to the problem are unreliable.

We will first outline these two approaches and then examine how they are unreliable.

Possible Models Approach

One approach is a particular mental models approach discussed by Krauss & Wang (2003a) and Tubau et al., (2003). Their solution first involves entertaining various "possibilities" about where the prize might be (or various "mental models", as they say). Then, one calculates the frequency with which switching doors would win the prize among those possibilities. In particular, these are

the two sets of mental models which Krauss and Wang (2003a) present in their solution to the Monty Hall problem. These are represented in their figures and tables below.

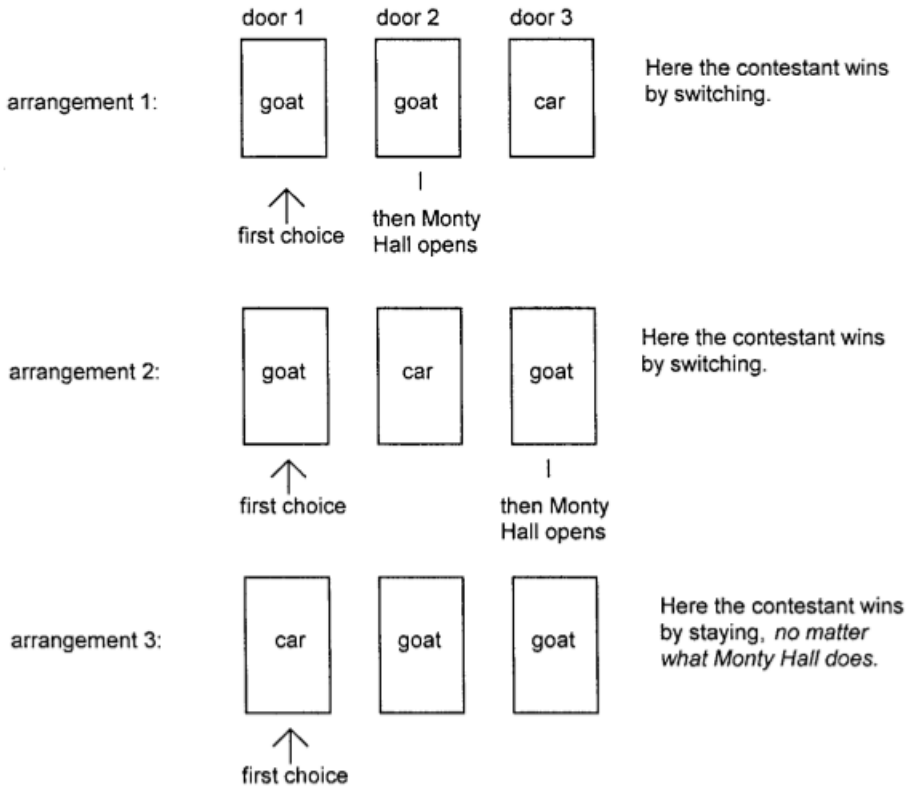


Figure 1. Explanation of the solution to the Monty Hall problem: In two out of three possible car–goat arrangements the contestant would win by switching; therefore she should switch.

Table 1
Mental Model Representation of the Monty Hall Problem

Mental model	Door 1 (chosen door)	Door 2	Door 3
1	car	open	
2	car		open
3		car	open
4		car	open
5		open	car
6		open	car

Note. Based on mental models from Johnson-Laird et al. (1999).

The idea here is that one can both obtain the correct probabilities *and* understand why they are the correct probabilities by considering the frequency with which switching yields the prize among either of these two sets of mental models.

Tubau and Alonso (2003) also report a similar experiment that involves getting participants to consider various possibilities and count the frequency with which switching yields a favorable result.

That, then, is one prominent approach to solving the problem. For ease of reference, we call it the *possible models* approach.

Probability Accrual Approach

There is also another way to approach the problem. Rather than entertaining all the possibilities, this approach focuses on the fact that the non-selected doors have conceal the prize two thirds of the time—or, equivalently, these doors have a 2/3 prior probability of concealing the prize. Of course, once Monty Hall opens one of the unselected doors, it becomes clear that that particular door does not conceal the prize. The special part of the approach, however, is to claim that the probability of 2/3 then *accrues* to the other unselected door. Put simply, the reasoning is this: the unselected doors conceal the prize with a prior probability of 2/3, but once one of these doors is opened, there is then a 2/3 probability that the only unselected door conceals the prize. Call this the *probability accrual approach*.

Such reasoning seems to be endorsed by Tubau et al. (2015) as a promising approach. They discuss why humans are vulnerable to the "equiprobability illusion" in the Monty Hall problem (that is, the illusion that it is equally probable that both unopened doors conceal the prize):

> In particular, susceptibility to the illusion is caused by a weak representation of the facts that: (a) the non-selected doors will hide the prize 2 out of 3 times, (b) among the non-selected doors it is certain that at least one is null, and (c) this null option will always be eliminated.(Tubau et al., 2015, 8)

The implication of this statement is that if participants are aware of conditions (a)-(c), then they will not be susceptible to the equiprobability illusion. This seems to implicitly reflect the probability accrual approach: if participants were sufficiently aware that the unselected doors conceal the prize two thirds of the time and that the opened door does not conceal the prize, then they would not think the remaining doors are equally likely, but would instead think the unselected door conceals the prize with a probability of 2/3.

Similar reasoning has been reported by participants in our pilot experiments. One such participant studied the Monty Hall problem before the experiment. They correctly answered that the probability that door B concealed the prize is 2/3. Their justification for this probability appeared to reflect the probability accrual approach:

> Essentially by choosing door A and switching, I'm choosing both doors B and C. It's just that I know one of the two won't have the prize. But that means switching still increases my chances of winning from 1/3 to 2/3

The Problem with These Approaches: Safety

The problem with these two approaches, however, is that they lack *safety*.

Safety is a concept which we borrow from discussions in epistemology. There, safety is occasionally advocated as *necessary condition* for a belief to qualify as knowledge. In brief, an agent has a safe belief in some proposition *p* when the following condition holds:

> If the agent were to believe that *p,* then *p* would not be false. (Ichikawa & Steup, 2018; Sosa, 1999)

In epistemology, some claim that a belief does not count as knowledge if it is unsafe.

Let us consider examples of safe and unsafe beliefs, examples adapted from Bertrand Russell (1948). Suppose someone looks at a functional clock which says it is 12pm, and they then form the belief that it is 12pm. This belief is safe in the sense that if the agent believes that it is 12pm, then this it would not be false that it is 12pm. And it is safe as such because it is based on a reliable method of getting at the truth—namely, telling the time from a reliable clock. Contrast this to the following example of an unsafe belief. Suppose someone looks at a clock which says that the time is 12pm, but the clock is *broken* and so it *never* changes with the actual time. The person does not know it is broken, and they then form the belief that it is 12pm. Coincidentally, it happens to be 12pm, and so their belief is true. However, their belief is not safe: if they used this method to arrive at their belief that it is 12pm, then it is not necessarily the case that it is 12pm rather than some other time. After all, they could have looked at the clock when it was 1pm and then falsely concluded that it was 12pm. The safety of a belief is closely related to how sensitive that belief is to the truth, or how well the belief tracks the truth (although sensitivity and truth-tracking have both taken on technical meanings and are surrounded with controversy in epistemology).

The concept of safety highlights the importance of using methods that get us to the truth, not by chance, but because *those ways track the truth*. A broken clock is not a good way of forming beliefs about time, even if it happens to get one to the right beliefs in some circumstances.

In this context, we claim that the aforementioned probabilistic methods are unsafe in this particular sense—they do not track the truth. More specifically, if someone were to use those methods to arrive at judgments about probabilities, then those judgments are not necessarily true. Instead, those judgments could be false.

The 90% Monty Hall Problem

We will illustrate this with the following case which we call the *90% Monty Hall problem.* This version is exactly the same as the Monty Hall problem in all respects except this: if you select a given door and it conceals the prize, then Monty Hall has a *90% probability of opening the right-*

most door that is unselected and does not conceal the prize. In this case, if you select door A, and if door A conceals the prize, then Monty Hall is going to open door C with a 90% probability or door B with a 10% probability. Hence, the likelihoods change. Suppose you select door A and Monty Hall opens door C. By Bayes's theorem, it then follows that door A and door B have a near equal probability of concealing the prize:

$$P(B|c) = \frac{P(c|B)P(B)}{P(c|A)P(A) + P(c|B)P(B) + P(c|C)P(C)}$$
$$= \frac{.9 \times \frac{1}{3}}{1 \times \frac{1}{3} + .9 \times \frac{1}{3} + 0 \times \frac{1}{3}}$$
$$= \frac{9}{19} = .53$$

So we have a case where the likelihoods change, and so too do the posterior probabilities about whether doors A and B conceal the prize. In this case, it *is nearly just as rational to stay* with your choice in door A as it is to switch. And we know this via exactly the same mathematical machinery that tells us to switch in the original Monty Hall problem scenario—namely, Bayes theorem.

To reinforce this point, we ran 100,000 computer simulations where Monty Hall had a 90% chance of opening door C if door A concealed the prize. Door B then concealed the prize about 53% of the time. We have included our code in an appendix so that others can reproduce this result if need be.

Note, however, that even though the probabilities have changed, it is not obvious that one would get the right answer if they were to follow the possible models and probability accrual approaches above. In other words, they might *falsely still conclude that the probability that door B conceals the prize is 2/3.*

Consider the possible models approach. Note that *all* of the possibilities are exactly the same. Nothing has changed about the space of mental models. For that reason, the counting procedure of the possible models approach would generate exactly the same probabilities.

And consider the probability accrual approach. In the right-most Monty Hall problem, the following facts still obtain: "(a) the non-selected doors will hide the prize 2 out of 3 times, (b) among the non-selected doors it is certain that at least one is null, and (c) this null option will always be eliminated"(Tubau et al., 2015, 8). According to Tubau et al. (2015, 8), awareness of such facts would stop participants from being susceptible to the "equiprobability illusion".

However, in the 90% Monty Hall problem, there is a *near* equal probability that the doors conceal the prize, but it is doubtful that the probability accrual approach would by itself help them to realize this.

Of course, one might think that even if the possible models and probability accrual approaches do not get the right answer in the 90% Monty Hall problem, perhaps we could modify or supplement them so that they do.

That may be a possibility, and we make no claims about its prospects. Instead, we merely claim that, *as they currently stand*, the possible models and probability accrual approaches do not help participants get to such judgments, and we present experimental findings later on that suggest this is the case.

Further, we claim that any safe approach to the Monty Hall problem should be sufficiently sensitive to the *likelihoods*: if the likelihoods change, then so will the posterior probabilities that are recommended by the approach. We later outline one approach that is sensitive as such: the mental simulations approach. And we provide some experimental evidence that it improves reasoning in the Monty Hall problem—by improving both the correctness of participants' answers and their understanding of why those answers are correct.

The Mental Simulations Approach to Probabilistic Reasoning

This section outlines what we call the 'mental simulations' approach. The approach has two purposes:

1) Facilitating the *calculation* of the correct probabilities
2) Providing an *intuitive justification* of the correct probabilities

By 2), we essentially mean helping humans to obtain an *intuitive grasp* of *why* Bayes's theorem delivers the right result or *why* the law of likelihood is correct. I will note how the approach achieves these purposes once it has been outlined.

The approach relies on two insights. The first is that computer simulations have successfully convinced skeptics of the correctness of the verdict delivered by Bayes's theorem; take the aforementioned example of mathematician Paul Erdõs, for example. The second is that reasoning in the population has occasionally been improved when probabilistic reasoning is carried out in terms of natural frequencies instead of probabilities (Hoffrage et al., 2015).

The mental simulations approach then combines these two insights. It involves participants running "mental" simulations of the probabilistic set up to thereby convert the probabilities into natural frequencies. Then, the probabilities of interest can be calculated by counting outcomes among these simulations. It essentially is just an attempt to apply and generalize some of the insights of Hoffrage et al. (2015).

An Example of Mental Simulations in the Monty Hall Problem

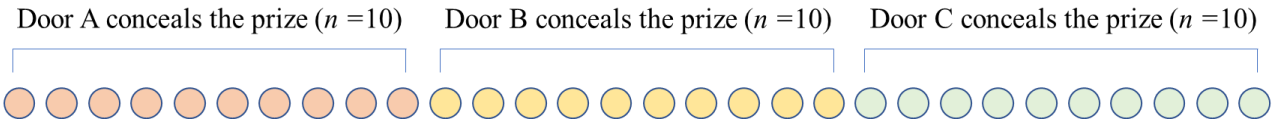
Let us go through this with an example before outlining the approach in more abstract generality. Consider the Monty Hall problem. Let us run 30 mental simulations, that is, let us imagine 30 situations in which the Monty Hall problem is played (and I will explain why I have chosen the

number 30 shortly). To help visualize this, let us denote these 30 simulations with circles, imagining that each circle represents a scenario where the Monty Hall problem is played out.

30 'mental' simulations of the Monty Hall problem

Now we know that the prior probability of any door containing the prize is 1/3, so $P(A) = P(B) = P(C) = 1/3$. The first step, then, is to translate these probabilities into natural frequencies among the 30 simulations—that is, to divide up these simulations so that the proportion of times that a prize is in a location corresponds to the probability that that location would conceal the prize. For example, since each door has a 1/3 probability of concealing the prize, a given door conceals the prize in a third of simulations.

Outcomes of the simulations proportioned to prior probabilities

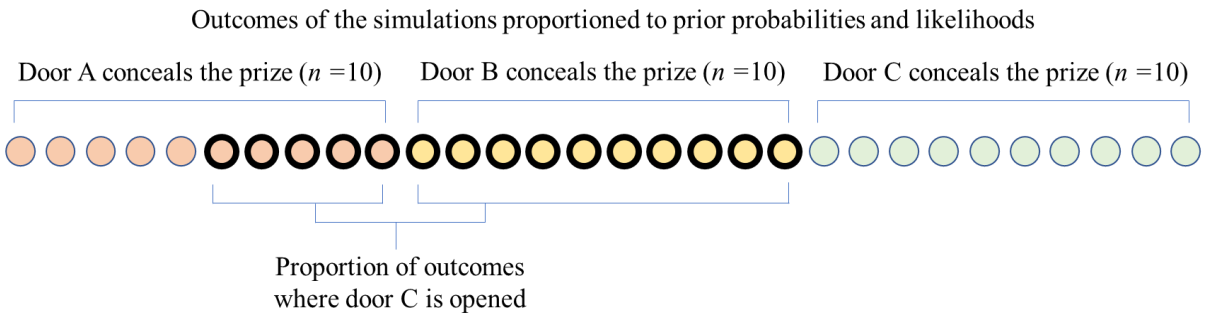


So now our simulations correspond to the prior probabilities: each outcome has a prior probability of 1/3, and each outcome is true in 1/3 of the simulations.

Now, we are going to translate the likelihoods of the evidence we observed into natural frequencies among these simulations. Recall that our evidence is *c*: that door C was opened after door A was selected, and it concealed a goat. Recall also the likelihoods:

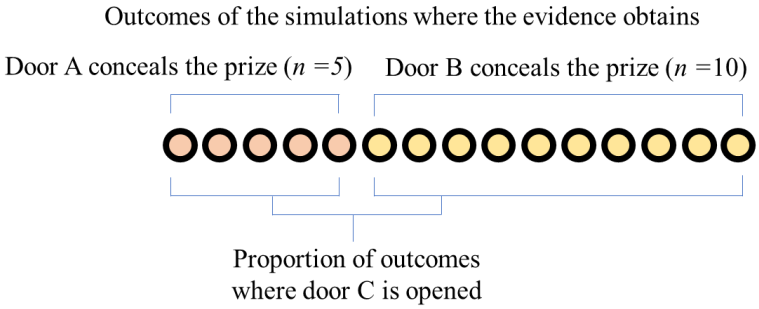
- $P(c|A) = \frac{1}{2}$, meaning that door C has a 50% probability of being opened if door A concealed the prize.
- $P(c|B) = 1$, meaning that door C has a 100% probability of being opened if door B concealed the prize.
- $P(c|C) = 0$, meaning that door C has a 0% probability of being opened if door C concealed the prize.

We then translate these probabilities into proportions of outcomes where the evidence *c* is true. For example, door C has a 50% probability of being opened if door A concealed the prize, so door C will be opened in 50% of the simulations where door A conceals the prize. Likewise, we then note the proportion of the other simulations where door C would be opened given the respective outcomes.



The outcomes where the evidence *c* is true are denoted by the circles with bolded outlines. We can see that the proportion of outcomes where the evidence is true is proportional to the likelihoods.

So now suppose we want to calculate the probability of door B concealing the prize given the evidence. To do this, we simply eliminate the simulations where the evidence does not obtain, and we then determine the proportion of remaining simulations where door B conceals the prize:



We can see that the remaining proportion of outcomes where door B conceals the prize is $\frac{10}{15}$, or $\frac{2}{3}$. This aligns precisely with the verdict of Bayes's theorem: the probability that door B conceals the prize given that door C is opened is $\frac{2}{3}$, so $P(B|c) = \frac{2}{3}$.

A Generalized Characterized of the Mental Simulations Approach

That, then, is an example of how to implement the approach.

Let us now characterize it in general detail. To calculate the probability of a hypothesis (or an outcome) given some evidence:

1. Generate simulations:
    - Imagine *n* number of simulations (and I will discuss what value *n* may take shortly)
2. Proportion according to priors:
    - For each possible outcome, make the proportion of simulations where that outcome is true correspond to the prior probability of that outcome
3. Proportion according to likelihoods:

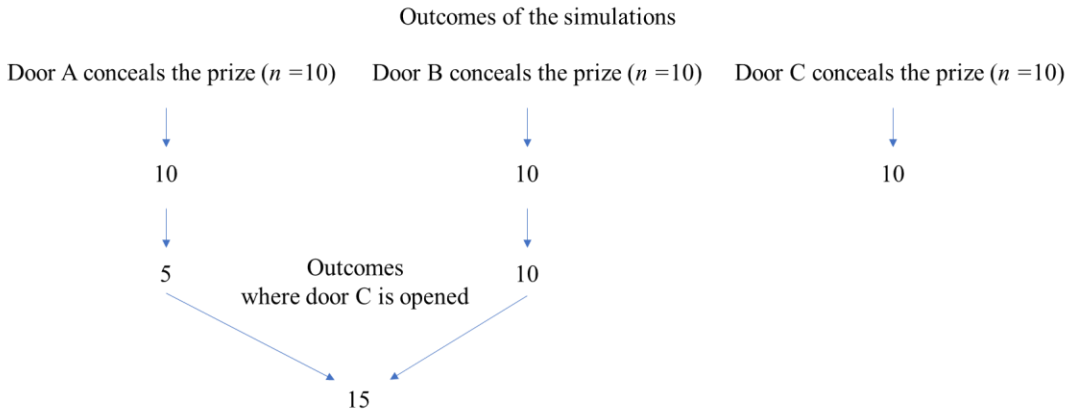
- For each set of simulations corresponding to a given outcome, make the proportion of simulations where the evidence obtains correspond to the likelihood of that evidence given that outcome

4. Eliminate irrelevant simulations:
    - Eliminate the outcomes where the evidence does not obtain
5. Calculate probabilities:
    - Determine the proportion of simulations where a particular outcome is true; this is the probability of that outcome given the evidence

There are three points to note about this approach.

The first is that any choice of *n* is appropriate so long as it allows for the appropriate division of simulations by the probabilities. One will not be able to make door A conceal the prize in a third of all simulations if there are four simulations altogether: one cannot divide 4 by a third in the approach. But one could carry out the above with simulations of 6 or any multiple of it.

The second is that the depiction of the simulations does not matter either. For instance, the simulations could be represented with numbers instead of circles corresponding to those numbers, like the following:



A third point to note is that the approach makes the idealizing assumption that the proportions of scenarios correspond perfectly to the relevant probabilities. In real life situations and in experiments, this assumption is often false: if a coin is tossed 10 times, the proportion of times it lands heads will often not be exactly 50%, even if the probability of heads is exactly 50%. Regardless, this assumption is harmless in the mental simulations approach, and it facilitates the calculation of proportions and probabilities. The assumption also enables the simulation of finite scenarios to reflect the proportion of outcomes of a probabilistic set up if they went to infinity. In other words, if you were to run the set up for an indefinitely increasing number of times, then the proportion of times that an outcome is true given some evidence corresponds to the result yielded among the finite simulations in the approach.

So that is an outline of the approach.

Strengths and Limitations of the Mental Simulations Approach

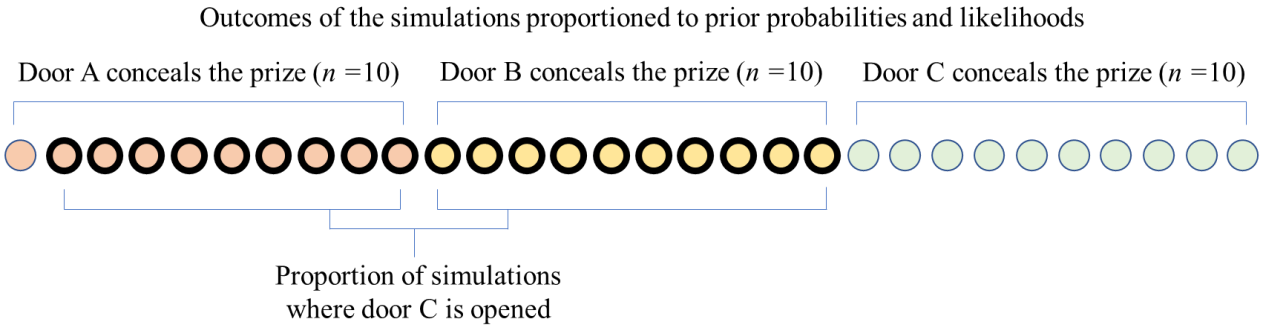
This approach has four strengths.

One strength is that, if carried out correctly, in ensures correct reasoning in accordance with Bayes's theorem, and it thus eliminates neglect of likelihoods and prior probabilities. Proof of this can be found in the appendix. Consequently, it achieves its first aim of helping humans to correctly calculate probabilities.

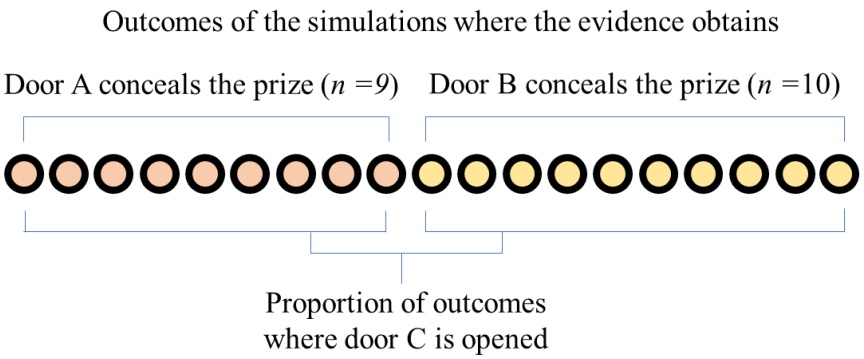
A further strength is that it can help people to understand *why* the probabilities are the correct ones. For example, from the simulations approach and the correspondence between its verdicts and what happens in the limit, we can make sense of why there is a probability of $\frac{2}{3}$ that the other door conceals the prize: because if the probabilistic set up was run an infinite number of times, then the other door would conceals the prize in two thirds of those times.

Another strength is that, as mentioned earlier, it draws on the greater facility that humans have with reasoning in terms of natural frequencies rather than probabilities, thus potentially making it easier for people to implement than calculations with Bayes's theorem.

A final strength emerges relative to alternative approaches, such as the possible models and probability accrual approaches. In particular, it can be adapted when the likelihoods take on different values. Recall the 90% Monty Hall problem, for instance. There, the probability that door C would be opened if door A concealed the prize was instead 90% rather than 50%. Then, without having to perform laborious calculations with Bayes's theorem, we can then determine the probability that door B conceals the prize: just adjust the proportion of simulations where the evidence obtains and where door A conceals the prize to 90%.



Then eliminate the simulations where door C has not been opened:



We can then see that the probability that door B conceals the prize is $\frac{9}{19}$, again the probability recommended by Bayes's theorem. Contrast this to the possible models approach. The possible models approach calculates only the proportion of times that switching would win the prize among a particular set of possible outcomes, and this proportion among this set is $\frac{2}{3}$ regardless of whether the likelihood $P(c|A)$ is 50% or 90% (Krauss, & Wang, 2003).

But the mental simulations approach also has its limitations. For one, it requires some training: this is not a technique that would come to someone naturally without any instruction. For another, it may be difficult to run simulations in cases where the probabilities are very small or go to many decimal places, since it would require mental operations with a potentially large number of simulations.

In any case, every approach has its strengths and weaknesses, and perhaps the strengths of this approach outweigh its limitations. Regardless, I think the strengths of the approach are such that it merits further testing.

Mental Simulations and the Mental Models Approach

One might wonder how the mental simulations approach relates to a prominent theory in the psychology of thinking: the mental models approach.

The mental models approach may mean different things to different authors, but we can compare the mental simulations approach to at least some ways of understanding mental models. Like mental models theory, the mental simulations approach involves generating mental entities that stand in for ways that the world might possibly be—or ways that it possibly could have been before we received some evidence about it. We call then call these mental entities "simulations", and they resemble mental models in this respect.

However, unlike the mental models theory (at least on some understandings of it), these simulations do not always denote *distinct* possibilities, nor is each possibility regarded as *equiprobable*. For example, consider how Johnson-Laird (2012) construes the mental models theory. He claims that, according to the theory, "Each mental model represents a *distinct*

possibility" (our emphasis, 137), and "Each mental model represents an *equiprobable* possibility unless there are reasons to the contrary" (our emphasis, 144).

The mental simulations approach is not like this. Two or more simulations may represent the *same* possibility: above, for example, we have 10 simulations representing the same possibility where door B conceals the prize and door C is opened. Furthermore, not every possibility is equally probable: the possibility where door B conceals the prize after door C is opened is more probable than the possibility where door A conceals the prize after door C is opened. And the reason for this is that probabilities are calculated differently: probabilities are not calculated by counting possibilities. Instead, simulations are generated from probabilities, the simulations inconsistent with the evidence are then eliminated, and the probabilities are then counted from these simulations—not from the possibilities which they denote.

To that extent, then, the mental simulations approach differs from at least some prominent ways of understanding the mental models theory.

In saying that, one might think that the mental simulations approach involves mental models *if* mental models are understood in a different way.

In any case, though, the mental simulations approach is not intended to be a descriptive claim about human psychology. It does not intend to *describe* how humans *actually do reason*, unlike some theories involving mental models. Instead, it is a claim about one way that humans *can* reason—especially if they receive some training. And we claim at most that this approach can improve reasoning and avoid particular biases—most notably, the likelihood neglect bias.

The purpose of this study is then to provide support for this claim, and we conducted two experiments to this end.