DOES THE PRINCIPAL PRINCIPLE NEED SUPERBABIES?

MICHAEL TITELBAUM

Please do not circulate or cite without permission.

One often hears David Lewis’s Principal Principle informally glossed as something like the following:

The Principal Principle (Quick Gloss). An agent with no indamissible evidence should set her credence in an event equal to what she takes to be its chance.

This quick gloss captures the intuition behind the principle, and is rough-and-ready enough for many applications. However, when we go back and look at the Principal Principle as Lewis actually formulated it, we find:

The Principal Principle. Let $C$ be any reasonable initial credence function. Let $t$ be any time. Let $x$ be any real number in the unit interval. Let $X$ be the proposition that the chance, at time $t$, of $A$’s holding equals $x$. Let $E$ be any proposition compatible with $X$ that is admissible at time $t$. Then

$C(A \mid X \& E) = x$

This formulation has a number of bells and whistles not found in our quick gloss. Two of these bells (whistles?) make the Principle much more broadly applicable than the quick gloss suggests.

First, the Principal Principle involves conditional credences instead of the unconditional credences suggested by the quick gloss. This means that the Principal Principle can be used to set our credences even if aren’t certain what the chance of $A$ is. Lewis offers this example about a coin flip:

Suppose you are not sure that the coin is fair. You divide your belief among three alternative hypotheses about the chance of heads, as follows.

You believe to degree 27% that the chance of heads is 50%.
You believe to degree 22% that the chance of heads is 35%.
You believe to degree 51% that the chance of heads is 80%. (266)

The Principal Principle yields a credence conditional on each of the possible chance values for heads you consider, then the probability calculus allows you to combine those conditional credences by weighting each according to your credence that that chance assignment is correct.

1All references in this paper are to page numbers in Lewis’s “A Subjectivist’s Guide to Objective Chance,” Studies in Inductive Logic and Probability, Volume 2, edited by Richard C. Jeffrey (Berkeley, Ca.: University of California Press). Throughout this paper I’ve altered Lewis’s notation by inserting ampersands he usually omits. Also, my apologies for quoting at greater length in some places than might be strictly necessary — I’m a bit of a sucker for Lewis’s prose.
Second, the Principal Principle time-indexes both the chance of A and the admissibility criterion for E. The quick gloss might seem applicable only to an agent who knows what the chance of the event is at the time she is assigning her credence. But the time-indexing in the Principal Principle allows it to apply to an agent who, for instance, does not know the current chance of an event but knows what its chance was some time ago.

For example, suppose a coin was flipped at noon. Marge was allowed a very careful inspection of the coin and the flipping apparatus at 11:59am, enough to satisfy her that the chance at noon that the coin would land heads was 1/2. However, Marge was banished from the flipping room just before the coin was flipped, and no evidence has come out of the room since then. It is now 1pm.

We can apply the Principal Principle to Marge’s situation by indexing it to noon. Let E be the total evidence Marge has learned since noon. Clearly E is admissible for heads at noon. Being explicit about the time-indexing, the Principal Principle tells us that

\[ C_{1\text{pm}}(\text{Heads} \mid (Ch_{\text{noon}}(\text{Heads}) = 1/2) \& E) = 1/2 \]

This reflects the fact that Marge’s credence at 1pm that the coin came up heads should be 1/2.

It’s important that the Principal Principle can be applied this way. If the Principal Principle could be applied only to your current knowledge about current chances, the fact that a chance event had occurred already but occurred in such a way that you could gain no evidence about its outcome would prevent your using the Principal Principle and what information you did have about its chances to set your credences. The Principal Principle would not be able to reproduce a simple inference involving chances that we make frequently and naturally; it would hardly live up to Lewis’s billing of including “all that we take ourselves to know about chance.” (276)

However, there is an element of the Principal Principle not in the quick gloss that might seem a bit mysterious: the description of C as “any reasonable initial credence function.” In a moment I will discuss what Lewis might mean by an “initial” credence function. But first let’s see why he couldn’t just leave that word out. Consider the following principle:

The Principal Principle (Defective). Let C be any reasonable credence function. Let t be any time. Let x be any real number in the unit interval. Let X be the proposition that the chance, at time t, of A’s holding equals x. Let E be any proposition compatible with X that is admissible at time t. Then

\[ C(A \mid X \& E) = x \]

This is the same as the original Principal Principle, with the word “initial” omitted. Here’s a counter-example to the defective principle, showing why it’s defective.

Suppose that in the Marge example I am the coin-flipper. After Marge is banished, I flip the coin and see that it comes up heads. At 1pm I assign a credence of 1 to the coin’s having come up heads; clearly my credence function at 1pm is

\[^2\text{A straightforward test of this claim: You could have described to Marge at 11:59am everything that would happen to her between noon and 1pm — the entire contents of E — and she still would have believed that the chance of heads at noon was 1/2.}\]
reasonable. Now let $X$ be the proposition that the chance of heads at noon was $1/2$, and let $E$ be the total evidence Marge has learned outside the room since noon. As we said above, $E$ is admissible for the chance of heads at noon. So applying the defective Principal Principle time-indexed to noon, we get a requirement that for any reasonable credence function $C$, $C(\text{Heads} \mid X \& E) = 1/2$. But my credence function is reasonable, and I have $C(\text{Heads} \mid X \& E) = 1$.

This is an example of what we might call the inadmissible background problem. The Principal Principle restricts itself to admissible $E$’s because it recognizes that inadmissible evidence can rightly skew our credences away from the chances. But an agent who already possesses inadmissible evidence will have worked it into the background of his current credence function, and therefore (rightly) set a credence of $A$ not equal to $x$ even conditional on $X$ and some other piece of evidence $E$ that is admissible.

Because the defective Principal Principle applies to any reasonable credence function, it is susceptible to counter-examples in which agents have inadmissible background evidence. I believe that Lewis’s restriction of the Principal Principle to reasonable initial credence functions is meant to avoid the inadmissible background problem.

Lewis does not tell us what he means by an “initial” credence function. He does, however, tell us what it is for an initial credence function $C$ to be “reasonable” (I am going to start numbering passages from the Lewis for ease of reference):

[1] In general, $C$ is to be reasonable in the sense that if you started out with it as your initial credence function, and if you always learned from experience by conditionalizing on your total evidence, then no matter what course of experience you might undergo your beliefs would be reasonable for one who had undergone that course of experience. I do not say what distinguishes a reasonable from an unreasonable credence function to arrive at after a given course of experience. We do make the distinction, even if we cannot analyze it; and therefore I may appeal to it in saying what it means to require that $C$ be a reasonable initial credence function. (268)

Notice the antecedent “if you started out with it as your initial credence function...” in the first sentence. This phrase seems to offer hope that an initial credence function might be something an actual agent could possess at the beginning of a credence development. But that hope fades in the light of this passage:

[2] Given a time $t$ and a world $w$, let us write $P_{tw}$ for the chance distribution that obtains at $t$ and $w$. For any proposition $A$, $P_{tw}(A)$ is the chance, at time $t$ and world $w$, of $A$’s holding. Let us also write $H_{tw}$ for the complete history of world $w$ up to time $t$: the conjunction of all propositions that hold at $w$ about matters of particular fact no later than $t$. Let us also write $T_{w}$ for the complete theory of chance for world $w$: the conjunction of all the conditionals from history to chance...that hold at $w$. We have

The Principal Principle Reformulated. Let $C$ be any reasonable initial credence function. Then for any
time $t$, world $w$, and proposition $A$ in the domain of $P_{tw}$

$$P_{tw}(A) = C(A \mid H_{tw} \& T_w)$$

In words: the chance distribution at a time and a world comes from any reasonable initial credence function by conditionalizing on the complete history of the world up to the time, together with the complete theory of chance for the world. (276-7, emphases in original)

Notice that both Lewis’s reformulated Principal Principle and his elaboration of it in words immediately after explicitly state that the reformulation holds of any reasonable initial credence function. In order for the reformulation to hold, none of the reasonable initial credence functions can be susceptible to the inadmissible background problem for any of the chance events covered by the reformulation. The inadmissible background problem arises whenever a credence function assigns a credence of 1 to the occurrence of a particular chance event whose chance was at some time less than 1. Thus none of the reasonable initial credence functions can assign a credence of 1 to the outcome of any chance event in the history of the world. If a reasonable initial credence function $C$ assigned a credence of 1 to the occurrence of a chance event $A$ in world $w$ whose chance at some time $t$ in $w$ was less than 1, for that $A$, $w$, $t$, and $C$ it would not be the case that $P_{tw}(A) = C(A \mid H_{tw} \& T_w)$. The left-hand side of the equation would be less than 1, while the right-hand side would equal 1.

Combine that with the following sentiment of Lewis’s:

[3] A reasonable initial credence function does not reject any possibility out of hand. It assigns some nonzero credence to any consistent theory of chance, no matter how unsystematic and incompressible it is. (276)

Admittedly, this quote is taken out of context; in context Lewis is discussing the issue of whether we should consider among the options for $T_w$ only “systematic” theories of chance expressible in language. Still, the sentiment is striking, especially in combination with the restriction on initial credence functions we’ve just developed from our analysis of passage #2.

Putting all this together, I think it is highly plausible that Lewis intends a reasonable initial credence function to be the kind of thing that would be possessed by a “superbaby.”

A superbaby is an agent who on the one hand is perfectly rational, yet on the other hand has no empirical evidence about the world. Thus a superbaby does not assign a credence of 1 to any empirical propositions. A superbaby does assign conditional credences of the form “Conditional on the physical laws of world $w$ being such-and-such, and the history of $w$ evolving in such-and-such way up to time $t$, my credence that $A$ will occur is....” However, conditional is all these credences can ever be for the superbaby as long as it remains a superbaby, for as such it can never learn what the physical laws of a world actually are and what actually happens in that world’s history.

The idea, then, would be that the reasonable initial credence functions mentioned in the Principal Principle are the credence functions of superbabies. The Principal Principle describes some conditional credences a superbaby assigns. It thereby

---

3 Alan Hájek reports in conversation that “superbaby” is a term Lewis used during a course at Princeton. Neither Alan nor I know of any published material in which Lewis uses the term.
indirectly describes some unconditional credences of reasonable agents, since the unconditional credences of a reasonable agent are superbaby credences condition-
alized on that agent’s total evidence.

The fact that the Principal Principle describes superbaby credences directly and
reasonable agent credences only indirectly makes it tempting to adopt an account
of probabilistic rationality that puts the superbaby in the driver’s seat. On this
account, we start by determining what credences a superbaby would assign to
various propositions. We then hold that an agent’s credences are reasonable just in
case they could be obtained from superbaby credences by conditionalizing on that
agent’s total evidence.

I think there are good reasons to be wary of this kind of account. The first
reason is that it is unclear why a reasonable agent’s credences are obtained from
superbaby credences by conditionalizing. Conditionalization has been argued for
(via Dutch Book arguments and the like) as a norm relating a reasonable agent’s
credences at different times to each other. Since no actual agent (reasonable or
otherwise) has ever been a superbaby, it is unclear why conditionalization should
represent a required relation between superbaby credences and reasonable agent
credences.

Second, it is unclear how many superbabies there are, with the contentious point
being whether there is a unique superbaby or whether there are multitudes. Put
another way, the question is whether it can be consistent with the requirements of
rationality for two agents with no empirical evidence to assign different uncondi-
tional credences (or conditional) to the same proposition (or ordered pair).

The defenders of a single-superbaby line must hold that the requirements of
strict rationality are incredibly strong, strong enough to take the superbaby from
a bare set of certainties about the a priori all the way to a unique mandated
credence distribution over the set of possible worlds.4 Indifference principles might
offer some hope here, but to this point an indifference principle that would do the
needed work without yielding contradictions or begging some important question
has the status of a subjectivist’s Philosopher’s Stone. On the other hand, those
who would admit multiple superbabies face the question of how it is compatible
with the requirements of rationality for a superbaby to assign a particular credence
to an empirical proposition for what looks like no reason at all.5

Third and finally, our understanding of superbabies may not be thick enough
to use judgments about them as the support for our entire system of reasonable
probabilistic agency. It would be one thing to see superbabies as just an idealized
extreme at one end of the spectrum of reasonable agents. On that approach, we
would develop requirements of rationality by considering the epistemic plights of
agents more normally situated, then generalize those requirements to all reasonable

---

4Even if you thought that all the superbaby assigned were conditional credences, this would
still yield a credence distribution over the set of possible worlds. For any world w the superbaby’s
credence C(w | T) (where T is the set of all possible worlds) will be the credence the superbaby
assigns to that world.

5Alan Hájek and John N. Williams expand upon this point in a manuscript entitled ‘p, and I
have absolutely no justification for believing that p’: the incoherence of Bayesianism. On a dif-
f erent front, one might ask how the multiple-superbaby-based theory could give rise to diachronic
norms for reasonable agents. A simple solution would be to require that once an agent assigns a
credence incompatible with having conditionalized from the credences of any of a particular set
of superbabies, he is barred thenceforth from adopting any credences that could be obtained by
conditionalizing only from superbabies in that set.
agents including superbabies. But to start with the superbabies, and derive rational requirements on other reasonable agents from our judgments about them, we would need a firm, direct grasp of what rationality requires of an agent with no empirical evidence. It is already difficult enough to make sense of superbabies as agents; the thicker understanding needed to execute this approach to Bayesian theory seems to me somewhat beyond our reach.⁶

For what it’s worth, while I do think Lewis means reasonable initial credence functions to be understood on a superbaby-like model, I do not think his view of reasonable probabilistic agency is the account we’ve just been considering that starts from superbabies and works its way up. One might read that account into various passages in the Lewis, such as the last sentence of the following:

[4] To the subjectivist who believes in objective chance, particular or general propositions about chances are nothing special. We believe them to varying degrees. As new evidence arrives, our credence in them should wax and wane in accordance with Bayesian confirmation theory. It is reasonable to believe such a proposition, like any other, to the degree given by a reasonable initial credence function conditionalized on one’s present total evidence. (285)

But I think the shaky confirmation given by this passage is overwhelmed by the evidence from passage #1 above and the following:

[5] Roughly speaking, what makes it be so that a certain credence function is your credence function is the very fact that you are disposed to act in more or less the ways that it rationalizes. (Better: what makes it be so that a certain reasonable initial credence function and a certain reasonable system of basic intrinsic values are both yours is that you are disposed to act in more or less the ways that are rationalized by the pair of them together, taking into account the modification of credence by conditionalizing on total evidence; and further, you would have been likewise disposed if your life history of experience, and consequent modification of credence, had been different; and further, no other such pair would fit your dispositions more closely.) No wonder your credence function tends to guide your life. If its doing so did not accord to some considerable extent with your dispositions to act, then it would not be your credence function. You would have some other credence function, or none. (288)

Passage #1 makes clear that Lewis is starting with the notion of a reasonable agent credence function, then developing the notion of a reasonable initial credence function (i.e. a superbaby function) from that. Passage #5 then makes clear that we are to start with the dispositions of a reasonable agent and develop out of those, via a representation-style approach, a characterization of her associated superbaby. (It is clear here and in many places elsewhere that Lewis is committed to the

⁶A defender of the superbaby-centric account might backpedal a bit and claim that superbaby talk is really talk about a credence function held by the reasonable agent at some early temporal point in his epistemic career. But if Lewis’s reformulation of the Principal Principle is meant to apply to all superbabies, that early superbaby cannot have any inadmissible evidence about the outcome of any chance events. And that seems a difficult requirement for the epistemic state of even a very young reasonable agent to meet.
existence of multiple reasonable initial credence functions.) The Principal Principle merely describes a feature of superbaby functions derived from reasonable agent dispositions.

This approach is consistent with Lewis’s stated aim in the paper: to elucidate our concept of chance. Lewis is not out to give an account of what makes an agent’s credence function reasonable; as is abundantly clear in passage #1, he is happy to take that notion for granted. He is after an account of chance, and given that project Lewis’s move to superbabies is quite sensible: his starting point is reasonable agent credences, but chances can’t be read directly off of those since they are often based on inadmissible background evidence, so he decides to read chances off of the credences of agents with no empirical evidence. We might have trouble making sense of such agents if we started with them as the beginning of a theory of something (such as rational subjective credence), but they are perfectly suitable as a theoretical element constructed from a given (reasonable agent dispositions) and used as a bridge to reach another goal (a theory of chance). Superbaby credences do a very nice job of allowing Lewis to derive the formal results about chance he is ultimately after.

But what about those of us who are after a theory of reasonable agent credences? I take it we all think reasonable credences are responsible not just for the sort of broad-brush correlation with current chance estimates suggested by a quick gloss of the Principal Principle, but also for the sort of fine-tuned conditional and time-indexed correlations suggested by the full Principal Principle. To get the responsiveness of reasonable credences to chances right, must we start from superbabies?

One approach might be to start with reasonable agents, but use superbabies as a handy construction along the way. Somewhat mirroring Lewis’s approach, we might start by considering cases in which we take an agent’s credence function to be reasonable, then systematize such cases by enumerating principles restricting the credences of superbabies and specifying that reasonable agents have credence functions matching conditionalizations on their total evidence from superbabies. The superbabies would play a purely instrumental role here, taken not even as agents but simply as theoretical elements in a formal machinery used to get from intuitions about reasonable agent credences in particular cases to suggestions about what would be reasonable in other cases.

Once we’ve taken all the life out of the superbabies in this way, I don’t have much objection to the approach. The theory is a theory about reasonable agent credences: it starts with intuitions about what would be reasonable for typical agents, it uses a particular machinery, then it yields verdicts about what would be reasonable for other typical agents. It does, however, move a bit indirectly. Instead of offering principles that we can apply directly to agents to assess whether they’re reasonable, it works through the medium of a superbaby construction. Might we instead find a version of the Principal Principle that applies directly to reasonable agent credences?

Earlier, we proposed a defective Principal Principle that attempted to impose a requirement directly on reasonable agent credence functions. Recall what went wrong there: it seemed reasonable to us for an agent to deviate her credence in \( A \) from the chance of \( A \) at \( t \) when her background evidence included information inadmissible for \( A \) at \( t \). But this suggests an immediate fix:
The Principal Principle (Direct). Let $t$ be any time. Let $x$ be any real number in the unit interval. Let $X$ be the proposition that the chance, at time $t$, of $A$’s holding equals $x$. Let $E$ be any proposition compatible with $X$ that is admissible for $A$ at time $t$. Let $C$ be the credence function of any reasonable agent whose total evidence is compatible with $X$ and $E$ and admissible for $A$ at time $t$. Then

$$C(A | X \& E) = x$$

Here we have avoided the inadmissible background problem by restricting the Principal Principle to the credence functions of agents with no inadmissible evidence. To do so, we have had to add the notion of an agent’s total evidence into the principle, but that notion would have been needed anyway on a superbaby account of reasonable credence.

It can easily be shown that this version of the Principal Principle generates all the verdicts about reasonable agent credences we were after to begin with. Moreover, I think it does a good job of expressing the intuition that makes the quick gloss appealing. While at a high level of detail it handles technical matters involving conditional credences and time-indexing, at a more general level this direct Principal Principle says just that a reasonable agent with no inadmissible evidence will set her credences equal to what she takes to be the chances.