Hedden – Time-slice rationality

Brian T. Miller

November 21, 2017

§1 – Introduction

Contrast two cases:

> **Fickle Frank:** Frank is a physicist who changes his mind constantly and frivolously. At breakfast, he is pretty sure that the Everett multiple universe hypothesis is the right interpretation of quantum mechanics. By mid-morning, he abandons that belief in favour of the Copenhagen interpretation. At lunchtime, he switches camps once again, siding with the de Broglie-Bohm theory. But that does not last, and by afternoon tea he is firmly convinced that some sort of hidden variable approach must be right. It is not that he keeps gaining new evidence throughout the day which supports different hypotheses. Rather, he just changes his mind.
>
> **The Frankfurt Physics Conference:** A major conference on quantum mechanics is being held in Frankfurt. In attendance are proponents of a wide range of interpretations of quantum mechanics. There is a team of researchers from MIT who believe that the Everett multiple universe hypothesis is the best explanation of the available data. Seated next to them is an eminent professor from Cambridge who advocates the Copenhagen interpretation. Further down the row is a philosopher of physics who recently authored a book arguing that the de Broglie-Bohm theory is correct. In all, the lecture hall is filled by advocates of at least a dozen competing quantum-mechanical views.

Common intuition: Frank is irrational, but the physicists at the conference are perfectly rational (for all we know)

BH rejects this intuition: Frank's time-slices stand in the same relations to one another as the time-slices of the various physicists at the conference.

Time-slice rationality:

Synchronicity: What attitudes you ought to have at a time does not directly depend on what attitudes you have at other times

Impartiality: In determining what attitudes you ought to have, your beliefs about what attitudes you have at other times play the same role as your beliefs about what attitudes other people have

§2 – *Against Conditionalization*

§2.1 – *Conditionalization and personal identity*

Conditionalization: It is a requirement of rationality that, for all H,
$P_1(H) = P_0(H \mid E)$

Conditionalization does not satisfy Impartiality because it treats *your* past beliefs differently from how it treats other people's beliefs.

That means that, in order to determine what you ought to believe, we first have to determine the correct theory of identity over time.

Problem:

> [Pre] enters the teletransporter in New York. Her body is scanned, and at the moment her body is vaporized, two different molecule-for-molecule duplicates of her are created, one in Los Angeles and the other in San Francisco. Call them 'Lefty' and 'Righty', respectively. Lefty and Righty are qualitatively just like Pre in all physical and mental respects. Now, there is a debate about whether Lefty, or Righty, or both, or neither is the same person as Pre. But what I want to emphasize is that in order to determine what Lefty and Righty ought to believe, following the double teletransportation, we do not have to first settle this debate about personal identity over time. If Lefty appears and immediately gains some new evidence,[1] we do not have to first figure out the correct theory of personal identity in order to determine what Lefty ought to believe. All that matters is what Lefty's present evidence is. But Conditionalization conflicts with this datum. It only says that Lefty's credences ought to be constrained by Pre's credences if Lefty is the same person as Pre; it is silent if Lefty and Pre are not the same person. If Lefty and Pre are not the same person, it is as if Lefty just suddenly came into existence, so it is compatible with obeying Conditionalization that Lefty choose any rationally permissible prior probability function and update it on her present total evidence, unconstrained by facts about Pre's credences (except in so far as they affect Lefty's present evidence). (456)

[1] Is that possible? On the Bayesian framework agents are always assumed to have a credence function, and that's in part what determines their posterior credences when they obtain new evidence. The alternative is to suggest that there might be a superbaby whose evidence precedes her prior credences, and I'm not sure how to make sense of that situation. The central question seems to be whether Lefty's credence function is the same as Pre's credence function, and once that question is settled the identity relation between Pre and Lefty is beside the point.

BH: we shouldn't have to settle personal identity questions in order to settle rationality questions

BTM:

Not everyone understands Conditionalization in the way that BH describes, and alternative understands avoid this problem. In particular, Richard Jeffrey thinks that Conditionalization is only rationally appropriate when your conditional probabilities don't change throughout your acquisition of new evidence. In that case Conditionalization is a trivial consequence of the norm of probabilistic coherence:

Total Probability is a theorem of the probability calculus and the definition of conditional probability:

Total Probability $P_{new}(A) = \sum_i P_{new}(A \mid B_i)P_{new}(B_i)$

In simple cases where one's evidence consists of a shift in credence in a single proposition e, Total Probability requires that:

$$P_{new}(A) = P_{new}(h \mid e)P_{new}(e) + P_{new}(h \mid \neg e)P_{new}(\neg e)$$

And when one becomes *certain* that e is true, and hence *certain* that $\neg e$ is false, that becomes:

$$P_{new}(h) = P_{new}(h \mid e)P_{new}(e)$$

Now, if obtaining new evidence e doesn't change any of my credences *conditional on* e, then:

$$P_{new}(h \mid e) = P_{old}(h \mid e)$$

Put the last two together and you get:

Conditionalization: $P_{new}(h) = P_{old}(h \mid e)P_{new}(e)$

Note that none of this required any assumptions about identity over time *of the agent*, only identity over time of one's probabilities conditional on the evidence.

The same point holds for Jeffrey Conditionalization:

Total Probability $P_{new}(A) = \sum_i P_{new}(A \mid B_i)P_{new}(B_i)$

together with the stipulation that there's no change to your probabilities conditional on *any* evidence proposition B_i, i.e. that:

$$B_i,\ P_{new}(A \mid B_i) = P_{old}(A \mid B_i)$$

yields Jeffrey Conditionalization:

Jeffrey Conditionalization $P_{new}(A) = \sum_i P_{old}(A \mid B_i)P_{new}(B_i)$

NB the sharp contrast between:

1. BH's understanding of Conditionalization as a norm of rationality that *always* applies to agents, and

2. RJ's understanding of Conditionalization as a norm of rationality that applies only when a certain precondition has been met: that you don't change your probabilities conditional on the evidence propositions

BH's Fission case is unproblematic for RJ's way of seeing things.

Question: is RJ's way of seeing things consistent with Impartiality?

End BTM

§2.2 – Conditionalization and Internalism

For BH the relevant version of Internalism is *mentalism*:

mentalism: what attitudes you ought to have supervenes on your **present** mental states[2]

[2] BH doesn't always specify the 'present' part, but it's quite important for his arguments for Synchronicity.

Case:

> **Two Roads to Shangri La:** There are two paths to Shangri La, the Path by the Mountains, and the Path by the Sea. A fair coin will be tossed by the guardians to determine which path you will take: if heads you go by the Mountains, if tails you go by the Sea. If you go by the Mountains, nothing strange will happen: while traveling you will see the glorious Mountains, and even after you enter Shangri La, you will forever retain your memories of that Magnificent Journey. If you go by the Sea, you will revel in the Beauty of the Misty Ocean. But, just as you enter Shangri La, your memory of this Beauteous Journey will be erased and be replaced by [an apparent] memory of the Journey by the Mountains. (459)

Suppose that you know the setup of the case and that you in fact travel by the Mountains.

BH's claims about what we ought to think:

- while en route you ought to be highly confident that you are going by the Mountains,
- upon entering Shangri La, your credence that you went by the Mountains should drop to 0.5, since
  - (i) you have no evidence that suggests that your apparent memory is real rather than illusory, and
  - (ii) whether your apparent memory would be real or illusory was determined by the toss of a fair coin

What Conditionalization says:

- you ought to start off with credence 0.5 that you will travel by the Mountains
- upon entering the mountains and conditionalizing on your new evidence you become highly confident that you are travelling by the Mountains
- upon entering Shangri La, you do not gain any new evidence that bears on whether you travelled by the Mountains, and hence Conditionalization does not kick in.
- So, according to Conditionalization, you ought to just retain your high credence that you travelled by the Mountains.

So, Conditionalization should be rejected.

§3 – Against reflection

According to reflection, one is rationally obliged to defer to the belief of your future self:

Reflection: It is a requirement of rationality that, for all H, $P_0(H \mid P_1(H)=n) = n$

Although Reflection is a sychronic rule,[3], it's vulnerable to counterexamples involving future irrationality[4] or loss of evidence[5], so:

Modified Reflection: It is a requirement of rationality that, for all H, $P_0(H \mid P_1(H)=n) = n$, unless you believe that at t_1 you will be irrational or will have lost evidence

Problem for Modified Reflection: it's insufficiently general.

1. Why should you always defer to your *future* self? Aren't there cases when my epistemic position will get worse over time?
2. Why should you always defer to your future *self*? Shouldn't I in some cases defer to someone else's beliefs, e.g. when that someone else is an expert and I'm not?

Conclusion: Reflection is insufficiently general, so it must be a special case of a more general, more fundamental principle.

[3] Rational belief is determined by what you now believe that your future self will believe, and hence nothing outside of your present mental states is being appealed to.

[4] E.g. you believe that your future self will be drunk

[5] E.g. you believe your future self will have forgotten something that you now know

§4 – Rebutting diachronic Dutch Book arguments

Dutch Book: a set of bets that together guarantee a loss

Synchronic Dutch Book arguments purport to show that probabilistic incoherence is irrational

Question: How much are you willing to pay for a bet? I.e., how much will you pay for a chance to win $10, with a 30% chance of winning?

Decision Theory's answer[6]: you should pay up to the *expected value* of the bet:

[6] Basic DT + a linear valuation of money, that is.

For a bet with two possible outcomes, A and B,

Expected Value = P(A) x Value(A) + P(B) x Value(B)

In this case the two possible outcomes are *win* and *lose*

Expected Value of taking the bet:

P(win) x V(win) + P(lose) x V(lose) = .3 x $10 + .7 x 0 = $3

Now suppose your beliefs are incoherent, and you think there's a 150% chance that you'll win that same bet: then the expected value of the bet = P(win) x V(win) + P(lose) x V(lose) = **1.5** x $10 + .7 x 0 = $15

So, you're willing to pay $15 for a bet that pays out a maximum of $10.

That's irrational. It's also probabilistically incoherent to think there's a 150% chance of anything. So: don't be incoherent.

Diachronic Dutch Book arguments purport to show that failing to conditionalize (or conform to Reflection) is irrational

NB: these arguments are quite different: one might be probabilistically coherent at each of a series of moments without transitioning moment to moment via conditionalization. In that case you're immune to *synchronic* Dutch Book arguments but (maybe) vulnerable to *diachronic* Dutch Book arguments.

Example:

$P_0(e) = .5$

$P_0(h|e) = .75$

But, if you gain evidence for e your credence in h will be .65.

NB this means that you didn't conditionalize properly!

At t_1, prior to learning that e, a bookie offers you 1 cent if you take both Bet 1 and Bet 2:

Bet 1: pays \$25 if $h\&e$, \$ − 75 if $\neg h\&e$, and \$0 if $\neg e$

Expected Value:

= [$P(h\&e)$ x $V(h\&e)$] + [$P(\neg h\&e)$ x $V(\neg h\&e)$] + [$P(\neg e)$ x $V(\neg e)$]

= [.75 x \$25] + [.25 x \$-75] + [.5 x \$0]

= \$0

Bet 2: pays \$5 if e and \$-5 if $\neg e$

Expected Value:

= [$P(e)$ x $V(e)$] + [$P(\neg e)$ x $V(\neg e)$]

= [.5 x \$5] + [.5 x \$-5]

= 0

Since the expected value of each bet is \$0, you should take them along with the bookie's \$.01 and (probably) come out slightly richer.

At t_2 you'll learn whether e. If e is true, then the bookie will offer to pay you 1 cent to take Bet 3:

Bet 3: pays \$-35 if h and \$65 if $\neg h$

Expected Value:

= [$P(h)$ x $V(h)$] + [$P(\neg h)$ x $V(\neg h)$]

= [.65 x \$-35] + [.35 x \$65]

= 0

As before, the expected value of your bet is \$0, so you should take the bet and (probably) earn a penny.

But now put them all together:

- if $\neg e$ is true, then:
  - Bet 1 pays \$0
  - Bet 2 pays \$-5
  - you are not offered Bet 3
    - Total: \$-5
- if e is true, then you've taken all three bets. Then your winnings depend on the truth of h:

	h	$\neg h$
Bet 1	\$25	\$-75
Bet 2	\$5	\$5
Bet 3	\$-35	\$65
Overall Winnings	\$-5	\$-5

So: you've failed to conditionalize properly, and as a result you've taken a series of bets in which you're guaranteed to lose money - you've take a Dutch Book. That's irrational! So conditionalize properly.

BH: Diachronic Dutch Book arguments are question-begging:

- they assume that it's irrational for different time-slices of the same agent to act to produce a mutually disadvantageous outcome (in this case, losing money).
- time slicer: different time-slices of a single agent should be treated no differently from time slices of different agents
- and it's uncontroversial that time slices of different agents might act perfectly rationally while producing mutually disadvantageous outcomes. Example: prisoner's dilemma

Prisoner's dilemma: two criminals A and B are at the police station, unable to communicate with each other. There isn't enough evidence to convict on all charges unless one criminal agrees to testify. So the prosecutor offers a deal: if you confess and your partner doesn't, you get one year and your partner gets twenty years. If both confess, each gets ten years. If neither confesses, then each will be convicted on lesser charges and serve two years.[7]

	A: confess	A: don't confess
B: confess	-10,-10	-1,-20
B: don't confess	-20,-1	-2,-2

The idea is that it's rational for both A and B to confess, even though the outcome would be much better (from their perspective) if they both refuse to confess. So it's a case in which distinct time-slices of agents act rationally while producing a mutually disadvantageous outcome.

BH:

> In the diachronic Dutch Book case above where E is true, your t_1 self prefers accepting Bets 1 and 2 (plus the penny), no matter what your t_2 self does. And your t_2 self prefers accepting Bet 3 (plus the penny), no matter what your t_1 self did. Moreover, your t_1 and t_2 selves each

[7] Here we've moved from *decision theory* to *game theory*, the difference being that in game theory the outcome of one's actions depend in part on the actions of others. Below is a *decision matrix*. The top row represents the actions available to A, and the leftmost column represents the actions available to B. The remaining boxes represent the value of each outcome for A and B given the relevant actions taken by each, so the '-1,-20' box indicates that if B confesses and A doesn't, then the value to B is -1 (year in prison) and the value to B is -20.

> prefer that the other reject the bets she is offered. But the outcome that results from your t_1 and t_2 selves each accepting the bets they are offered is worse by each of their lights than the outcome that would have resulted from their declining those bets. So, the diachronic Dutch Book case is an intrapersonal Prisoner's Dilemma, with your t_1 self as Prisoner A, your t_2 self as Prisoner B, accepting the bets offered as defecting, and rejecting the bets offered as cooperating.
>
> ...The Prisoner's Dilemma is just a case where two people predictably wind up with a mutually dispreferred outcome without anyone being irrational. We should say the same thing about Lewis's and van Fraassen's in*tra*personal Prisoner's Dilemmas. (466)

§5 – Replacing Conditionalization

Intuition: Fickle Frank is irrational, and so is anyone else whose beliefs fluctuate wildly.

So, seems like we need some sort of principle by which Frank's t_2 time-slice is doxastically constrained by his t_1 time slice.

The is hard for those who accept *Permissivism*:

Permissivism: Given a body of total evidence, there are multiple unique doxastic states that it is rational to be in

Permissivism contrasts with *Uniqueness:*

Uniqueness: Given a body of total evidence, there is a unique doxastic state that it is rational to be in

Given Uniqueness, as long as Frank's total evidence doesn't change radically between t_1 and t_2, what it's rational for him to believe can't change radically either; in that case there's no need for some further principle of rationality to constrain his doxastic revisions

For Permissivists, Conditionalization can serve that role:

- Given a total body of evidence, there are multiple unique doxastic states that it is rational to be in because there are multiple unique prior credence functions that are rationally permissible
- But, once Frank's credences are set at t_1, Frank's new evidence together with Conditionalization determine what credences he must have at t_2
- By some measures, Conditionalization is the belief-revision method that accommodates new evidence with the least disruption to pre-existing doxastic state

- so, no wild fluctuations in doxastic state

Second option: embrace Uniqueness and Conditionalization

Bayesians who think that there's a unique rational prior[8] credence function are called *Objective Bayesians*. Objective Bayesians can still embrace Conditionalization, and rely upon it constrain changes in Fickle Frank's doxastic states.

But BH has argued that Conditionalization should be rejected, so how is he going to constrain Franks beliefs?

1. Embrace Uniqueness
2. Embrace *Synchronic Conditionalization*[9]

   Synchronic Conditionalization: Let P be the uniquely rational prior probability function. If at time t you have total evidence E, your credence at t in each proposition H should equal P(H | E).

Recall (diachronic) Conditionalization:

Conditionalization: It is a requirement of rationality that, for all H, $P_1(H) = P_0(H \mid E)$

The principle is diachronic because it defines $P_1(\cdot)$ in terms of $P_0(\cdot)$[10]

Synchronic Conditionalization doesn't do that: it only mentions a single moment, t.

So how does this rule help constrain Fickle Frank?

It doesn't: the rule isn't what's doing the work. Frank's rational doxastic states are stable because they are determined by his total evidence, and his total evidence is stable: it grows or shrinks gradually (usually).

§6 – Replacing Reflection

Modified Reflection needs to be replaced with a new deference-to-experts principle.

Where P_{you} is your credence function and $P^A_{ex}(H) = n$ is the proposition that A is an expert (someone who is rational and has more evidence than you do) with credence n in H, we get:

Expert Deference: It is a requirement of rationality that, for all H, $P_{you}(H | P^A_{ex}(H) = n) = n$

[8] Unfortunately, 'prior' is ambiguous in this context. First disambiguation: a prior credence function is just the one you had before you conditionalized on new evidence. Used in this sense 'prior' is a relative term: relative to t_2, my credence function from t_1 was my prior credence funtion, relative to t_3... Second disambiguation: a prior credence function is the one I had in the absence of all evidence, i.e. before conditionalizing even once. (Agents who have no evidence are called 'super-babies'.) Used in this sense 'prior' is not a relative term. Objective Bayesianism is characterized by the thesis that there's a uniquely rational prior (in the second, non-relative sense) credence function.

[9] This is essentially the route that Williamson suggests in Ch. 10 of KAIL.

[10] Recall that the subscripts refer to the times at which the credence functions are held.

Expert Deference avoids the problems with Modified Reflection:

- it allows you to regard your future self as an expert, but doesn't force you to do so in problem cases (e.g. forgetting, inebriation)
- doesn't *require* the expert to be you, or to be you in the future
- no need to solve the problem of personal identity over time to determine whether one ought to defer, because it doesn't matter whether the expert is you or not

Furthermore, the best argument for Modified Reflection[11], if sound, also established Expert Deference

Expert Deference is a kind of generalization of Modified Reflection: it allows you to treat your future self as an expert (as Modified Reflection recommends), and also allows you to treat your past self, or another person, as an expert.

Hence Expert Deference is the more fundamental principle.

[11] That argument is complex and we haven't done the background reading to evaluate it properly, so no details here.