Identity failure, functional forgetting and bogus stopping: a defense of conditionalization

by Randall G. McCutcheon

In this note I will comment on Frank Arntzenius's paper "Some problems for conditionalization and reflection" [1]. The purpose of that paper was to present various puzzles designed to show that rational agents can update credences in manners that "violate Bayesian conditionalization and Bas van Fraasen's reflection principle," allegedly owing to "the fact that there are two as yet unrecognized ways in which the degrees of belief of rational people can develop." Others (e.g. [2]), in commenting on the paper, have advocated a conservative position concerning the novelty or significance of these anomalies for rational action. Although there is nothing particularly new or different here, I'm offering my own less technical version because I'm not convinced that the objections have been particularly well understood by philosophers.

First puzzle: the road to Shangri La. You can only enter Shangri La if, after entering, you don't remember which of the two paths you took to get there (mountains or sea). So you conspire with the guardians as follows. First, you flip a coin. If heads, you go by the mountain path and enter Shangri La normally. If tails, you go by sea and, just as you cross the threshold, the guardians replace your memory of the journey by sea with a false memory of a journey by the mountain path. Suppose the coin lands *heads*. How does your credence in *heads* evolve? On the mountain path, it is 1. But later, inside Shangri La, it is one-half. If credences evolve by conditionalization alone, that can't happen. That's not necessarily very interesting. Obviously we forget things from time to time, and insofar as we do, our credences evolve in ways other than by conditionalization. According to Arntzenius, however, this case is more interesting than that, as in the *heads* scenario, "the development of your memories will be completely normal."

The development of one's memories in this example is perhaps normal in the sense of neural geometry, but in the functional sense it is not, and that is the sense on which aptness of the conditionalization paradigm depends. Any system of reliable correlations (strings tied around fingers, say) can play the memory role, as far as conditionalization is concerned. On the other hand, if "memoric episodes" lack such correlations, they cannot. In this puzzle, memoric episodes outside Shangri La as of journeys on the mountain path correlate perfectly with such journeys, whereas memoric episodes inside Shangri La as of journeys on the mountain path do not. Obviously the conditionalization paradigm fits at most insofar as said memoric episodes do correlate with the supposed targets of their intentionality. The functional role of memory is to establish this correlation, and to the extent that a candidate mechanism fails to implement this role, its workings don't constitute memory at all. The traveler on the mountain path has failed to lay down a neural pattern that will, inside Shangri La, correlate with having been on the mountain path. Perhaps it was just impossible or unprofitable for him to do so. Very well...in that case, it was just impossible or unprofitable for him to remember (in the functional sense). Which means that this puzzle falls into the uninteresting category of (functional) forgetting.

Second puzzle: a condemned prisoner is hoping for a stay of his execution. If he is to get it, a guard will turn a light off at midnight. Otherwise the light stays on all night. At 6 P.M. the prisoner's credence in getting the stay is one-half. The prisoner's cell is not outfitted with a clock, and his internal clock is imperfect. At 11:59 P.M. it seems likely that he will assign non-zero probability to it being past midnight. Since the light will still be on, his credence in the stay will accordingly be, almost surely, less than one-half. Moreover he knows this in advance. So, at 6:00, E(credence in *stay* at 11:59) is strictly less than one-half. This is supposed to be a violation of conditionalization and/or reflection.

It isn't that.[1] Here is the cause of the confusion: all conditionalization says is that the probability of B, in light of new knowledge A, should be updated to $P(B|A) = \frac{P(B \cap A)}{P(A)}$. Conditionalization implies that when one conditions event B on a measurable partition (that is, updates according to which cell of the partition obtains), the expected probability of B doesn't change. The contrapositive of this statement is that if the expected probability of an event B, conditioned on some measurable partition, differs from $P(B)$, then conditionalization fails. It does *not* follow that conditionalization fails if the expected probability of B one minute before midnight is different from $P(B)$. The paradigm of conditionalization is a paradigm of rational agency; in order to demonstrate that it fails, one must give an example in which, in light of new knowledge A, $P(B)$ should be updated to something other than $\frac{P(B \cap A)}{P(A)}$. One's expected credence in B at one minute before midnight is only relevant, therefore, when knowledge obtained prior to one minute before midnight is sufficient to determine when it is in fact one minute before midnight (in technical terms, when one minute before midnight is a *stopping time*, i.e. when our agent has it in his power to say *stop* at that time).

This point may be illustrated by scenarios simpler than that of the prisoner: consider a two-cell partition $\{A, A^c\}$. When I condition $P(B)$ on event A, I update my credence in

[1] What gives the paradox its kick is that the prisoner's credence in *stay* as it evolves over time should, according to principles of conditioning, be a bounded martingale (X_k). Then by standard convergence theorems (cf. e.g. [1, Paragraph 278]), if T is a stopping time (i.e. a random number with event $T = k$ depending only on the prisoner's evidence up to and including time k), one must have $E(X_T) = E(X_0) = \frac{1}{2}$. This appears to be violated by taking $T =$ 11:59 P.M. As the theorem is valid, the question arises as to which hypothesis fails: credence in *stay* being a martingale or $T =$ 11:59 P.M. being a stopping time. It is easy to see that the martingale hypothesis fails when k counts out real time but the prisoner loses track of it, relegating the paradox to the uninteresting case of forgetting, while the stopping time hypothesis fails when k counts out subjective time, and 11:59 comes as a surprise to the prisoner. If we are not to hold our rational agents to impossible standards, subjective time makes for the superior model, where expected credence in *stay* is indeed $\frac{1}{2}$ conditional on any future subjective time, and the prisoner's observed decreasing actual credence in *stay* is a consequence of conditionalization on the additional evidence that the light remains on.

B to $P(B|A) = \frac{P(B \cap A)}{P(A)}$. Similarly when I condition B on A^c, I update my credence in B to $P(B|A^c) = \frac{P(B \cap A^c)}{P(A^c)}$. So, if I ask whether A or A^c obtains and condition $P(B)$ on what I learn, my expected credence in B after conditioning is

$$P(A) \cdot P(B|A) + P(A^c)P(B|A^c) = P(A)\frac{P(B \cap A)}{P(A)} + P(A^c)\frac{P(B \cap A^c)}{P(A^c)} = P(B), \qquad (*)$$

as expected. Consider though a non-uniform protocol whereby if A obtains I will be informed of this at 11:59, whereas if A^c obtains I will be informed of this at 12:01. At 6:00, what is E(credence in B at 12:00)? If A obtains, I will have learned this by 12:00, so my credence in B at that time will be $P(B|A)$. If A^c obtains, my credence in B will be fairly close to $P(B)$, as I will assign high probability, at that time, to the actual time being before 11:59 (since I haven't heard anything yet); I won't update to $P(B|A^c)$ until 12:01, and in the meantime (*) is apt to fail. Is this a violation of conditionalization? Certainly not...in fact, without assuming that conditionalization is correct, we'd have no argument that (*) fails! The hypothesis of the theorem about the expected probability of B staying the same upon conditionalization with respect to a measurable partition simply hasn't been met. At least, not with respect to the partition $\{A, A^c\}$. At midnight we'll be in the process of conditioning on that partition, but may well not have finished. The puzzle about the prisoner is analogous. If the light goes out at midnight, he finds this out at midnight. If the light does not go out at midnight, this dawns on him by degrees.

It deserves some discussion how one might formalize a system in which a constant time turns out not to be a valid stopping time. Usually, one can spot time-valued random variables that aren't stopping times a mile away by their indeterminate definitions. Suppose for example I were to search a beach for a lost ring. The search ends when I find the ring or complete a single pass over the beach. Say my initial credence is $\frac{1}{2}$ that I will recover the ring. There seems to be very little mystery in the fact that, at start, E(credence in *ring recovered* one second before end of search) is substantially less than one-half. The given time is obviously not a valid stopping time; there is no way for me to know when it's exactly one second before the end of my search. By contrast, given a fixed constant $0 < \alpha < 1$, E(credence in *ring recovered* when proportion α is combed) is $\frac{1}{2}$ (it is certainly in my power to yell *stop* when proportion α has been combed).

How then can the "determinate time" 11:59 be, for the prisoner, more like the indeterminate time *one minute before end of search* than it is like the determinate time *when proportion α is combed*? Well, the problem is what one means by *determinate time.* In this case, what one should mean is *determined by information obtained up to that time*, and by this criterion, of course, 11:59 fails to be a determinate time. A determinate time for the prisoner, by contrast, would be something like "when my internal clock is centered on 11:59". Keep in mind: the beach comber is continually updating his credence in *ring recovered* according to the proportion of the beach that has been combed together with the fact that the ring has not been found, while the prisoner is continually updating his credence in *stay* according to the center[2] of his internal clock together with the fact that the light is still on. It would be news if there were some state Σ of the prisoner's internal clock, conditioned on which his expected credence in the stay were different from one-half. But of course there is no such state Σ; the prisoner's expected credence in the stay when his internal clock is centered on 10:30, or 12:45 or 11:59, is one-half.[3]

The upshot: though it's an instructive example, the case of the prisoner is ultimately no more of a challenge to conditionalization or reflection than are more commonly cited consequences of stopping time protocol violations, such as the fact that my expected balance at a fair gambling table just prior to my first losing bet is strictly positive.[4]

Fourth puzzle: Sleeping Beauty. A fair coin is flipped on Sunday night. Beauty's credence in *heads* on Sunday night is $\frac{1}{2}$. She wakes up Monday, is asked her credence in *heads*, and if in fact *heads* obtains that's the end of the experiment. If *tails* however, she is put back to sleep and a drug is administered that erases memory of her Monday awakening. When she wakes up on Tuesday, she has a subjectively identical experience in which she is again asked credence in *heads.* Then that's the end of the experiment. In any event, Beauty's memory of the experiment is erased at the end. *Halfers* say that since Beauty's credence in *heads* is $\frac{1}{2}$ on Sunday and she doesn't learn anything new relevant to *heads* by Monday morning, her credence in *heads* should still be $\frac{1}{2}$ on Monday. *Thirders* (most of them) say that she does learn something new that is relevant to *heads* (say, for example, that it's not the *heads Tuesday* scenario) when she wakes up Monday, and that, upon conditioning on this new information, Beauty should assign *heads* a credence of $\frac{1}{3}$.

Arntzenius is a thirder, but not a typical one. He says both that Beauty learns nothing relevant to *heads* and that Beauty should assign *heads* a credence of $\frac{1}{3}$ on Monday morning. And, that this violates both reflection and conditionalization. For an argument, he imagines a scenario in which Beauty is a lucid dreamer, who, when not awakened at 9:00 A.M., will always dream that she is awakened. She then pinches herself, which hurts if she is actually awake and which doesn't hurt (and doesn't awaken her) if she is asleep. In this scenario, Arntzenius argues, Beauty should upon pinching herself and finding herself to be awake condition on elimination of *Tuesday heads* and update credence in *heads* to $\frac{1}{3}$. Later in life, Beauty loses the habit of dreaming and at that point knows herself to be

[2] Also the variance, assuming a normal distribution, of course.

[3] I am assuming, of course, that the operation of the prisoner's internal clock is independent of the light's being on or off.

[4] Arntzenius mentions gambling in regard to this puzzle thusly: "At 6 P.M. you will be willing to accept a bet on (*stay*) at even odds, and at 11:59 P.M. you will, almost certainly, be willing to accept a bet on (*no stay*) at worse than even odds. And that adds up to a sure loss. And that means you are irrational." The status of this passage is unclear to me (it may be interlocution), but of course in order for a bookie to take advantage of the prisoner by this series of bets they would have to know when it was 11:59 P.M., and it isn't interesting that a bookie who knows more than the prisoner can gain an advantage.

awake in the morning without the need for a pinch. As her epistemic position now upon awakening is the same as it was previously after the pinch, her credence in *heads* should still be $\frac{1}{3}$, but this is not arrived at via conditionalization by elimination of alternatives.

As an argument for the one-third solution, this fails, and as an attack on conditionalization, fails to avoid falling into the uninteresting case of (functional) forgetting. Beauty has two valid readings of *credence* at her disposal, leading to distinct credence functions: *propositional credence* and *de se credence*. Adoption of propositional credence as one's subjective credence function in Sleeping Beauty scenarios was developed by David Lewis, and though it is somewhat counterintuituve, it does cohere with accepted maxims of probability.[5] According to his account, there is an important difference between non-existent and eliminated alternatives. The "commutativity" argument (if result of evidence is the same, it doesn't matter what order it was gathered in) of Arntzenius fails as an argument against Lewisian halfing because, for Lewis, Dreaming Beauty and Sleeping Beauty don't wind up in analogous situations. Dreaming Beauty knows, after pinching herself, that if *heads* then she will, tomorrow, again have an experience as of an experimental awakening (which for Lewis reduces the weight of her current conviction that the coin lies *heads*), whereas Sleeping Beauty knows that, if *heads*, she will not experience any such thing.

As an attack on the conditionalization paradigm, Arntzenius's one-third solution also founders, owing to the fact that Beauty functionally forgets that this is her first awakening since Sunday night. That is, she has failed to record a brain state that correlates reliably with it being such an awakening. If she was prevented from doing so, that's okay– all it means is that she was prevented from remembering. So, just as in the Shangrai La case, where credence in P dropped from one to one-half upon unelimination of a *not* P alternative to stand alongside a single P alternative, here credence in Q drops from one-half to one-third upon unelimination of a *not* Q alternative to stand alongside a pair consisting of a *not* Q alternative and a Q alternative. The only difference here is that the alternatives are centered worlds, but that doesn't matter; thirders utilize *de se* credences, and *de se* credences treat centered worlds as propositional credences do uncentered worlds.

Fifth puzzle: Shiva and Brahma. First version. Vishnu tells you that one month ago, two identical humans, of which you are one, were created by Brahma. At the same time, Shiva tossed a coin. If it landed *heads*, Shiva will destroy one of the two humans one month from now. You are advised to check your mail. If there is a copy of Fred Dretske's *Knowledge and the Flow of Information* waiting for you, you will be destroyed. You check your mail and find no such book. Your credence in *heads* is now of course $\frac{1}{3}$. Second version. If the coin landed heads, Shiva destroyed one of the humans one week ago. It wasn't you obviously, so your credence in *heads* is again $\frac{1}{3}$. Third version. One of the humans had existed previously, the other was a duplicate. No difference. Still $\frac{1}{3}$. Fourth

[5] Although he does not say as much, Lewis plainly views the fact that Beauty experiences two awakenings in case of *tails* as a sampling bias (*tails* worlds are oversampled); his solution just fixes the bias in the most natural way.

version. Rather than destroy one of the humans, Shiva, in case of heads, simply prevents Brahma from creating the duplicate. According to Arntzenius, this makes no difference, and credence in *heads* is still $\frac{1}{3}$. Fifth version. Same as fourth version, but the duplicate (in case of *heads*) will not be created until one month from now. Therefore in one month, you'll have credence $\frac{1}{3}$ in heads, yet right now you have credence $\frac{1}{2}$ in *heads*. This is supposed to be a violation of reflection (and conditionalization, presumably).

First, the similarity of the third and fourth versions is contentious; a Lewisian could mark a difference between destruction and prevention and assign credence $\frac{1}{2}$ to *heads* in the fourth version. Possibly Lewis would condone such a solution, though there are differences between this case and that of Sleeping Beauty, where the Lewisian solution compensates for Beauty's double tails jeopardy by only counting her estimated credence in *heads* against her once, via assignation of half weight to each of her Monday and Tuesday tails awakenings. That's a tough sell in version four, as you and your duplicate are probably distinct agents now. There are however some perspectives that make it plausible (souls are split in two and thus diminished by duplication, souls aren't duplicated or split, but who winds up with them is uniformly random, etc.), so as an argument for a one-third solution, what Arntzenius has to say may still fail.

Perhaps that's not so bad. It's the one-third solution that seems to suffer a violation of reflection; nor is this seeming violation a great fit for functional forgetting. True, at the end of the month you can't remember whether you were a person yesterday or a steaming vat of organic molecules in Brahma's laboratory. However, this is an accidental feature of the puzzle, which could as easily have had it that in one month's time Shiva destroys you and creates *two* duplicates. If these duplicates have forgotten anything now, it would be difficult to say what. Is this then a novel failure of conditionalization? Not necessarily. It's true that before the month is up, you have credence $\frac{1}{2}$ in *heads*, and after the month is up, your duplicates have credence $\frac{1}{3}$ in *heads*. But, what if you ceased to exist and your duplicates aren't you? Call this *identity failure*. Identity failure dodges the conclusion that anyone's personal credence function was updated in violation of conditionalization by cogently maintaining that no one's personal identity function was updated at all.

References

[1] F. Arntzenius, Some problems for conditionalization and reflection, *The Journal of Philosophy* **100** (2003), 356-370.

[2] M.J. Schervish, T. Seidenfeld and J.B. Kadane, Stopping to reflect, *The Journal of Philosophy* **101** (2004), 315-322.

rmcctchn@memphis.edu