A New Problem with Mixed Decisions, or: You'll Regret Reading This Article, But You Still Should

Benjamin Plommer

Penultimate draft, 7 June 2015

Abstract

Andy Egan recently drew attention to a class of decision situations that provide a certain kind of informational feedback, which he claims constitute a counterexample to causal decision theory (CDT). Arntzenius and Wallace have sought to vindicate a form of CDT by describing a dynamic process of deliberation that culminates in a "mixed" decision. I show that, for many of the cases in question, this proposal depends on an incorrect way of calculating expected utilities, and argue that it is therefore unsuccessful. I then tentatively defend an alternative proposal by Joyce, which produces a similar process of dynamic deliberation but for a different reason.

Andy Egan [2007] recently drew attention to a class of decision situations which, he claims, constitute a counterexample to causal decision theory (CDT). In these situations, the process of making a decision provides the agent with informational feedback that makes all options *unratifiable*: no matter what option the agent chooses, in choosing it she will receive information that will make her prefer some other option. Frank Arntzenius [2008] and David Wallace [2010] have both sought to vindicate forms of CDT through the use of mixed decisions, in which the agent chooses a probability distribution over acts rather than a single act. In this way, they claim, an agent in the situations Egan describes can make a mixed decision that she will have no reason to want to change. In this paper I show that, for a significant number of the cases in question, this proposal does not work because it rests on an incorrect way of calculating the expected utility of a mixed decision.

In section 1 I introduce expected utility theory and describe a case involving a conditional bet that is unfair, despite being composed of two fair unconditional bets. This unfair conditional bet illustrates a general phenomenon which, as I show later, leads Arntzenius (systematically) and Wallace (in one case) to miscalculate the expected utility of mixed decisions. In section 2 I introduce causal decision theory, the version of expected utility theory defended by Arntzenius and Wallace. In section 3 I describe the problem cases and the instability to which they give rise. In section 4 I introduce mixed decisions and show how, in some cases, they solve the problem of instability. In section 5 I argue that, for a large subclass of the problem cases, mixed decisions do not constitute a solution. The solutions that have been proposed rest on a way of calculating expected utilities that is incorrect, for the same reason that the conditional bet described in section 1 is unfair. Arntzenius is concerned mainly with the cases in which my objection applies, whereas Wallace is concerned mainly with those in which it does not; consequently, in this section I engage mainly with Arntzenius's proposal, though Wallace's is similar. In section 6 I describe, and tentatively accept, a different but related proposal by James Joyce [2012]for dealing with these cases.

1 Evidential expected utility and conditional bets

In this section we shall consider a series of easy problems.

> Problem 1. You are faced with two closed boxes, A and B. The boxes are transparent; you can see that box A contains \$10 and box B contains \$20. You may open one box and keep the contents. What do you do?

This problem is not difficult: assuming that you prefer having more money to having less, you should take box B.

> Problem 2. You are faced with two closed boxes, A and B. Box A is transparent, and contains \$10. Box B is opaque; you confidently believe that its contents were determined by the toss of a fair coin, such that it contains \$21 if the coin landed heads but is empty if the coin landed tails. You may open one box and keep the contents.

The expected utility calculation for this case is straightforward, but it is worth introducing some terminology at this stage for later use. We define the *choice partition* $\mathcal{A}$ as comprising the actions available to the agent and the and *outcome partition* $\mathcal{O}$ as specifying the different ways that events might unfold, in a way that is sufficiently fine-grained to capture all aspects of the situation that the agent cares about. In this case the the agent has two options, *take box A* (A_A) and *take box B* (A_B). Given we care only about money, there are three possible outcomes: *win \$0*, *win \$10* and *win \$21.* We also need to provide a *utility function*, $U(\cdot)$, giving a (cardinal) numerical measure of the desirability of each outcome. The relation between dollars and utilities depends on the agent's attitude to risk, but for the remainder of this paper we will assume agents are risk-neutral, valuing money at a rate of 1 dollar = 1 utility point. The agent's utility function for cash prizes is then $U(\text{win } \$x) = x$. In fact, to simplify matters, we will build the agent's utility function into our outcome partition by representing outcomes as *value-level propositions* of the form $V = v$, where v is the numerical desirability of the outcome in question (C.f. Briggs, 2010). $\mathcal{O}$ is thus represented as the set $\{V = 0, V = 10, V = 21\}$. The expected utility of an act A, then, is simply the expectation value of V given that the act is performed: that is, an average of the desirabilites of the possible outcomes, with each outcome weighted according to its subjective probability given that A occurs: $EEU(A) = \sum_{o \in \mathcal{O}} Cr(o/A)\,U(o)$. In Problem 2, taking box A has an expected utility of 10, whereas taking box B has an expected utility of 10.5, and so you should take box B.

> Problem 3 (The Boxing Match). You are considering placing a bet of \$10 on a boxing match between Bill, the reigning champion, and Ted, the underdog. If you bet on Bill and he wins, you make a profit of \$2; if you bet on Ted and he wins, you make a profit of \$30; if you bet on the loser, you lose

> \$10. You expect Bill to win with a probability of 0.7. Thus, a bet on Bill has an expected utility of $0.7 \times 2 + 0.3 \times (-10) = -1.6$, whereas a bet on Ted has an expected utility of $0.3 \times 30 + 0.7 \times (-10) = 2$. However, you do not have the option of choosing which boxer to bet on. Instead, you are offered a conditional bet. If you accept then a fair coin will be tossed; if it lands heads you must bet on Bill, and if it lands tails you must bet on Ted. Should you bet?

We can calculate the expected utility the conditional bet by enumerating the possible outcomes and taking an average of their values, weighted according to their probabilities. The possible outcomes are a winning bet on Bill, which has a value of 12 and a probability of 0.35; a winning bet on Ted, which has a value of 40 and a probability of 0.15; and a losing bet, which has a value of -10 and a probability of 0.5. The weighted average of these is $0.35 \times 2 + 0.15 \times 30 + 0.5 \times (-10) = 0.2$, so the bet has a positive expected utility, indicating (by a small margin) that you should take it.

There is also a second way we can calculate the conditional bet's expected utility. Instead of directly calculating a weighted average of all possible outcomes, we can first calculate the expected utility of each unconditional bet and take an average of these, with each unconditional bet weighted according to the probability that it would eventuate from the conditional bet. Because the result of the toss is probabilistically independent of the result of the match, the result is the same: $0.5 \times (-1.6) + 0.5 \times 2 = 0.2$. This second approach is often clearer and more convenient.

> Problem 4 (The Pet Bet). You have decided to adopt two pets from the local animal shelter, which will be delivered later today. Because you like surprises, you have asked the staff there to decide what animals to send by tossing a fair coin; they will send cats if it lands heads and dogs if it lands tails. You have been told that you will be sent two brothers, Tom and Tim, but you do not know what their species is. Because you are a compulsive gamblers, you and a friend agree to bet on Tim's species: one of you will bet

> that he is a cat, the other that he is a dog, and whoever loses must give \$100 to the winner. To make the bet even more exciting, your friend suggests the following procedure for deciding who will bet on which species: if Tom is a cat, then you must bet that Tim is a dog; is Tom is a dog, then you must bet that Tim is a cat. Is this a fair bet?

The answer is obviously no, because you are sure to lose, but your friend might defend the bet by arguing as follows. There are two unconditional bets you might make: CAT or DOG. Since Tim is equally likely to belong to either species, each bet has an expected utility of $0.5 \times 100 + 0.5 \times (-100) = 0$. Your friend argues that, as with problem 3, we can then calculate the expected utility of the conditional bet by taking a weighted average of the expected utilities of the unconditional bets, with each weighted according to its probability of eventuating from the second-order bet. Tom is equally likely to be a dog or a cat, so each first-order bet has a weighting of 0.5; thus, the expected value of the second-order bet is $0.5 \times 0 + 0.5 \times 0 = 0$, making it a fair bet.

This argument is fallacious because Tom's species is not probabilistically independent of that of his brother Tim; although each brother has a 0.5 chance of being either a cat or a dog, the conditional probability of Tom being a cat given that Tim is a cat is 1. Thus, if we wish to calculate the second-order bet's expected value as a weighted average of those of the first-order bets, then we must not use each first-order bet's unconditional expected utility, but rather its expected utility conditional on its eventuating from the second-order bet. In this case, because you are sure that Tom is a cat if and only if Tim is, each first-order bet is sure to be lost given that it eventuates from the second-order bet, and thus has an expected utility of -100, which is therefore also the second-order bet's expected utility. This illustrates a general fact about expected utilities:

> FACT. If B is a second-order bet over a set of first-order bets, then it is not necessarily the case that the expected utility of B is a weighted average of the unconditional expected utilities of those first-order bets.

My central claim is that a substantial part of the literature on mixed decisions fails to take this fact into account, making an error exactly analogous to that made by your friend in the Pet Bet.

2 Newcomb's problem and causal expected utility

> Problem 5. You are a contestant on a game show, faced with a transparent box containing $10. Sat opposite you is the game show host, who has two envelopes, one empty and one containing $100. You may, if you wish, take the contents of the box. The host will observe your choice and give you an envelope, whose contents (if any) you can keep. You confidently believe that he will give you $100 if you do not take the contents of the box, but noting if you do.

In this case, it is clear that you should not take the contents of the box: if you do so then you can expect to win only $10, whereas if you don't then you can expect to win $100.

> Problem 6 (Newcomb's problem). As in problem 5, you are faced with a transparent box containing $10. You may, if you wish, take the contents of the box. Again there is a game show host, but this time, instead of two envelopes, he has an opaque box, whose contents have already been determined and which will receive regardless of your choice. However, you confidently believe that he is an extremely reliable predictor of human behaviour, who has decided the contents of the box based on an accurate prediction of your choice (perhaps using a brain scanner and a supercomputer to simulate your decision-making process): more specifically, if he predicted you would take the $10 then he left the box empty, but if he predicted you would leave the $10 then he put $100 in the box.

From the point of view of evidential expected utility, this is indistinguishable from problem 5. You are (by stipulation) completely confident in the host's powers of foresight, so your credence in the box containing \$100 is 1 conditional on your leaving the \$10, and zero conditional on your taking it. So once again, EEU-maximization prescribes leaving the \$10 and taking just the contents of the opaque box. And indeed, this course of action has been defended on the grounds that, by taking just one box, you can expect to win \$100, whereas by taking both boxes you can only expect to win \$10.

There is, though, a crucial difference between problems 4 and 5. In problem 4, when you are choosing whether to take the \$10 it has not yet been determined whether you will get the \$100; your decision on the former matter is what determines the latter. In problem 5, by contrast, when you make your decision it is too late to exert any causal influence on the contents of the opaque box: if it contains \$100 then you will get \$100 regardless of your choice, and if it doesn't then you won't. (I am excluding, by stipulation, time travel, crystal balls, and other forms of backwards causation.) This observation leads to the following argument for taking both boxes, which has convinced most decision theorists. Either the opaque box contains \$100 or it doesn't. If it contains \$100, then taking both boxes will earn you \$110, whereas taking one will only earn you \$10. If the opaque box is empty, then two-boxing will get you \$10, and one-boxing will get you nothing. Either way, you'll be better-off by \$10 if you two-box. Granted, one-boxing will provide you with evidence in the light of which you can *expect* a higher payoff than you could expect if you took both boxes, but leaving the \$10 in order to obtain this evidence would amount to "an irrational policy of managing the news" [Lewis, 1981].

These observations motivated the development of causal decision theory (CDT), in which expected utilities are weighted, not according to conditional probabilities, but according to the probabilities of subjunctive conditionals which are intended to capture an act's propensity to causally influence the outcome of a situation. Modelling a situation (or rather, an agent's understanding of a situation) using CDT requires, in addition to act and outcome partitions, a set of *states of nature* or *dependency hypotheses* which constitute "the locus of all the agent's uncertainty"[Joyce, 1999]: the state of nature that obtains, together with the agent's act, determines what the outcome of the situation will be. The exact nature of these varies among versions of CDT, but for present purposes we will model each state as a set of (non-backtracking) subjunctive conditionals, one for each available action, specifying what would happen if that action were performed. An action's A's causal expected utility is is a weighted average of the desirabilities of the possible outcomes, with each outcome o weighted according to the agent's credence that the state of nature obtaining contains a conditional with A as antecedent and o as consequent: $CEU(A) = \sum_{o \in \mathcal{O}} Cr\left(A \mathrel{\Box\!\!\rightarrow} o\right) U\left(o\right)$.[1] In this problem there are two possible states: S_1, in which the opaque box contains \$100, represented as $\{A_1 \mathrel{\Box\!\!\rightarrow} V = 100,\, A_2 \mathrel{\Box\!\!\rightarrow} V = 110\}$ (where A_1 is the act of taking one box and A_2 the act of taking both), and S_2, in which the opaque box is empty, represented as $\{A_1 \mathrel{\Box\!\!\rightarrow} V = 0,\, A_2 \mathrel{\Box\!\!\rightarrow} V = 10\}$. This is represented graphically by the payoff matrix below. If the agent has credence p that the opaque box contains \$100, the CEU of taking one box is $Cr\left(A_1 \mathrel{\Box\!\!\rightarrow} V = 100\right) \times 100 + Cr\left(A_1 \mathrel{\Box\!\!\rightarrow} V = 0\right) \times 0 = 100p$, while by similar calculations that for taking both boxes is $100p + 10$. This reflects the fact that, whatever the contents of the box, two-boxing will always earn the agent an extra \$10.

	S_1	S_2
A_1	100	0
A_2	110	10

The value of p, of course, is not fixed. Once you decide for sure to take both boxes, you can be sure that the predictor will have foreseen this and left the opaque box empty, so that taking two boxes will then have a CEU of just 10; but since your choice has no effect on the contents of the box, taking one box will then have an even lower CEU, of zero. Conversely, if you decide for sure to take only one box you can be sure that the box will have \$100 in it, so that one-boxing has a CEU of 100; but two-boxing will then have a greater CEU still, of 110. Your deliberation provides you with informational feedback about the state of nature, and hence about the outcomes that would result from the actions available to you. But this feedback is not relevant to your decision: deciding to take both boxes gives you evidence that you are unlucky enough to have an empty box in

[1] Incidentally, if an agent is sure she will perform A then A's CEU and EEU are the same, since $Cr\left(A \mathrel{\Box\!\!\rightarrow} o\right)$ and $Cr\left(o/A\right)$ are then both equal to the agent's unconditional credence in o. Thus, once a decision has been made, we can speak unambiguously of *the* expected utility of that decision.

front of you, but gives you no reason to change your judgement that you can do better by taking both boxes than by taking only one. In the next section, we shall consider some situations in which the decision-making process provides informational feedback that does seem to be relevant to the agent's decision.

3 Decision instability

> Problem 7. You are a contestant on a game show. In front of you are two opaque boxes, A and B. You know that one box contains \$100, but you don't know which. You may choose one box to open, and keep its contents, if any; if you pick box A, you will receive an extra \$10, regardless of the box's contents. As in Problem 4, you confidently believe the host to be an extremely reliable predictor of human behaviour; and once again, you believe that the host has determined the boxes' contents in such a way as to frustrate you, this time by putting the \$100 in whichever box he foresaw you would not open.

In this case, the actions' causal utilities depend on the agent's beliefs about the boxes' contents. Specifically, if the agent has a credence of p that the money is in box A (state S_A) and $1 - p$ that it is in box B (S_B), then taking box A has a CEU of $100p + 10$ and taking box B has a CEU of $100\,(1 - p)$. It follows that taking box A has greater CEU than taking box B iff p is $^9/_{20}$ or greater; so CEU advises taking box A if your credence in its containing \$100 is greater than $^9/_{20}$, taking box B if it is less, and being indifferent if it is exactly $^9/_{20}$.

	S_A	S_B
A_A	110	10
A_B	0	100

CDT thus does not provide unqualified advice in this case, but only advice relative to an agent's credence function. We might stipulate that you have an initial credence of $^1/_2$ that the \$100 is in box A, perhaps because you have watched the show many times and the money has been in box A half the times. Given this credence, CDT then prescribes taking box A. But this prescription is *unstable*, in the following sense. If you decide to take box A, you will come to believe that you are going to do so. As in Newcomb's problem, this provides informational feedback about the likely contents of the boxes; but unlike in Newcomb's problem, the information you gain is relevant to your preference ordering over the boxes.

Specifically, when you decide to take box A you come to expect the host to have foreseen this and placed the money in box B; in the light of this, taking box B will seem more desirable, and, once you have updated your credences accordingly, this is what CDT will recommend relative to your new credence function. But if you then decide to take box B, something similar will happen: you will come to expect the host to have foreseen this and placed the money in box A, leaving you in the same predicament. CDT thus appears to give no stable advice in this situation; if you consult it for advice, plan to follow that advice, and update your credences accordingly, then the advice will inevitably be revoked. In short, each option has the property of being *causally unratifiable*: from the point of view of an agent who has decided to perform it and updated her credences accordingly, it is less desirable than some alternative. This gives rise to two worries. The first is a *normative* worry: does CDT give correct advice in this case, and if so, *which* piece of advice is correct? The second is a worry about dynamics: even if CDT does give good advice then, given the instability of that advice, how can an agent who is trying to maximise CEU ever settle on a decision rather than oscillating uselessly between the options?

The existence of cases providing no ratifiable option is noted by Gibbard and Harper [1978], who contend that this "instability of rational decision" is an oddity of the cases rather than of CDT: a claim that is of little use to those of us who wish to work out what to do in such a situation. More recently, Egan (op cit) (after whom these cases have come to be known as Egan cases) has claimed that an agent in a situation like this should do what he "confidently expects" will cause the better outcome - in the case of problem 5, this would mean taking box A, and thus expecting to get \$110, rather than taking box B and expecting to get \$100. This leads him to reject CDT, although he does not propose a theory of his own. In any case, I will not engage further with Egan's views here, but instead will examine another proposed solution, namely "mixed decisions", which Arntzenius [2008] and Wallace [2010] have both recently proposed as answers to the challenge raised by Egan cases.

4 Mixed decisions

Arntzenius objects to the performance of unratifiable acts on the basis of what he calls *Piaf's maxim*: "a rational person should not be able to foresee that she will regret her decisions"(p. 277). If an agent commits to unratifiable act and updates her credences accordingly, her updated causal utilities will be such as to make her foreseeably prefer some other act; so there is a clear sense in which she will regret her decision[2]. He seeks (unsuccessfully, as we shall see) to make regret-free actions possible in Egan cases by allowing *mixed decisions.*

A mixed decision can be thought of as a probability distribution over acts. The most (ostensibly) straightforward interpretation of this formal notion is that to make the mixed decision to perform A with probability p is simply to make a decision to act "randomly" in some sense. This is usually understood as being analogous with making a decision by means of a coin toss or some other external generator of randomness, except that the randomness is generated by some means internal to the agent. For now, we need only note that, immediately after making the mixed decision to perform A with probability p, an agent will have a credence of p that she is going to perform A.

How does allowing mixed decisions help in problem 7? To answer this we need to specify the host's behaviour towards agents who make mixed decisions. The most important

[2] Op cit, pp. 290-1. This invocation of Piaf's maxim to criticise the instability of CDT should not be conflated with his invocation of the principle of weak desire reflection (which he describes (p. 277) as a "version of Piaf's maxim") to criticise evidential decision theory. Weak desire reflection states that an agent's evaluation of an option's desirability at a given time should be equal her expectation, at that time, of her evaluation of its desirability at any given future time, provided that her evaluations change only as a result of updating her credence function according to new information. Arntzenius shows that EDT does not satisfy weak desire reflection (pp. 278-282) but CDT does (pp. 282-5), which he offers as a demonstration of CDT's superiority over EDT. His criticism of CDT's instability is based on an entirely separate application of the notion of regret, that does not involve desire reflection.

matter to clarify is whether he can foresee the "first-order" outcome of mixed decisions: for example, if a contestant takes box A as a result of a mixed decision to take box A with probability p and box B with probability q, whether the host can foresee that the contestant takes box A or only that she mixes with probabilities p and q. Call predictors who can foresee the first-order act that eventuates from a mixed decision as type-1 predictors, and those who can foresee only the assigned probabilities type-2 predictors. Likewise, call cases featuring type-1 predictors type-1 Egan cases, and those featuring type-2 predictors type-2 Egan cases. By stipulating that the host is a type-2 predictor, and specifying his behaviour when he predicts a mixed decision, we arrive at the following type-2 case.

> Problem 8. You are a contestant on a game show. In front of you are two opaque boxes, A and B. You know that one box contains \$100, but you don't know which. You may open one, and keep its contents, if any; if you pick box A, you will receive an extra \$10, regardless of the box's contents. You have the option to make a mixed decision. You believe that the host has foreseen your decision and placed the money according to the probability p he predicted you would assign to box A: if he predicted p would be greater than ${}^{11}\!/\!_{20}$ he placed the money in box B, if less than ${}^{11}\!/\!_{20}$ he placed it in box A, and if exactly ${}^{11}\!/\!_{20}$ then he randomised, putting the money in box A with probability ${}^{9}\!/\!_{20}$ and box B with probability ${}^{11}\!/\!_{20}$.

It is clear that no mixed decision assigning probability greater than or less than ${}^{11}\!/\!_{20}$ is ratifiable: if you assign box A greater probability than ${}^{11}\!/\!_{20}$ then you can be sure the money will be in box B, which you then prefer to take with certainty, and if you assign box A a lower probability than ${}^{11}\!/\!_{20}$ then you can be sure the money will be in box A, which you will likewise then prefer to take with certainty. If you assign box A a probability of exactly ${}^{11}\!/\!_{20}$ then you can expect the host to have randomised the location of the \$100, putting it in box A with probability ${}^{9}\!/\!_{20}$, so that each box has an expected value of \$55. But in problem 8 (and generally in problems involving type-2 predictors), despite the fact about expected utilities noted in section 1, the CEU of a mixed decision is simply a weighted average of the unconditional CEUs of the available pure acts. Once you have decided to assign box A a probability of $^{11}/_{20}$ and adjusted your credences accordingly, any decision, pure or mixed, therefore has the same CEU of 55, which means you have no incentive to revoke your decision: uniquely among the decisions available to you, this decision is ratifiable. If we are prepared to accept the concept of mixed decisions, then they seem to provide the means to deal with cases featuring type-2 predictors, not just in problem 8 but generally.[3] This is not true of cases featuring type-1 predictors. Consider a second variant of problem 7, featuring a type-1 predictor:

> Problem 9. You are a contestant on a game show. In front of you are two opaque boxes, A and B. You know that one box contains $100, but you don't know which. You may open one, and keep its contents, if any; if you pick box A, you will receive an extra $10, regardless of the box's contents. You have the option of making a mixed decision, but you believe the host is a type-1 predictor who has determined the boxes' contents in such a way as to frustrate you, by putting the $100 in whichever box he foresaw you would not open.

Arntzenius's proposal, which he calls Deliberative (Causal) Decision Theory (DDT), follows Skyrms [1990] in modelling deliberation as a dynamical process. On this model the agent starts out with a set of credences about the state of nature and how she is going to act, which she then adjusts according to a rule that changes the latter in a way that "seeks the good" and updates the former accordingly. For example, in problem 9, the agent might start out with equal credences that she is going to take box A and box B. Accordingly, she sets her credences in the money being in box A and in its being in box B - $Cr\left(S_A\right)$ and $Cr\left(S_B\right)$ respectively - to $^{1}/_{2}$ each. She now calculates three CEUs: that of taking box A, that of taking box B, and that of the "default mixed act" associated with her current set of credences, i.e. that which assigns each act a probability equal to her current credence that she is going to perform it. Taking box A and taking box B have CEUs of 60 and 50, respectively. DDT assumes (without argument and, as we shall see,

[3] Wallace [2010] provides a detailed treatment of the use of mixed decisions in cases involving type-2 predictors, though he also applies them - erroneously, as we shall see - to a case involving a type-1 predictor, namely the "psychopath button" case (see below). Both Arntzenius and Wallace prove that there is always a ratifiable mixed decision in these cases.

wrongly) that the CEU of a mixed act is a weighted average the unconditional CEUs of those of the component pure acts, weighted according to the probabilities it assigns them, thus assigning the default mixed act a CEU of 55. Finding box A more desirable, and box B less desirable, than the status quo, she increases her credence that she is going to take box A, decreases her credence that she is going to take box B, and adjusts her credences about the location of the \$100 accordingly. She then recalculates CEUs; if it is still the case that one of the boxes is more desirable than the status quo, she updates her credences once again. In general, this updating of credences about how she is going to act takes place according to a rule that "increases the credence only of actions whose causal utility is greater than the status quo" and "raises the sum of the probabilities of all acts with causal utility greater than the status quo". This process of deliberation continues until the agent reaches a fixed point, i.e. a point at which the agent's updating rule tells her not to adjust her credences because no pure act has a utility greater than that of the status quo. In this case, that happens when the agent has a credence of $^{11}/_{20}$ that she is going to take box A.[4]

The idea of DDT is that, having arrived at these credences, the agent can make the corresponding mixed decision, assigning probability $^{11}/_{20}$ to box A, without regretting that decision. Taking either box has a CEU of 55, and so (DDT assumes) does the mixed decision. Committing to the mixed decision will leave the agent's credences, and hence these CEUs, unchanged; having done so, the agent will therefore have no reason to wish she had chosen differently. (It is easy to see that, in cases of this sort, the equilibrium credences will be such as to make the two pure acts have equal CEU; for only then will neither pure act's CEU be greater than a weighted average of the two, and hence only

[4] Arntzenius describes his own proposal as "little more than an exposition" of Skyrms's work (p. 292), which I have therefore referred to to resolve some minor unclarity in Arntzenius's paper. Some of the details mentioned here are not explicit in Arntzenius but are explicit in Skyrms. Most importantly, Arntzenius is not explicit about the assumption that the CEU of a mixed act is a weighted average the unconditional CEUs of those of the component pure acts, but Skyrms (op cit, pp. 29-30) asserts this without argument by stating that "A state of indecision, P, carries with it an expected utility, the expectation according to the probability vector $P = \langle p_1 \ldots p_n \rangle$ of the expected utilities of the acts $A_1 \ldots A_n$" (Unlike Arntzenius, Skyrms draws a notional distinction between the agent's state of indecision about how to act, which he formalises as a probability vector, and her credences about how she is going to act. This difference is not important for present purposes.) Arntzenius also does not state explicitly what he means by the expected utility of the status quo, but Skyrms (pp. 29-30) makes it clear that the expected utility of what he terms a "state of indecision" is that of the default mixed act of that state.

then will DDT tell the agent not to readjust his credences.)

One worry that might arise at this point is that this does not really eliminate regret. Granted, when the agent has made a mixed decision but has not yet determined which pure act to perform as a result, she will have no reason to rue her choice. But in due course, the decision will have to culminate in her taking one box or the other. And once she finds herself taking a box she will come to expect the predictor to have foreseen this and placed the money in the other box; regardless of having made a mixed decision, she will now wish to revoke her decision and open the other box instead. This is a serious objection, to which I do not think either Arntzenius or Wallace has an adequate response, but it is not the objection I wish to pursue here.[5] Instead I shall pursue a separate and more fundamental objection, which is that DDT's way of calculating the expected utility of mixed acts is completely wrong when a type-1 predictor is involved.

5 DDT is wrong about type-1 cases

In section 1 I gave an example (namely the Pet Bet) of a conditional bet that has a lower expected value than any of the unconditional bets from which it is composed, because the event determining the values of the unconditional bets is not probabilistically independent of the the event that determines which unconditional bet eventuates from the conditional bet. An act in a decision situation is effectively a bet on the state of nature, and a mixed decision is a conditional bet that results in an unconditional bet through whatever chance process determines which pure act is performed as a result of the mixed decision. And just as a conditional bet's expected value is sometimes not a weighted average of the unconditional expected values of the component unconditional bets, a mixed decision's expected utility is in some cases not a weighted average of the unconditional expected utilities of the component pure acts. Problem 9 and other Egan cases, as we shall see, are cases of this sort.

In Problem 9, contrary to DDT, the mixed decision M of taking box A with probability

[5] Wedgwood [2011, §3] pursues this and related objections in more detail.

${}^{11}/_{20}$ is not ratifiable. This can be seen as follows. An agent who is sure she will do M has a ${}^{11}/_{20}$ credence that the \$100 is in box B, and hence that, if she were to take box B, she would get \$100; consequently, taking box B has an expected payoff of \$55. By similar reasoning, taking box A has the same expected payoff. Nonetheless, the agent's credence that she is *actually* going to get the \$100 is zero. For by hypothesis, she is confident that the host has foreseen which box she will actually take, and has left it empty. Since she is sure she will do M, and is also sure she won't get the \$100, it follows, by the general principle that p and q jointly imply $p \mathrel{\Box\!\!\rightarrow} q$, that she should also be sure of the subjunctive conditional that if she *were* to do M she *would* not get the \$100. In view of the probabilities assigned by M to the pure acts, it follows that she must have ${}^{11}/_{20}$ credence that if she were to do M she would take box A and get \$10, and ${}^{9}/_{20}$ credence that if she were to do M she would take box B and get nothing, giving M a CEU of ${}^{11}/_{20} \times 10 = 5.5$, making either pure act preferable to M by a margin of 49.5 utility points.

The same point can be made as an argument by contradiction. Suppose, on the contrary, that you are facing problem 9 and are able to perform mixed decisions that resolve into pure acts in a way that is independent of the location of the money, so that a mixed decision's CEU is a weighted average of those of the pure acts. It follows that your credence that M would result in each of the four possible outcomes - taking box A and getting \$110, taking box B and getting \$100, taking box A and getting \$10, and taking box B and getting nothing - can be found by multiplying your credence that M would result in your taking that box with your credence that the box's contents are as specified. Once you are sure that you will do M you thus have, for example, a credence of ${}^{9}/_{20}$ that M would result in your taking box B and a credence of ${}^{11}/_{20}$ that box B contains \$100; you should thus have a credence of ${}^{9}/_{20} \times {}^{11}/_{20} = {}^{101}/_{400}$ that, if you did M, you would take box B and get \$100. But you are sure you *are* going to do M; so by modus ponens, you should have a credence of ${}^{101}/_{400}$ that you are going to take box B and get \$100. But this contradicts our earlier stipulation that you are sure the predictor will have foreseen which box you would take and put the money in the other box. Given the existence of a predictor who can foresee your action and adjust the state of nature accordingly, it is *incoherent* to suppose that you have the ability to make a mixed decision that works in such a way that the outcome of a pure act is independent of whether it eventuates from that decision.

This observation generalises to many of the cases discussed in recent literature, which feature either type-1 predictors or natural processes that behave analogously. One such case is Death in Damascus (discussed by Gibbard and Harper [1978]): the predicament faced by a man who has a date with Death, whose appointment book contains infallible predictions of where his victims will be at the appointed time. The point of this story is that you cannot cheat death: no matter how cunningly you try to model his reasoning process in order to predict where he will be waiting for you, he will always be one step ahead of you. But this can only be the case if Death can foresee where the agent will actually go: if he could foresee only the choice of probability distribution, it would always be possible to have a 1/2 chance of cheating him by performing a mixed decision and thus leaving him uncertain of where you will be. This possibility would clearly defeat the point of the story.

The case of Psycho Paul, introduced by Egan, is analogous. In this case, Paul is contemplating whether to press the "kill all psychos" button. He thinks it would be good for all psychos to die, provided that he himself is not a psycho; but he has a very strong desire not to die himself. He thinks it very unlikely (credence 0.05) that he himself is a psycho; but he thinks pressing the button would be very strong evidence that he is a psycho: his credence in being a psycho conditional on pushing is 0.95, whereas his credence conditional on not pushing is 0.05. Either decision - pushing or not pushing - is thus one that he will regret: if he decides to push then he will come to have a high credence that he is a psycho and hence that he should not push, whereas if he decides not to push he will have a high credence that he is not a psycho, and hence that he can safely push.

	Psycho	Not Psycho
Push	-100	10
Don't Push	0	0

DDT tells Paul to make the mixed decision to push with probability $^1/_{22}$.[6] If he does so, then his credences in the possible eventualities are $^{21}/_{22}$ that he will refrain from pushing, $^{19}/_{440} \approx 0.043$ that he is a psycho who will push and thereby kill himself, and just $^1/_{440} \approx 0.0023$ that he is a non-psycho who will push and thereby enjoy a psycho-free future. The expected utility of this set of prospects is $^{-189}/_{44} \approx -4.3$: the chance of being a non-psycho who will push is 19 times smaller than that of being a psycho who will push, whereas it would need to be ten times greater in order for the prospect of living without psychos to compensate for the risk of dying.[7] As a result, the mixed act compares unfavourably with either pure act: Since Paul has $^{10}/_{11}$ credence that he is not a psycho, pushing with certainty would give a $^{10}/_{11}$ chance of the most favourable outcome and only a $^1/_{11}$ chance of dying, a far better set of prospects than those conditional on his pushing as a result of the mixed decision. Meanwhile, refraining with certainty from pushing would guarantee his survival; the only scenario in which this would be worse than the mixed act is that in which he is not a psycho but will push anyway, but this is a scenario in whose actuality he has a credence of less than a quarter of one percent: not nearly great enough to compensate for the risk of dying. Consequently, the mixed decision, like

[6] This can be shown as follows. Updating his credences according to a $^1/_{22}$ credence that he will push gives him credence $^1/_{22} \times 0.95 + {}^{21}/_{22} \times 0.05 = {}^1/_{11}$ credence that he is a psycho; the CEU of pushing is then $CU\left(PUSH\right) = {}^1/_{11} \times (-100) + {}^{10}/_{11} \times 10 = 0$, while that of not pushing is always 0. The mixed act's CEU is, according to DDT, a weighted average of those of the pure acts, and thus also equal to 0; this set of credences is thus a fixed point in Paul's deliberation.

[7]

$$\begin{aligned} Cr\left(Push \,\&\, Psycho\right) &= Cr\left(Psycho/Push\right) Cr\left(Push\right) \\ &= 0.95 \times {}^1/_{22} \\ &= {}^{19}/_{440} \end{aligned}$$

$$\begin{aligned} Cr\left(Push \,\&\, \neg Psycho\right) &= Cr\left(\neg Psycho/Push\right) Cr\left(Push\right) \\ &= 0.05 \times {}^1/_{22} \\ &= {}^1/_{440} \end{aligned}$$

$$\begin{aligned} CU\left(Push\right) &= -100 Cr\left(Psycho\right) + 10 Cr\left(\neg Psycho\right) \\ &= -100\left[Cr\left(Psycho/Push\right) Cr\left(Push\right) + Cr\left(Psycho/\neg Push\right) Cr\left(\neg Push\right)\right] \\ &\quad +10\left[Cr\left(\neg Psycho/Push\right) Cr\left(Push\right) + Cr\left(\neg Psycho/\neg Push\right) Cr\left(\neg Push\right)\right] \\ &= -100\left(0.95 \times {}^1/_{22} + 0.05 \times {}^{21}/_{22}\right) + 10\left(0.05 \times {}^1/_{22} + 0.95 \times {}^{21}/_{22}\right) \end{aligned}$$

the pure decisions, is unratifiable: as soon as he has made it, and before he has started determining whether he will actually end up pushing, he will wish he had instead picked one of the pure decisions of pushing or not pushing.

It might be objected at this point that "medical" cases like the Psycho Button are really type-2 rather than type-1 cases: Paul's decision provides evidence about whether he is a psycho, and so the greater the probability he assigns to pushing, the greater credence he should have that he is a psycho; but once he has made a given mixed decision, the eventual outcome - whether he ends up pushing or not - gives no *further* evidence about his mental health. After all, the argument might go, making a mixed decision is something like acting on the toss of a coin or the roll of a die - and whereas an infallible predictor might be able to predict the outcome of such a process, and thus correlate his behaviour with it, there is no way that the outcome of a dice roll could be correlated with Paul's mental health.

My response to this objection is twofold. Firstly, the Psycho Button case threatens to become incoherent if we think of Paul as having access to mixed decisions whose outcomes are uncorrelated with his mental health. For Paul has a high credence that he is a psycho conditional on his pushing, and a much lower credence that he is a psycho conditional on his not pushing; hence, he has a much higher credence that he will push, conditional on his being a psycho, than conditional on his not being a psycho. But if he really takes himself to be making a mixed decision whose outcome is uncorrelated with his mental health, then his credences about his action given his mental health should be equal to his unconditional credences; hence, so too should his credences about his mental health given his actions. But this simply contradicts the initial stipulation that he takes his actions to provide evidence about his credences.

It may be objected that this response makes the unwarranted presupposition that Paul has introspective knowledge of his credences, utilities, and decision procedure. But my second response is that, in any case, I don't really care whether it is *possible* to conceive of Paul as having access to mixed decisions of a sort that decorrelate his acts from the state of nature. The version of Psycho Button case that *I* am talking about is, by *stipulation*, one that works the way I have described. Nor does this interpretation originate with me; Wallace writes that "we're not assuming that Paul thinks whether he's a psychopath depends [evidentially] on what (potentially mixed) strategy he chooses, but only on the end result: what button he presses, that is." (p. 258) This seems to be the only type-1 case Wallace considers in a paper that otherwise constitutes a correct analysis of the use of mixed decisions in type-2 cases.[8] Moreover, it is clear that cases involving a reliable predictor, such as Problem 9 and Death in Damascus, must be interpreted as type-1 cases. DDT was supposed to provide a general theory for dealing with Egan cases; so if it cannot handle these cases, it is not fully adequate.

To sum up, the mixed decisions that were supposed to solve the problem of regret in Egan cases are in fact unratifiable themselves, so DDT does not provide a way for an agent facing such a situation to make a decision that she will not regret in Arntzenius's sense. Might there, though, be some *other* mixed decision available in these cases, which, unlike that recommended by DDT, *is* ratifiable? No. Consider, for example, problem 9 again. Suppose you decide to take box A with probability p (with $0 < p < 1$) and box B with probability $1 - p$, and adjust your credences accordingly. Then you have credence p that the \$100 is in box B ad $1-p$ that it is in box A. But conditional on your taking either box, you have credence 1 that the money is in the other: so you have credence p that you will take box A and get \$10, credence $1-p$ that you will take box B and get nothing, and credence zero that you will get the \$100: so the CEU of going ahead with your plan is $10p$. On the other hand, the CEU of taking A for sure is $100\,(1-p)+10$, and that of taking B for sure is $100p$: both are greater than that of the mixed act, regardless of the value of p, which is thus unratifiable. Providing a similar argument for the Psycho Button case is left as an exercise for the reader. In the next section I provide a general argument to show that a large class of Egan cases provide no ratifiable option; in the following section I extend the analysis to consider, from the point of view of an agent who has settled on

[8] Wallace also considers Death in Damascus, but it is not clear whether he is regarding it as a type-1 or type-2 case. He writes that "[a]ssuming that Death is a very good predictor, the optimal strategy is to choose at random. (If Death's powers go beyond prediction into actual prophecy, I'm less sure if the analysis applies.)" The mention of "actual prophecy" seems to refer to the possibility of death being a type-1 predictor; if so, he is correct that his analysis does not then apply.

some particular mixed decision, the expected utility of some *other* mixed decision. These sections are somewhat more technical than the rest of the paper; less technically-inclined readers may skip to section 8, in which I argue that, despite the preceding critique of DDT, there is, nonetheless, reason to think that DDT may be right about the credences with which a rational agent should finish her deliberations, though for the wrong reason.

6 A general case

So far I have argued that, in two specific cases, the mixed decisions prescribed by DDT are unratifiable, and that, in one, no mixed decision is ratifiable. In this section I will argue that any mixed decision, in any Egan case, is unratifiable, provided the case satisfies a certain constraint: namely, that the agent's conditional credence in each state, given her action, is fixed relative to her credences in how she is going to act: in other words, it is independent of her choice of (pure or mixed) act.[9] First, some more terminology. Let A be an action (pure or mixed), and P be a proposition. Then the *conditional CEU of A given P, $CU\left(A/P\right)$*, is the CEU of A, relative to $Cr\left(\cdot/P\right)$, the agent's conditional credence function given P. In particular, $CU\left(A/A'\right)$ is the CEU of A, from the point of view of an agent who has decided to perform A' and has updated her credences accordingly by conditioning. The *conditonal CEU of A*, without further specification, means the conditional CEU of A given A. When two pure acts, A_1 and A_2, are available, I will write M^p for the mixed decision that assigns probabiliity p to A_1 and $1-p$ to A_2.

The general form of an Egan case is a decision situation with two acts, A_1 and A_2, and two states, S_1 and S_2, such that both acts are unratifiable. This unratifiability is captured formally by a pair of inequalities, $CU\left(A_1/A_2\right) > CU\left(A_2/A_2\right)$ and $CU\left(A_2/A_1\right) > CU\left(A_1/A_1\right)$. As I am restricting my attention to cases in which the agent's credences in states, conditional on acts, are fixed, we can abbreviate $Cr\left(S_1/A_1\right)$ as q_1 and $Cr\left(S_2/A_2\right)$ as q_2. Denoting as v_{nm} the value of the outcome resulting from performing the act A_n in

[9] This is subject to the proviso that, in standard probability theory, probabilities conditional on an event with probability zero are not well-defined. This proviso is not relevant here, since probabilities conditional on an action with probability zero are irrelevant for expected utility calculations.

state S_m, we can calculate the CEU of each pure act, conditional both on itself and on the other pure act:

$$CU\left(A_1/A_1\right) = Cr\left(S_1/A_1\right)v_{11} + Cr\left(S_2/A_1\right)v_{12} = q_1 v_{11} + \left(1-q_1\right)v_{12}$$

$$CU\left(A_1/A_2\right) = Cr\left(S_1/A_2\right)v_{11} + Cr\left(S_2/A_2\right)v_{12} = \left(1-q_2\right)v_{11} + q_2 v_{12}$$

$$CU\left(A_2/A_2\right) = Cr\left(S_1/A_2\right)v_{21} + Cr\left(S_2/A_2\right)v_{22} = \left(1-q_2\right)v_{21} + q_2 v_{22}$$

$$CU\left(A_2/A_1\right) = Cr\left(S_1/A_1\right)v_{21} + Cr\left(S_2/A_1\right)v_{22} = q_1 v_{21} + \left(1-q_1\right)v_{22}$$

How, though, do we calculate the conditional CEU, $CU\left(M^p/M^p\right)$, of a mixed act? We have enough information to calculate the expected utility *simpliciter* associated with the agent's credence and utility functions, $\sum_{o\in\mathcal{O}} Cr\left(o\right)U\left(o\right)$. The CEU of M^p, $\sum_{o\in O} Cr\left(M^p \mathrel{\Box\!\!\rightarrow} o\right)U\left(o\right)$ differs from the agent's unconditional CEU in that each outcome o is weighted according to the agent's credence in the subjunctive conditional that o would occur if M^p were performed. In the previous section I appealed to the centring principle that p and q jointly imply $p \mathrel{\Box\!\!\rightarrow} q$. Since we are considering the case in which the agent is sure she will perform M^p, this principle has the consequence that the agent should be sure that, if a given outcome o occurs, the conditional $M^p \mathrel{\Box\!\!\rightarrow} o$ obtains. Conversely, by modus ponens, the agent should be sure that, if $M^p \mathrel{\Box\!\!\rightarrow} o$ holds, o will occur. Thus, the agent's credence in the conditional $M^p \mathrel{\Box\!\!\rightarrow} o$ should be equal to her unconditional credence in o, and so the conditional CEU of M^p is equal to the expected utility *simpliciter* associated with the credence function $Cr\left(\cdot/M^p\right)$.[10]

The agent's credence in the outcome $V = v_{nm}$, associated with act A_n and state S_m,

[10] Although strong centring follows from counterfactual excluded middle and modus ponens, it might be rejected by somebody who rejects conditional excluded middle, in which case this argument would not go through. Since the formulation of CDT we have been using presupposes counterfactual excluded middle, an alternative formulation would be needed, such as that provided by Lewis [1981]. I will show here that Lewis's formulation also implies the equivalence, argued for in the text, between an act's conditional CEU, given it is performed, and the expected utility *simpliciter* of the agent's conditional credence function, given the act is performed.

Lewis formulates the CEU of an act A as $CEU\left(A\right) = \sum_{k\in K} Cr\left(k\right)V\left(A \,\&\, k\right)$, where $V\left(\cdot\right)$ denotes evidential expected utility the elements of K are propositions specifying aspects of the state of the world, insofar as they are relevant to the outcome of the decision situation, that the agent cannot causally influence. Since we are considering the case in which the agent is sure she will perform A, this reduces to $CEU\left(A/A\right) = \sum_{k\in K} Cr\left(k\right)V\left(k\right)$. Expanding this by substituting in the definition of evidential expected utility, we have $CEU\left(A\right) = \sum_{k\in K} Cr\left(k\right)\sum_{o\in O} Cr\left(o/k\right)U\left(o\right)$. But rearranging and simplifying this gives $\sum_{o\in\mathcal{O}} Cr\left(o\right)U\left(o\right)$, the agent's expected utility *simpliciter*, which is thus equivalent to A's conditional CEU on Lewis's formulation as well as on the formulation I am using.

given that she does M^p, is $Cr\left(A_n \,\&\, S_m/M^p\right) = Cr\left(A_n/M^p\right) Cr\left(S_m/A_n \,\&\, M^p\right)$. Since, by stipulation, the agent's conditional credences in states, given acts, are independent of her choice of mixed act, this reduces to $Cr\left(A_n/M^p\right) Cr\left(S_m/A_n\right)$. The conditional CEU of M^p is thus:

$$\begin{aligned}
CU\left(M^p/M^p\right) &= \sum_{o\in\mathcal{O}} Cr\left(o/M^p\right) U\left(o\right) \\
&= Cr\left(A_1 \,\&\, S_1/M^p\right) v_{11} + Cr\left(A_1 \,\&\, S_2/M^p\right) v_{12} + \\
&\quad Cr\left(A_2 \,\&\, S_1/M^p\right) v_{21} + Cr\left(A_2 \,\&\, S_2/M^p\right) v_{22} \\
&= Cr\left(A_1/M^p\right)\left[Cr\left(S_1/A_1\right) v_{11} + Cr\left(S_2/A_1\right) v_{12}\right] + \\
&= Cr\left(A_2/M^p\right)\left[Cr\left(S_1/A_2\right) v_{21} + Cr\left(S_2/A_2\right) v_{22}\right]
\end{aligned}$$

But her conditional credences in her actions given she does M^p are $Cr\left(A_1/M^p\right) = p$ and $Cr\left(A_2/M^p\right) = 1 - p$. Denoting, once more, $Cr\left(S_n/A_n\right)$ as q_n, we thus have:

$$CU\left(M^p/M^p\right) = p\left[q_1 v_{11} + \left(1 - q_1\right) v_{12}\right] + \left(1 - p\right)\left[\left(1 - q_2\right) v_{21} + q_2 v_{22}\right]$$

But this is just a weighted sum of the causal utilities we calculated earlier for each act, *conditional on that act being performed*:

$$CU\left(M^p/M^p\right) = pCU\left(A_1/A_1\right) + \left(1 - p\right) CU\left(A_2/A_2\right)$$

The assumption underlying DDT, that the CEU of a mixed decision M^p is a weighted sum of those of the component pure acts, turns out not to be so far from the truth: unfortunately, the relevant CEU of each pure act A is not the CEU of that pure act conditional on the mixed act is performed, $CU\left(A/M^p\right)$, but its CEU given that it actually ends up getting performed, $CU\left(A/A\right)$. It is easy to show that M^p is thus unratifiable. To do this we need to calculate the CEU of a pure act, e.g. A_1, conditional on M^p:

$$\begin{aligned}
CU\left(A_1/M^p\right) &= Cr\left(S_1/M^p\right)v_{11} + Cr\left(S_2/M^p\right)v_{12} \\
&= \left[Cr\left(S_1/A_1\right)Cr\left(A_1/M^p\right) + Cr\left(S_1/A_2\right)Cr\left(A_2/M^p\right)\right]v_{11} + \\
&\quad \left[Cr\left(S_2/A_1\right)Cr\left(A_1/M^p\right) + Cr\left(S_2/A_2\right)Cr\left(A_2/M^p\right)\right]v_{12} \\
&= \left[q_1 p + \left(1-q_2\right)\left(1-p\right)\right]v_{11} + \left[\left(1-q_1\right)p + q_2\left(1-p\right)\right]v_{12} \\
&= p\left[q_1 v_{11} + \left(1-q_1\right)v_{12}\right] + \left(1-p\right)\left[\left(1-q_2\right)v_{11} + q_2 v_{12}\right]
\end{aligned}$$

But this is simply a weighted sum of the CEUs we calculated previously for A_1, conditional on itself and ocnditional on A_2:

$CU\left(A_1/M^p\right) = pCU\left(A_1/A_1\right) + \left(1-p\right)CU\left(A_1/A_2\right)$

Subtracting this from the conditional CEU previously calculated for M^p gives $CU\left(M^p/M^p\right) - CU\left(A_1/M^p\right) = \left(1-p\right)\left[CU\left(A_2/A_2\right) - CU\left(A_1/A_2\right)\right]$. But since this is an Egan case, $CU\left(A_1/A_2\right)$ is less than $CU\left(A_2/A_2\right)$; thus (provided p is less than 1) $CU\left(M^p/M^p\right) - CU\left(A_1/M^p\right)$ is negative, and we have $CU\left(A_1/M^p\right) > CU\left(M^p/M^p\right)$: given that the agent assigns A_1 probability less than 1, she prefers to perform A_1 with certainty. By an exactly parallel argument, A_2 will also be preferred to M^p, given M^p is chosen, provided that p is greater than zero. Thus, *any* mixed act, like the pure acts, is unratifiable.

7 Counterfactual mixed decisions

So far we have seen, for a range of Egan cases, how to calculate, conditional on on a given mixed decision M^p being made, the causal utilities both of M^p and of the component pure acts. But one might also ask a further question: given that an agent performs M^p, what is her CEU for some *other* mixed act, M^q, where $q \neq p$? For example, in problem 9, once you have decided on M^p, the mixed decision assigning probability p to box A and $1-p$ to box B, the CEUs are as follows. You have credence $1-p$ that the money is in box A, so taking box A has a CEU of $100\left(1-p\right)+10$; you have credence p that the money is in box B, so taking box B has a CEU of $100p$; and you have a credence of zero that mixing with probability p would result in you getting the money, so mixing with probability p has a CEU of $10p$. But what, in general, is the CEU of M^q given that M^p is performed, $CU\left(M^q/M^p\right)$?

Perhaps most straightforward are two "corner" answers. The first of these is that taking box A with probability q has a CEU conditional on M^p that is simply a weighted average of the CEUs of of the pure acts conditional on M^p: $CU\left(M^q/M^p\right) = qCU\left(A_A/M^p\right) + (1-q)\,CU\left(A_B/M^p\right)$. In this case this is equal to $q\left[100\left(1-p\right)+10\right] + (1-q)\left[100p\right]$. When p is chosen in accordance with DDT as $^{11}/_{20}$, so that the two pure acts have equal CEU given that the agent mixes with probability p, this formula implies that the CEU of mixing with probability q is independent of q: $CU\left(M^q/M^{11/20}\right) = 55$. This first corner answer would follow from the assumption that the predictor's accuracy is "counterfactually brittle": given that you actually mix with probability p he can perfectly predict your action, but switching to any other probability would completely eliminate the correlation between his prediction and your action. This answer seems unsatisfactory. For one thing, the formula it gives for $CU\left(M^q/M^p\right)$ is not the same as the formula for $CU\left(M^p/M^p\right)$, even when $p=q$. This implies a discontinuity in $CU\left(M^q/M^p\right)$ as a function of q: for example, $CU\left(M^{11/20}/M^{11/20}\right)$ is 5.5 but $CU\left(M^{11/20+\varepsilon}/M^{11/20}\right)$, for arbitrarily small ε, would be 55. Given the vague nature of real agents' credence functions, it seems implausible both that utilities should be discontinuous like this, and that an arbitrarily small shift in the probability assigned to an act should completely confound a reliable predictor.

A second corner answer is adopt the formula we derived for $CEU\left(M^p/M^p\right)$, substituting q for p: $CEU\left(M^q/M^p\right) = 10q$. This implies that the predictor's accuracy is completely counterfactually robust: even if you mixed with a probability different from your actual choice, the predictor still would have perfectly predicted this. Provided your choice of mixed decision actually has some causal effect on your outcome, then, since we are ruling out backwards causation, this is not plausible. Given you are mixing with probability p, if q is different from p, you should have some nonzero credence that the act you would perform if you mixed with probability q is different from the act you will actually perform. But you also think that the Predictor predicted the act you will actually perform; and, given the non-backtracking interpretation of counterfactuals that is relevant to CEU calculations, the prediction he would have made if you acted differently is the same as the one he actually made. So, if mixing with a different probability would change your action, it would also falsify the prediction.

If we reject these two corner answers, then it is impossible to give a general answer without making further specifications about the causal structure of a particular situation. We can, however, make progress in this direction by specifying that the relationship between a mixed decision and the act that results from it is *monotonic* in a certain sense: if a mixed decision M, assigning probability p to a pure act A, would result in A being performed, then so too would any mixed act M' assigning probability p' such that $p' > p$. This assumption of has the consequence that predictions have a certain degree of counterfactual robustness, especially in cases involving a predictor who perfectly predicts the agent's actual behaviour. To conceptualise this, it may be helpful to conceive of the agent's credence function as embodying a probability space of epistemically possible worlds. Each world in this space consists of an epistemically possible assignment of truth values to propositions (both unconditional propositions about what is actually the case, and conditional propositions). If an agent has credence p in a given proposition, then worlds at which that proposition is true will occupy a region of her epistemic possibility space with measure p.[11]

Thus, consider an agent who is certain she will perform M^p, performing A_1 with probability p and A_2 with probability $1 - p$, and is certain that her action will be correctly predicted. Then her epistemic possibility space will consist of two regions: a region with measure p, consisting of worlds at which she performs A_1 and this is correctly predicted, and a region with measure $1-p$, consisting of worlds at which she performs A_2 and this is correctly predicted. For values q different from p, the possibility space has a region with measure q of worlds at which the agent *would* perform A_1 if she *were* to make the mixed

[11] This space of epistemically possible worlds must, of course, be sharply distinguished from the modal space of metaphysically possible worlds. Indeed, each epistemically possible world comes with its own modal universe. For example, it is epistemically possible that Hesperus is not Phosphorus; in epistemically possible worlds at which this is true, it is metaphysically neecessary, and so each such world comes with a modal universe consisting entirely of worlds at which Hesperus is not Phosphorus.

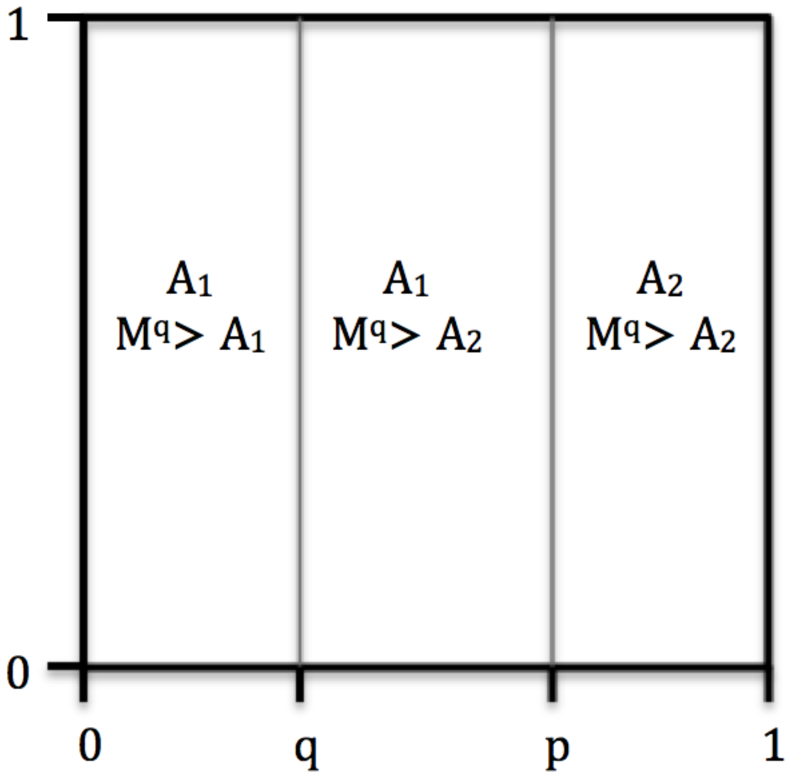

Figure 1:

decision M^p. The monotonicity assumption states that, if q is less than p, then this region consists only of worlds at which the agent actually performs A_1. This is shown in figure 1: the epistemic possibility space contains a region of measure p in which the agent performs A_1, which has a subregion of measure q in which the agent would still perform A_1 if she did M^q. (In the diagrams, "$>$" denotes a subjunctive conditional.) The remainder of the A_1 region, and the entirety of the A_2 region, together having measure $1-p$, consists of worlds at which the agent would perform A_2 if she did M^q.

In problem 9, for example, we know that, given the you actually take box A with probability p, the money is in box A if and only if the mixed decision M^p would result in you taking box B. What does this tell us about some other mixed decision M^p? If $q > p$, then M^q assigns greater probability than M^p to box A: so by monotonicity, if M^p would result in box A being opened, then so would M^q. And, if M^p would result in box A being opened, the money is in box B. So, since you have a credence of p that M^p would result in box A being opened, you should also have a credence of p that M^q would result in your opening box A and finding it empty. Your remaining credence, $1-p$, is that M^p would result in your

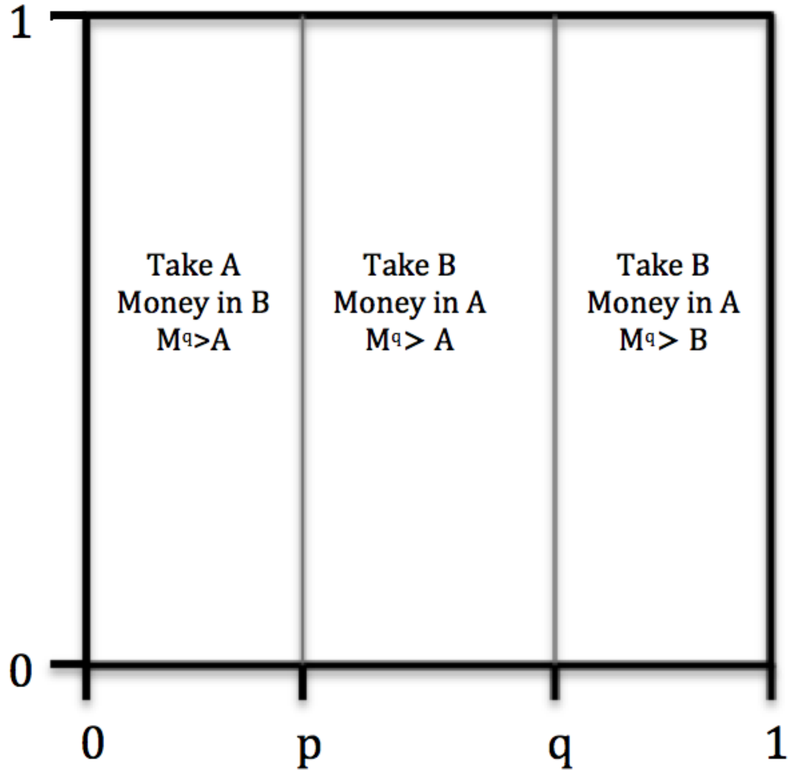

Figure 2:

taking box B, and hence that the money is in box A. Since $q > p$, you thus have credence $q - p$ that M^q would result in your opening box A and finding money in it, and credence $1 - q$ that it would result in your opening box B and finding it empty. This is shown in figure 2: in the middle region, which has measure $q - p$, doing M^q would result in you taking box A and getting the money. Thus we have $CU\left(M^q/M^p\right) = 10q + 100\left(q - p\right)$. If $q < p$ then, by similar reasoning, M^q would result in a $p - q$ likelihood of opening box B and finding money in it, giving $CU\left(M^q/M^p\right) = 10q + 100\left(p - q\right)$. The general formula is thus $CU\left(M^q/M^p\right) = 10q + 100\left|p - q\right|$, which is consistent with previous calculations for the cases where q is equal to 1, 0, or p. For $q \neq p$ we thus always have $CU\left(M^q/M^p\right) > CU\left(M^p/M^p\right)$ Moreover, maximising CEU will always require q to be 1, 0, or either, depending on p: given that M^p is performed, any other mixed decision M^q will be preferred, but the most preferred option will always be a puue act. Given the monotonicity assumption, $p - q$ represents the chance that, by switching from M^p to M^q, you would behave differently from your actual behaviour and thus confound the Predictor. The greater $p - q$, the better.

7.1 Imperfect predictors

Generalising this answer to cases involving imperfect predictors is not entirely straightforward. In the previous case, two kinds of world were epistemically possible: worlds at which you take box A but the money is in B, and worlds at which you take box B but the money is in A. Given the monotonicity assumption, there was thus only one kind of world to which to assign "extra" credence in opening a given box when considering counterfactual mixed decisions: namely, ones at which you open the other box and find it empty. Things are more complicated in cases in which the state of nature does not perfectly predict the agent's action. In Psycho Button, for example, there are four epistemic possibilities: Johnny might be a psycho and push, be a psycho and not push, not be a psycho and push, or not be a psycho and not push. Suppose Johnny is certain he will mix with probability p; In accordance with his conditional credences, he thus has a credence of $0.95p$ that he will push and is a psycho and $0.05p$ that he will push and is not a psycho. If either of these possibilities obtains then, by the monotonicity assumption, Johnny would still push, with the same outcome, if he mixed with some other probability, q, greater than p. But he must, additionally, have credence $q - p$ that he is at a world at which he does not actually push, but would if he did M^q. There are two kinds of such world to consider: worlds at which he is a psycho, and worlds at which he is not; a question thus arises as to how this extra credence that he would push should be distributed between these two kinds of world.

Possibly the most natural answer is simply to assign the extra credence to the two kinds of world in proportion with their probability measure.[12] Given that he does not (actually) push, Johnny has credence 0.05 that he is a psycho and 0.95 that he is not; on this scheme, we thus assign an additional probability of $0.05\,(q - p)$ to worlds at which Johnny is a psycho and would push if he did M^q, and $0.95\,(p - q)$ to worlds at which he is not a psycho and would push. This is shown in figure 3. Adding additional terms to represent Johnny's credence of p in worlds at which he actually pushes, and hence at which the outcome of

[12] I am *not* claiming that this is the "correct" answer. There are many possible answers, and singling one out as "correct" woudl require an analysis of the causal structure of the situation that is beyond the scope of the present discussion.

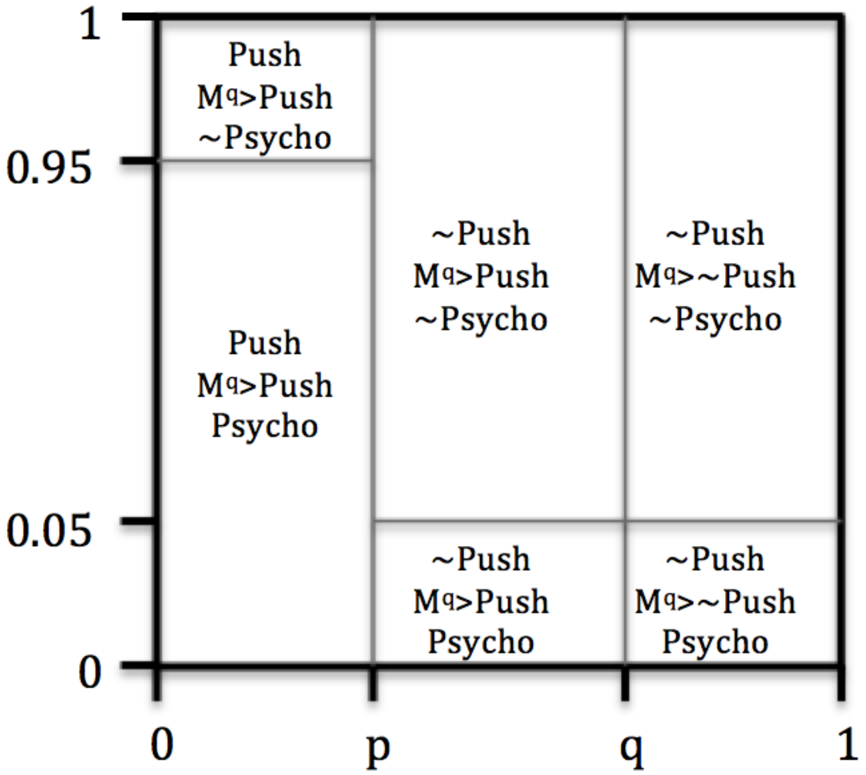

Figure 3:

M^q would be the same as the actual outcome, he thus has credence $0.95p + 0.05(q - p)$ that he is a psycho who would push if he did M^q, $0.05p + 0.95\,(q - p)$ that he is a non-psycho who would push if he did M^q, and $1 - q$ that he would not push if he did M^q. The conditional CEU for M^q given M^p when $q > p$ is thus $-100\,[0.95p + 0.05(q - p)] + 10\,[0.05p + 0.95\,(q - p)]$, which simplifies to $4.5q - 99p$.

By similar reasoning, if q is less than p then Johnny has credence $1 - p$ that he actually won't push and an additional credence of $p - q$ that he actually will push but would not push if he did M^q; this is divided between a credence of $0.95\,(p - q)$ in worlds at which he is a psycho and $0.05\,(p - q)$ in worlds at which he is not; his remaning credence, of q, is assigned to worlds at which he would push if he did M^q; he thus has credence $0.95q$ that he would push if he did M^q and is a psycho, and $0.05q$ that he would push if he did M^q and is not a psycho. So the conditional CEU for M^q given M^p when $q < p$ is $-100\,(0.95q) + 10\,(0.05q)$, which simplifies to $-94.5q$.

Again, these results are consistent with our previous calculations. In the case where

$q = p$, both formulae give the CEU of M^q as $-94.5p$, which is the same as the utility we originally calculated for M^p. Likewise, the formulae for when $q > p$ and for when $q < p$ are consistent with our previous calculations for the conditional CEUs, given M^p of the pure acts of pushing and not pushing, respectively. As in the case of perfect predictors, $CU\left(M^q/M^p\right)$ has its minimum value when $q = p$ and increases linearly as q approaches zero or one. In summary, then, considering counterfactual mixed decisions does not fundamentally change the previous analysis. Having settled on a mixed decision M^p, the agent will prefer another mixed decision M^q to the status quo, but her most prefered option will be a pure act.

8 Joyce

Like Arntzenius and Wallace, Joyce [2012] claims that an agent facing an Egan case should undergo a dynamic process of adjusting her credences that culminates in her being indifferent between the available pure acts. But Joyce's grounds for this are fundamentally epistemic. If the agent's credences do not make her indifferent between the acts, then there is at least one act such that she has positive credence that she will perform it, but which does not maximise CEU relative to her current credence function. In Psycho Button, for example, Johnny's initial credence of 0.05 that he is a psycho entails, given his conditional credences, a $^1/_{19}$ credence that he will push. But since pushing maximises CEU, it follows that he has only a $^1/_{19}$ credence that he will maximise CEU; whereas, as a rational agent, Johnny should believe himself to be a CEU-maximiser. Thus, there is an inconsistency in his beliefs, and so he should adjust them. When neither act is ratifiable, the only credence function that will avoid this inconsistency is one relative to which the two acts have equal CEU. So provided that the agent does indeed believe himself to be a CEU-maximizer, and has introspective access to his CEUs, theoretical rationality alone will lead him to form the same credences as those prescribed by DDT. This adjustment of credences, Joyce argues, is required by standard CDT: the agent's initial credences fail to incorporate information that is freely available to him, and hence the act of incorporating such information into his credences (which is assumed to be costless) is initially preferable to either of the acts explicitly specified in the decision problem.

So we have found a plausible way to salvage a substantial part of DDT, that concerned with credences. What about actions? Joyce claims, plausibly, that once the agent has adjusted her credences to incorporate the information freely available to her, maximising CEU is both necessary and sufficient for acting rationally. This means that either pure act is rationally permissible. Where Joyce goes wrong is to claim that all mixed acts are also rationally permissible, based seemingly on the misapprehension that a mixed act's CEU is a weighted average of those of the pure acts. This leads him to falsely conclude that "you can't go wrong" in Egan cases, "whatever you do". Believing the mixed decision recommended by DDT to be uniquely ratifiable, Joyce finds it necessary to back this conclusion up with arguments aiming to dispel the intuition, based on the desire to avoid regret, that this decision is normatively preferable to the alternatives. I can give ratificationism shorter shrift: there *is* no ratifiable decision, either pure or mixed, so since "ought" implies "can", ratifiability cannot be a constraint on rational choice.[13]

Where does this leave us? Once the agent's credences are in equilibrium, the default mixed decision corresponding with her credence function doesn't maximise CEU. If we hold that CEU-maximization is a *necessary* condition for rational action - a maxim that I do not intend to call into question - then an agent with these credences should prefer one or the other pure decision. This is not a major problem for Joyce's normative theory: he still gets to be a causal decision theorist and to hold that there is no normative problem with making unratifiable decisions. What might be more worrying is that we have no solution to the worry that Egan cases raise for decision dynamics: pure decisions are still unratifiable; if an agent makes a pure decision and comes to expect herself to go through with it, she will end up with credences that make her want to change her mind. Joyce mentions in passing that the mixed decision that assigns probabilities equal to the agent's equilibrium credences might be a "particularly salient way of picking". And were it indeed

[13] One might still be a lexical ratificationist, thinking that ratifiable decisions are preferable to unratifiable ones with the same CEU; but Egan (pp. 111-112) provides a convincing counterexample to this claim.

the case that this decision is ratifiable, that would constitute a solution to the dynamics problem: an agent who came to believe she would make this decision would have no need to adjust her credences, and would have no incentive to deviate from it.[14] But since no ratifiable decision is in fact available, we need some story about how an agent will get herself to execute an unratifiable decision. This seemingly leaves two possibilities. The first is that she retains her equilibrium credences until after she has irrevocably made her decision. This doesn't seem very plausible: surely, if Paul finds his hand moving towards the button, this will at least *somewhat* increase his credence that he is going to push it; and *any* increase in this credence above the equilibrium level will suffice to make him prefer to refrain. The other possibility is that any decision must be *resolute*: one that the agent will go through with *despite* coming to see it as suboptimal. This again might make us uncomfortable: why would she do such a thing?

9 Conclusion

So DDT is right about the credences with which a rational agent will complete her deliberations. On Arntzenius's official conception of mixed decisions, this amounts to a vindication of the entirety of DDT: "decision theory is theory of what *credences* one ought to have in one's actions ... not a theory that tells one which *actions* are rational and which are not". Indeed, Joyce's epistemic justification for adjusting credences fits well with this conception of decision theory: better, perhaps, than a pragmatic justification based on the desire to avoid regret. But the ambition of avoiding regret must be abandoned, and the the worry about the dynamics of executing an unratifiable decision remains unresolved.

[14] Here I am once more setting aside a worry about the dynamics of mixed decisions: namely that of how, once a mixed decision has been resolved into a pure act, the agent is to go through with that act. I think the worry I raise here, about how to go through with an unratifiable pure decision that does not result from a mixed decision but is itself, qua pure decision, the outcome of rational deliberation, is harder to set aside.

References

Frank Arntzenius. No regrets, or: Edith piaf revamps decision theory. *Erkenntnis*, 68(2): 277–297, 2008.

Rachael Briggs. Decision-theoretic paradoxes as voting paradoxes. *Philosophical Review*, 119(1), 2010.

Andy Egan. Some counterexamples to causal decision theory. *Philosophical Review*, 116 (1), 2007.

Allan Gibbard and William L. Harper. Counterfactuals and two kinds of expected utility. In *Foundations and Applications of Decision Theory*, pages 125–62. D. Reidel Publishing Company, 1978. Reprinted in Campbell & Sowden (1985); page references are to that version.

James Joyce. *The Foundations of Causal Decision Theory*. CUP, 1999.

James Joyce. Regret and instability in causal decision theory. *Synthese*, 187(1):123–145, 2012.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, July 1981.

B Skyrms. *The Dynamics of Rational Deliberation*. Harvard University Press, 1990.

David Wallace. Diachronic rationality and prediction-based games. *Proceedings of the Aristotelian Society*, 110(3):243–66, 2010.

Ralph Wedgwood. Gandalf's solution to the newcomb problem. *Synthese*, 2011. doi: 10.1007/s11229-011-9900-1. URL `http://dx.doi.org/10.1007/s11229-011-9900-1`.