

Cognitive decision theory

Hilary Greaves

October 9, 2009

DRAFT — PLEASE DO NOT CITE
COMMENTS ARE VERY WELCOME
hilary.greaves@philosophy.ox.ac.uk

Epistemic consequentialism is the analogue of ethical or practical consequentialism in the epistemic domain. Epistemic consequentialists recognise a notion of *epistemic value*, analogous to ethical value or utility: a state of affairs is one of high epistemic value for a given agent just in case it is a state of affairs in which there is a good degree of fit between that agent's beliefs and the truth. Where ethical or practical consequentialists evaluate actions such as lying (resp. carrying umbrellas) for (resp.) moral rectitude or practical rationality, the epistemic consequentialist evaluates 'epistemic acts' — acts such as believing or 'accepting' particular propositions, or adopting a particular credence function — for epistemic rationality or irrationality. Such 'acts' count as epistemically rational to the extent to which they do, or could reasonably be expected to, bring about states of high epistemic value. (The issue of whether or not they are under voluntary control is orthogonal to that of their epistemic rationality.)

In the practical domain, consequentialist ideas are nicely precisified and formalised by decision theory, in which the notion of good being pursued is captured by a utility function, and the notion of an act being 'reasonably expected to bring about states of high value' is captured by a formula for calculating any given act's expected utility. This suggests the project of developing an analogous *cognitive decision theory* to precisify and formalise epistemic consequentialism. This would be a theory in which a *cognitive utility function* makes quantitative the notion of 'degree of fit between the agent's belief states and the truth', and according to which, in any given epistemic predicament, the epistemically rational (epistemic) action is the one that maximises the subjective expectation value of this cognitive utility.¹

¹The term 'cognitive decision theory' is often applied to the study of the *psychological*

Cognitive decision theories (CDTs) have been developed fairly extensively for evaluating acts of *acceptance*: such theories take the epistemic acts under evaluation to be acts of accepting, rejecting or suspending judgment about particular propositions, and they have the agent weigh the epistemic risk of accepting falsehoods against the prospect of the epistemic gain of accepting truths. Typically, the agent also has given credences in the propositions in question, and the issue of whether or not she should accept a proposition in which she has a given credence depends on the extent to which her epistemic values recommend taking such risks. (CDTs of this type have been developed by, e.g., Levi (1967) and Maher (1993).) Such theories, however, suffer from the objection that their central notion of ‘acceptance’ is idle: all the work (in guiding action and determining future belief states, as well as in determining which propositions should now be ‘accepted’/‘rejected’) is done by the agent’s credences, and whether or not a given proposition is ‘accepted’ is at best epiphenomenal, at worst meaningless. (This objection is pressed by, e.g., Stalnaker (2002).) I think that this objection is correct: once the credences have been settled, the work is done. The interesting project, therefore, is to develop a cognitive decision theory in which the epistemic acts under evaluation are acts of *adopting a particular probability distribution as one’s credence function*.

The existing literature contains, to my knowledge, only a limited amount of speculation on the prospects for developing such cognitive decision theories of *this* type. Three attitudes can be identified (although I attribute them to particular authors only tentatively). The first (arguably implied by Thomas Kelly (2003)) is that the prospects for capturing epistemic rationality by such a theory would be dim: epistemic rationality (according to Kelly) just *is not* ‘a species of means-ends rationality, where the ends in question are epistemic ends’, as any cognitive decision theory would surely require it to be. The second (defended fairly explicitly by Maher (1993)) is that while it is not immediately obvious that such a project is doomed to failure, results we get when we try to develop it (for probabilist epistemic acts, as opposed to acts of ‘acceptance’) show that in fact it does not work. The third (apparently the view of Stalnaker (2002) and Percival (2002)) is that the idea is interesting, but requires further development.

I am sympathetic to the third of these attitudes. The objections that ‘Kelly’ and Maher raise are weak; we are not (yet?) in a position to see that cognitive decision theory cannot capture epistemic rationality. What we

processes underlying actual human decision-making. This, obviously, is a very different and unrelated use of the term.

must do is develop a probabilist cognitive decision theory as far as possible, make serious attempts to solve its problems and draw out its consequences, and only then assess the extent to which it can and cannot capture epistemic rationality. The present paper is intended as a contribution to this project.

The structure of the paper is as follows. Section 1 sketches a CDT that is structurally somewhat similar (although not isomorphic) to the practical decision theory (PDT) expounded by Savage (1972). We note that earlier work (Greaves & Wallace, 2006) shows how, within this framework, the Bayesian updating rule of conditionalisation can be justified: under independently motivated constraints on the cognitive utility function, conditionalisation is the unique updating policy that maximises expected cognitive utility. This result indicates that CDT is to be taken seriously: it seems, so far at least, to deliver results that are very much in accord with pre-existing notions of epistemic rationality. However, it is easy to come up with cognitive decision problems that our simple Savage-like framework does *not* seem able to handle: for example, problems in which the ‘state’ of the world (in Savage’s technical sense) is causally dependent on the agent’s choice of epistemic act. This sort of problem is familiar; in the history of PDT, Savage’s own theory faced such an objection. The reaction, in the context of PDT, has been a series of increasingly refined decision theories, able to deal correctly with an increasingly wide range of problem cases. These developments in PDT are reviewed (very briefly) in section 2. The task undertaken in the remainder of the paper is to develop the analogous problem cases and refinements for CDT. Section 3 presents a cognitive analogue of the ‘deterrence’ case that motivated the development of evidential PDT. Section 4 develops an evidential CDT, and applies it to the case in question. Section 5 presents an epistemic analogue of the ‘Newcomb problem’ that pushed practical decision theorists from evidential to causal PDT. Section 6 develops causal CDT, and shows how it handles the cognitive Newcomb problem. Section 7 presents a cognitive analogue of Andy Egan’s ‘Psycho Johnny’ problem, which causal CDT (as it stands) does not seem able to handle adequately. Section 8 reviews the ‘deliberational causal [practical] decision theory’ that Frank Arntzenius has recently proposed in response to the Psycho Johnny problem, develops deliberational causal CDT, and applies it to the problem case of section 7. Section 9 presents a new problem case, with no analogue in PDT, and shows that deliberational causal CDT makes the counterintuitive prediction that, in cases like this, one should violate a principle of reflection. Section 10 discusses the circumstances under which, according to deliberational causal CDT, rational agents *obey* a principle of Reflection. Section 11 discusses the prospects for replacing the thin notion of epistemic

utility with a thicker notion, if deliberational causal CDT’s violation of Belief Reflection is unacceptable (but is largely negative). Section 12 is the conclusion; I conclude that cognitive decision theory, appropriately refined as PDT has been refined, offers an elegant and plausible unifying account of probabilist epistemic norms, and that its surprising consequences should be accepted as teachings of the theory rather than counted as refutations thereof.

1 ‘Savage’

A simple cognitive decision theory operates as follows. We work with a set \mathcal{S} of States of the world. This set is, as in the case of practical decision theory, a set of mutually exclusive and jointly exhaustive propositions; the agent is uncertain as to which State obtains. Let $\mathcal{P}(\mathcal{S})$ be the set of probability distributions on \mathcal{S} . We suppose that at any given time, a given rational agent has some particular credence function, represented by some particular $Cr \in \mathcal{P}(\mathcal{S})$. Her *cognitive utility*, if she has credence function Cr and the world is in fact in state S , is given by her *cognitive utility function*, $U : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \rightarrow \mathbb{R}$.² We take a *cognitive act* to be a function from States to credence functions. We assume that epistemically rational agents maximise expected cognitive utility (ECU), where the ECU of a cognitive act $a : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{S})$ with respect to credence function p is given by

$$ECU_p(a) = \sum_{s \in \mathcal{S}} p(s) \cdot U(s, a(s)). \quad (1)$$

At first sight, the range of cognitive decision problems appears to be rather limited. Three sorts of problems suggest themselves:

1. Suppose you currently have credence function p . You receive no new information. Should you retain your current credence function, or jump to a new one?

²Note that at this point, our CDT is structurally *disanalogous* to Savage’s PDT. For Savage, the set of Consequences (i.e., the domain of the utility function) is another primitive, bearing no special relationship to the sets of States or credence functions; the agent may value anything she likes. For our ‘Savage-like’ cognitive decision theory, however, we identify the set of Consequences with the Cartesian product $\mathcal{S} \times \mathcal{P}(\mathcal{S})$. The difference reflects a difference between practical and epistemic rationality: practical utility may depend on anything whatsoever, but we require that ‘how well our agent is doing, epistemically’ depends only on (i) which State the world is in fact in and (ii) what her credence function is.

2. Suppose you currently have credence function p , and you are about to receive new information from some known set $\{E_i\}$ of mutually exclusive and jointly exhaustive propositions. Which newly available cognitive act $a : \{E_i\} \rightarrow P(\mathcal{S})$ should you perform?
3. Suppose you currently have credence function p , and you receive new information in the form of a new constraint that your posterior credence function must meet: for example, something happens to constrain your posterior credences in the elements of some given partition $\{E_1, \dots, E_n\}$ to take some particular values $\{p_1, \dots, p_n\}$. Which full posterior credence function, from those meeting this constraint, should you adopt?

We accept as a constraint on a reasonable cognitive decision theory that the answer to the first question should be ‘stick with p ’, for every $p \in P(\mathcal{S})$. That is, it must be the case, for every $p \in P(\mathcal{S})$, that the expected cognitive utility *calculated with respect to p* of having credence function p is higher than the expected cognitive utility *calculated with respect to p* of having any distinct credence function q . In other words, the epistemic utility function must be a ‘proper scoring rule’. Throughout the present paper, we will assume that the epistemic utility functions of the agents we deal with are proper scoring rules.³

The second question was analysed in (?, ?). There, we showed (*inter alia*) that, if the epistemic utility function is indeed a proper scoring rule, then *conditionalization* on the new piece of information received is the unique cognitive act that maximises expected cognitive utility. In this sense, cognitive decision theory provided a justification for the classical Bayesian rule of updating by conditionalisation.

The third question has been examined in (?, ?). Leitgeb and Pettigrew argue that the cognitive decision-theoretic perspective actually *undermines* the usual answer to this question, viz. that in such circumstances, one should update by Jeffrey conditionalization, but that it suggests an alternative updating policy which they endorse for such cases. The assessment of Leitgeb and Pettigrew’s argument is beyond the scope of this paper.

It might seem that, once the analysis of these three questions has been worked out, there will be little more to say in the realm of cognitive decision theory. But this is not the case. We can probe further, and thereby work towards a deeper understanding of the nature of epistemic rationality, by examining problem cases that seem to lie outside the scope of the simple

³Footnote to give examples of proper and improper scoring rules; further refs.

cognitive decision theory discussed so far: cases in which which State obtains depends, in one way or another, on which epistemic Act the agent performs.

For example, consider the following problem:

The Super-Evil Demon. You currently have credence function p . You are about to learn which proposition from some partition $\{E_i\}_{i \in I}$ is true. Your cognitive decision is a choice of updating rule: you must choose which posterior to move to if you learn E_i , for each $i \in I$. Normally, we know, you would choose conditionalisation. But in this case there is a catch: iff you conditionalise, the Demon will interfere with the world in numerous ways to make very many of your *other* credences (i.e. your credences in propositions logically independent of the $\{E_i\}$) as far from the truth as possible.

The idea, of course, is that this is supposed to be a case in which you might attain higher expected cognitive utility, overall, by updating *other* than by conditionalization. Yet we claimed to be able to prove that conditionalization maximised expected epistemic utility. What to think?

We will not immediately attempt a full analysis of this problem. (We make some further remarks on it in section ??.) Our point, in mentioning it at this stage, is merely that the reason this case is problematic is that it is a case in which the State of the world depends, causally and (hence) probabilistically, on the agent's choice of cognitive Act. The simple framework we have developed so far assumed that there is no such dependence, and hence gives us no guidance on the requirements of epistemic rationality in such situations.

The quandary in which this places us, as cognitive decision theorists, is familiar: practical decision theory, in the early days of its foundations, suffered a very similar quandary. Practical decision theory has since moved on: successively more refined theories have been developed, to take account of a wider and wider range of decision problems. We will therefore follow the obvious research strategy: we will recall those developments in practical decision theory, and attempt to develop their analogues for our cognitive case.

2 A brief history of practical decision theory

I take the (very!) brief history of the foundations of practical decision theory to be as follows. (For a fuller discussion, see, e.g., (Joyce, 1999).)

1. We first have Savage's (1972) theory, in naive form: the rational act is that with highest expected utility, where the expectation value is given by the formula

$$EU_p(a) = \sum_{s \in \mathcal{S}} p(s)U(s \wedge a). \quad (2)$$

2. Jeffrey pointed out that this theory could not in general deal adequately with decision problems in which States depended on Acts. For example:

Deterrence Problem. You park your car in a dodgy neighbourhood. A hooligan approaches you, and tells you (in effect) that he will probably smash your windscreen while you're gone unless you pay him \$10 now; if you do pay, he'll probably leave your car alone. The Acts are {pay, don't pay}. Your utility function is linear in money. What should you do?

In this case, Savage is supposed to be committed to a Dominance argument in favour of not paying the \$10; but this is the wrong answer.

3. Jeffrey (1965) developed an alternative theory, in which Acts, States and Consequences are all propositions about which the agent has credences, and the expected utility of an Act is calculated with respect to her *conditional* credences in States, conditional on the proposition that she performs the act in question:

$$EU_p(a) = \sum_{s \in \mathcal{S}} p(s|a)U(s \wedge a). \quad (3)$$

(In the light of the Newcomb problem, this theory later comes to be called *evidential decision theory*.)

4. Jeffrey's theory gives the right answers in cases like the Deterrence Problem. However, Jeffrey's theory gives the wrong answer in many cases in which causal and probabilistic dependence come apart — for example, the Newcomb problem:

Newcomb problem. There are two boxes, one transparent and one opaque. The transparent box contains \$1,000. The opaque box contains either nothing, or \$1,000,000. Your decision is whether to take only the opaque box, or both boxes.

There is a Predictor. The contents of the opaque box depends on what the Predictor predicted your decision would be. Iff she predicted you would take both boxes, she put nothing in the opaque box; iff she predicted you would take only the opaque box, she put \$1,000,000 in that box.

In this case, Jeffrey's theory predicts that you should take only the opaque box. But this, as is *generally* (if not universally) accepted, is the wrong answer.

5. The Newcomb problem motivates the development of *causal decision theory*. In one version of this theory (Lewis (1981)), the problems in Savage's theory are fixed by adding a stipulation about which partition of possible worlds may count as a State partition: we require the States to be 'causal dependency hypotheses'.
6. Causal decision theory deals well with all of the cases discussed so far; the mundane cases, the Deterrence Problem and the Newcomb Problem. However, it seems to give the wrong answer in cases in which there is some act A that maximises causal expected utility with respect to the agent's *initial credences*, but not with respect to the credences that she moves to if she becomes certain that she will perform A — for example, Andy Egan's (2007) Psycho Johnny problem:

Psycho Johnny. Johnny has a button. Iff he presses it, all psychopaths will die. Johnny's decision is whether or not to press the button. Johnny would like to kill all psychopaths as long as he isn't one himself, but he strongly prefers not to die. Johnny currently has very low credence that he is a psychopath. However, he also has high credence that only a psychopath would press the button.

In this case, causal decision theory predicts that Johnny should press the button, if his current credence that he is a psychopath is sufficiently low. However, Johnny is in a position to predict that once he has made this decision and updated on the proposition that he has decision to press, he will regret that decision: with respect to his *updated* credences, not pressing will have a higher causal expected utility than pressing. Vulnerability to this sort of decision instability seems to indicate irrationality.

7. Frank Arntzenius (2007) has recently suggested that a ‘deliberational decision theory’ along lines developed by Brian Skyrms (?) can solve problems like ‘Psycho Johnny’. According to DDT, a rational agent should not necessarily perform the act that has highest expected utility according to her *initial* credences; rather, she should allow her credences to develop according to a specified dynamical rule (which rule involves the expected utilities of the acts under consideration), and she should perform the mixed act with probabilities equal to her equilibrium credences. This theory gives an adequate treatment of all problem cases considered to date.

This progress in the foundations of practical decision theory can serve as a guide for the development of cognitive decision theory. In sections 3–8 we will consider analogues of the Deterrence problem, the Newcomb problem and the Psycho Johnny problem for the cognitive case, and we will develop (respectively) evidential, causal and deliberational CDT in response to these problems. To limit our task to a manageable size, here we will consider only decision problems in which the agent is to receive no new information between the time of deliberation and the time at which she adopts her chosen posterior credence function. Thus, we will be considering only cognitive acts that are ‘constant’, in the sense that the agent adopts the same posterior credence function regardless of which State is actual.

3 A cognitive analogue of Jeffrey’s deterrence problem

Consider the following case.

Promotion. You are up for promotion. Your boss, however, is a deeply insecure type, and you understand his psyche: he’s more likely to promote you if you yourself come across as lacking in confidence. Furthermore, you’re useless at play-acting, so you know that you’ll come across that way iff you really do have a low degree of belief that you’re going to get the promotion. Specifically, you think the chance of your getting the promotion will be $(1 - x)$, where x is whatever degree of belief you choose to have in the proposition P that you’ll be promoted. Your cognitive decision problem is: What credence in P is it *epistemically* rational for you to adopt?

This case is analogous to Jeffrey’s deterrence case in the following sense: if we take the State partition to be $\{P, \neg P\}$, then States depend, causally

and in consequence also probabilistically, on (cognitive) Acts.

This case poses a problem for Savage-style CDT in the following way. According to the latter, the agent has some initial credence function over the algebra generated by $\{P\}$, but there are no constraints on that credence function (aside from probabilistic coherence), and the decision theory (since it uses a proper scoring rule) will advise her to retain those same credences. Intuitively, however, some probabilistically coherent credence functions are bad: for example, if the agent has a close-to-extremal credence (i.e. credence close to 0 or to 1) in P , then her belief is almost certainly wrong, and she knows this. Intuitively, the correct response to this problem is to adopt credence $\frac{1}{2}$ in P , for then and only then will one's credences match the known chances. To have any hope of modelling this stronger constraint, a decision theory that is able to model probabilistic dependence of States on Acts. Such a theory will be developed in section 4.⁴

4 Evidential cognitive decision theory (evidential CDT)

Evidential CDT. As in the case of evidential PDT, we will regard States and (cognitive) Acts *both* as propositions. Our agent's credence function is defined on an algebra containing both states and acts (as well as conjunctions thereof, and so on). The *evidentially expected cognitive utility* (EECU) of adopting a particular final credence function \overline{Cr} , calculated with respect to current credences Cr , is given by

$$EECU(\overline{Cr}) = \sum_{s \in \mathcal{S}} Cr(s|\overline{Cr})U(s \wedge \overline{Cr}). \quad (4)$$

This theory, like Jeffrey's own, is *partition invariant*: the EECU of a given cognitive act is independent of the choice of State partition, and hence we can simply choose whichever partition is most convenient, without worrying that we are thereby smuggling in illicit assumptions.

We now wish to apply this theory to the Promotion problem developed in the previous section, and see which credence function it recommends that our agent adopt in that case.

⁴In Jeffrey's original deterrence problem, more could apparently be said about what Savage's theory would predict: there was a dominance argument favouring one act over another, i.e. the first act would lead to higher utility than the second independent of which State obtained. This feature is not present in our cognitive analogue. But that feature was anyway not essential to the key point being made, viz. that we apparently need to be able to model dependence of States on Acts.

ECDT’s analysis of the Promotion problem. We take the States to be given by $\mathcal{S} = \{P, \neg P\}$, where P is the proposition that the agent will be promoted. We assume the following cognitive utility function: for all $s \in \mathcal{S}, \overline{Cr} \in \mathcal{A}$,

$$U(s, \overline{Cr}) = -(\chi_s(P) - \overline{Cr}(P))^2, \quad (5)$$

where $\chi_P(P) = 1, \chi_P(\neg P) = 0$. (Here, \overline{Cr} itself is the proposition that the agent adopts a particular credence function; with slight abuse of notation, we also write $\overline{Cr}(\cdot)$ for the credence function concerned.) This is a proper scoring rule in the sense that each credence function *on the set of States* recommends itself; it is, however, insensitive to the accuracy of the agent’s degrees of belief concerning which epistemic act she will perform.

We take the agent’s prior credence function Cr to be such that the conditional credences $Cr(\text{state}|\text{act})$ respect the specification of the case, in the sense that for all possible final credence functions \overline{Cr} , $Cr(P|\overline{Cr}) = 1 - \overline{Cr}(P)$. (This suffices to specify all relevant features of the initial credence function. As in evidential PDT, the agent’s initial credences in her own *acts* are irrelevant to the decision theory.)

It is straightforward to establish that, with the cognitive utility function (5) and relative to any prior Cr that obeys the constraints just stated, the evidentially expected cognitive utility (as given by (4) and (5)) is maximized by setting $\overline{Cr}(P) = \frac{1}{2}$. That is, our evidential CDT delivers the intuitively correct result for our ‘Promotion’ problem case. (The details of the calculation that establishes this are given in the Appendix.)

5 A cognitive analogue of the Newcomb problem

We now wish to parallel the next development in practical decision theory: to come up with an ‘epistemic Newcomb problem’, such that (i) evidential CDT and a yet-to-be-developed *causal cognitive decision theory* (causal CDT) will give different answers; (ii) our understanding of epistemic rationality sanctions the verdict that it is causal CDT that is getting the answer right.

The structural features of a Newcomb problem are that there is some State partition with the features that

1. States depend probabilistically, *but not causally*, on Acts.
2. There is a dominance argument favouring one particular Act.
3. There is a ‘good’ state and a ‘bad’ state (‘good’ (resp. ‘bad’) in the sense that the higher ‘the probability’ of this state, the higher (resp. lower) the EU of any given act.

4. The ‘bad’ state is positively probabilistically correlated with the dominating act.

In decision problems with this structure, evidential decision theory advises performing a dominated act, because that act is evidence for the good state, and hence evidentially expected utility is higher than for the dominating act. *Causal* decision theory always prefers a dominating act, if there is one (with respect to a state partition satisfying condition (1) above).

The following cognitive decision problem exhibits features (1)–(4), and hence serves as a cognitive analogue of the Newcomb problem and will push us from evidential to causal cognitive decision theory.

Epistemic Newcomb problem. A colleague of yours is accused of embezzling funds. You happen to know that she’s guilty. You are to be interviewed by the disciplinary tribunal. Insofar as your colleague thinks that you believe her guilty, she’s had an incentive to randomise the contents of several otherwise informative files that you’ve just read (files, let us say, that the tribunal will want to examine if you give a damning testimony): specifically, she will have randomised the files with chance x , where x is her prediction for your degree of belief that she’s guilty. You know her to be a very reliable predictor of your credence functions. Your cognitive decision is a choice of a credence function on the algebra generated by the proposition G that she is guilty and various propositions $\{P_i\}_{i \in I}$ about which the files purport to be evidence.

The thought here is that your *overall* evidentially expected cognitive utility will be highest if you assign degree of belief zero to the proposition that your colleague is guilty; for, in that case, it is highly likely that the files you have just read contain reliable indicators of the truth-values of other propositions P_i , and so by also adopting the corresponding extremal credences in the P_i you will almost certainly do epistemically very well vis-a-vis those propositions. Meanwhile, conditional on the proposition that you adopt the truth-matching degree of belief 1 that your colleague is guilty, it is highly likely that the files have been randomised, and hence no degrees of belief you might adopt concerning the P_i will give you such high expected utility. Hence, evidential CDT recommends, *inter alia*, adopting credence 0 that your colleague is guilty. (These thoughts are made precise, and backed up by a calculation, in the Appendix.)

But this recommendation seems wrong. As in the practical Newcomb problem, regardless of the causal State of the world (i.e. regardless of

whether or not your colleague has randomised the files or of the truth-values of the P_i), it is true of the agent who behaves as the evidential cognitive decision theorist recommends that she *would have had higher cognitive utility* (in the appropriate, i.e. non-backtracking, sense of that counterfactual) if she had believed the colleague to be guilty (and her credences concerning the propositions on which the files contain information were just as they actually are). We need a causal cognitive decision theory.

6 Causal cognitive decision theory

In the practical case, Lewis arrives at a decision theory that gives the correct answer to both the Deterrence Problem and the Newcomb Problem by adding, to *Savage's* theory, the stipulation that the States must be causal dependency hypotheses. Rational preferences over acts are then given by a maximization of expected utility principle, with expected utility calculated using *unconditional* credences over States as in ‘Savage’s’ original theory; unconditional credences in Acts, and conditional credences $Cr(state|act)$, are not used, and need not be defined. Our next task is to parallel *this* development for the cognitive case, and apply the resulting “causal cognitive decision theory” to our Cognitive Newcomb Problem.

‘Causal dependency hypotheses’ are propositions that specify how every outcome that the agent cares about depends *causally* on her choice of act. The standard prescription for constructing causal dependency hypotheses is to identify such hypotheses with conjunctions of causal counterfactuals of the form $s_j \equiv \bigwedge_i (a_i \square \rightarrow o_{ij})$: ‘if I performed act a_1 then I would obtain outcome o_{1j} and if I performed act a_2 I would obtain outcome o_{2j} and ...’. However, in general this is not possible⁵: the current state of the world may suffice to determine only the *chances* of various outcomes, conditional on acts, and there may be no fact of the matter as to which outcome I *would* attain if I performed some given act. We therefore take the causal dependency hypotheses (instead) to be conjunctions of causal counterfactuals of the form $s_j \equiv \bigwedge_i (a_i \square \rightarrow Ch_{ij})$, where each Ch_{ij} is a distribution of chances over the propositions about which the agent is to choose her credences.

We therefore take the general setup for a cognitive decision problem to be as follows. We have some ‘base algebra’ \mathcal{D} of propositions. The set \mathcal{A} of *acts* is the set of probability functions on \mathcal{D} . The set \mathcal{S}_C of *causal states* is the set of all conjunctions $s_j \equiv \bigwedge_i (a_i \square \rightarrow Ch_{ij})$, where each Ch_{ij} is itself a

⁵Footnote about how it may be possible, but anyway no advantage to doing things any other way.

probability function on \mathcal{D} (giving the chances of propositions in \mathcal{D} 's being true if causal state s_j and act a_i is performed).

Our agent begins with some credence function Cr on the combined algebra “ $\mathcal{D} \times \mathcal{A} \times \mathcal{S}_C$ ” (by which we mean the algebra that is generated from $\mathcal{D} \cup \mathcal{A} \cup \mathcal{S}_C$ by closing under Boolean operations). (We require the credence function to satisfy the Principal Principle. It follows that the whole credence function on $\mathcal{D} \times \mathcal{A} \times \mathcal{S}_C$ is fixed by its values on $\mathcal{S}_C \times \mathcal{A}$. We do not *yet* impose any other constraints on Cr (other than probabilistic coherence, of course).) This credence function, in particular, assigns credences to elements of \mathcal{S}_C . With respect to these credences in causal states, our agent can evaluate the EU of each act in \mathcal{A} . Causal cognitive decision theory will instruct her to perform the act that has the highest causally expected cognitive utility.

We suppose that the ‘utilities’ $U(s, a)$ in Savage’s formula (1) are now given by the *objective* expectation values

$$U(s_j, a_i) = \sum_{d \in \mathcal{D}} Ch_{ij}(d)U(d, a_j) \quad (6)$$

The *causally expected cognitive utility* (CECU) of a given act a_i is then given by:

$$\begin{aligned} CECU(a_i) &= \sum_{s \in \mathcal{S}_C} Cr(s_j)U(s_j, a_i) \\ &= \sum_{s \in \mathcal{S}_C} Cr(s_j) \sum_{d \in \mathcal{D}} Ch_{ij}(d)U(d, a_i). \end{aligned}$$

It is then straightforward, for a given cognitive decision problem, to evaluate the CECU of any given act, and to show which has the highest CECU.

- In the Promotion problem, our agent is certain which causal state obtains: that is, she knows exactly how the chances Ch_{ij} of her getting promoted depend on her choice of cognitive act a_i (since it is precisely that information that is given in the specification of the case). The analysis of the decision problem thus consists solely in optimising the *objective* expected utility (6) of each act, given that particular causal state. And, mathematically, this analysis is identical to that required to solve the corresponding problem in evidential cognitive decision theory; thus causal decision theory sanctions the (intuitively correct) verdict that evidential CDT gave us for the Promotion case.
- In the epistemic Newcomb problem, our agent is *uncertain* which causal state obtains: she has some particular, and unconstrained, credence function over $\mathcal{S}_C \times \mathcal{A}$. If she obeys the Principal Principle, these

credences over $\mathcal{S}_C \times \mathcal{A}$ also induce first-order credences over the base algebra \mathcal{D} . We claim (and we prove in the Appendix) that, as a consequence of the fact that the cognitive utility function is a proper scoring rule, the cognitive act that maximizes the causally expected cognitive utility (6) is one according to which the agent’s final credence function \overline{Cr} agrees with her current credence function Cr on \mathcal{D} . In particular, it is one according to which her final credence in the proposition G that her colleague is guilty is unity: $\overline{Cr}(G) = 1$. This is despite the fact [also proved in the Appendix] that the act with the highest *evidentially* expected cognitive utility is one that assigns final credence zero to G .⁶

7 A cognitive analogue of the ‘Psycho Johnny’ problem

Causal decision theory offers a very simple, one-step prescription: perform the act that has the highest causally expected cognitive utility with respect to your initial credences, whatever those initial credences may be. This works well in simple cases, but, in cases of a certain type, leads to the agent regretting her cognitive decisions.

To see this, consider again an agent starting with some (coherent) credence function Cr on $\mathcal{D} \times \mathcal{S} \times \mathcal{A}$ that satisfies the PP. As we noted above, the acts in \mathcal{A} will be ranked by their causally expected cognitive utility (CECU) with respect to Cr . Suppose (say) that the agent becomes certain that she will perform that particular act that has the highest EU with respect to Cr , whereas Cr itself assigned nonzero credence (strictly speaking, in this case: credence density) to more than one element of \mathcal{A} . In this process, her credence function must change, but we have not yet specified exactly how, since we have not yet specified how her credences over \mathcal{S} and \mathcal{D} are to be adjusted in order to retain coherence. We require such specification to be part of the specification of the decision problem: that is, we stipulate that a cognitive decision problem is ‘well-posed’ only if there is given a precise rule for moving from old credence functions on $\mathcal{A} \times \mathcal{S} \times \mathcal{D}$ and new marginal

⁶There is, as in the practical Newcomb problem, a misguided objection to be disposed of: ‘If you’re so smart, why ain’t you got high epistemic utility?’ The answer, in this epistemic as in the practical case, is that is [epistemically] unfortunate for our agent that she is predictably [epistemically] rational, since this is a situation in which [epistemic] irrationality is [epistemically] rewarded. But, here as there, this doesn’t give her an [epistemic] reason to be [epistemically] irrational. Cf Joyce (1999, pp.151-4).

credences on \mathcal{A} to new credence functions on $\mathcal{A} \times \mathcal{S} \times \mathcal{D}$ (i.e., for revising credences in causal states in the light of arbitrary adjustments to credences in elements of \mathcal{A}), and we are interested only in solving well-posed decision problems.

Here is the point: it is not *a priori* obvious, and it is not in general true, that such a process of adjustment will leave marginal credences over \mathcal{S}_C invariant. And if credences in causal States change, then, in general, so too will the ranking of Acts by their expected utility: for example, the act $a \in \mathcal{A}$ that has highest CECU with respect to the agent's *initial* credence function Cr may not be the act that has highest CECU with respect to the new credence function Cr' that the agent moves to if she becomes certain that she will perform a and 'adjusts the rest of her credence function accordingly'. (This doesn't happen in Newcomb cases only because there, there is a dominant act.) So we have the possibility of regret: the agent performs a because that has highest CECU with respect to Cr , but ends up with a credence function according to which some other act has higher CECU.

The following is an example of a case in which this can occur.

Arrogance problem. You are wondering whether or not you are arrogant: your cognitive decision is the choice of a degree of belief in the proposition A that you are arrogant. But you take a low (resp. a high) degree of belief that one is arrogant to be evidence for the proposition that one in fact is (resp. is not) arrogant: specifically, your credence that you are arrogant, conditional on the proposition that you adopt final credence x in A , is $1 - x$, for all $x \in [0, 1]$.

Now, suppose I start off virtually certain that I will adopt final credence 0.3 (say) in A . Then, by the description of the case, I have **initial** credence 0.7 in A . There is something odd about this initial credence function (specifically, it violates a principle of reflection — on which more below). However, it is probabilistically coherent (and trivially satisfies the Principal Principle, since there are no chances in this case). Hence, nothing in our existing causal cognitive decision theory rules it out.

In *this* problem, the causal states are just $\{A, \neg A\}$; hence, since we are using a proper scoring rule, the Act that maximises CECU with respect to Cr is that of retaining credence 0.7 in A . But now suppose I become certain that I will perform this act. Since I have a stable relationship between my degrees of belief in A and my degrees of belief concerning what my *final* degree of belief in A will be, in the way stated above, to adjust my credences in causal states 'accordingly' means to become virtually certain

that the chance of A is 0.3. But now we have two problems. First, revising my beliefs to accommodate virtual certainty that I will perform the initially-recommended act has the consequence that my degrees of belief in causal states change (to 0.3 in A), and hence the initially-CECU-maximising act in question does not remain the CECU-maximizer after the agent has become certain she will perform it. Second, and relatedly, revising my beliefs to accommodate virtual certainty that I will perform the initially-recommended act has the consequence that I am not now performing it: it is impossible both to perform it and to update ‘appropriately’ on the proposition that I perform it at the same time.

Note that the credence function Cr violates the following principle of *Belief Reflection*:

Belief reflection. For any proposition P , $Cr(P) = \sum_{a \in \mathcal{A}} Cr(a) \cdot a(P)$.

It is plausible to conjecture that all problems of this ilk would go away if we restricted to credence functions satisfying Belief Reflection. Hence, one possibility would be to supplement our existing cognitive decision theory with a principle of Belief Reflection. However, to do so would be to give up an important part of the game. Recall our aim: we are attempting to sketch a cognitive decision theory, assuming no more than a suitable precisification of the (‘epistemic consequentialist’) idea that epistemic rationality consists in taking steps that can reasonably be expected to bring about epistemically good outcomes, from which as many epistemic norms as possible can be derived. We are therefore playing a different game: we will (in a suitably limited sense; see section 10) *derive* Belief Reflection from an improved cognitive decision theory.

8 Deliberational causal cognitive decision theory

All is not [yet?] hopeless; again, work in the foundations of practical decision theory has encountered and solved analogous problems, and hence can be our guide. Frank Arntzenius (2007) has recently suggested that, just as the Newcomb problem drove us from evidential to causal (practical) decision theory, similarly Egan’s ‘Psycho Johnny’ problem should drive us from our existing causal (practical) decision theory to *deliberational* causal (practical) decision theory (DCPDT).

According to DCPDT, if initial credences are unconstrained, then one should not necessarily choose the action that maximises causally expected utility with respect to one’s initial credences. Rather, one should allow

those initial credences to evolve under a ‘deliberational dynamics’. The basic idea of such a deliberational dynamics is as follows: credences are, as in evidential decision theory, defined on an algebra that includes both states and acts. Starting from arbitrary initial credences, one calculates the expected causal utility of each available act. One then revises one’s credences about which act one will in fact perform, in a (precisely specified) way that tends to increase one’s credence that one will perform the act that currently has the highest expected utility. In ‘Psycho Johnny cases’ (by definition), this process of revision alters one’s (marginal) credences about which State obtains. One then recalculates the expected utilities of the various available acts, using one’s updated credences in States. This process continues until equilibrium is reached. One performs the ‘mixed act’ corresponding to one’s equilibrium credences in Acts.

As before, let \mathcal{A} be the set of pure Acts. Let $\mathcal{P}(\mathcal{A})$ be the set of probability distributions on \mathcal{A} . In the context of *practical* decision theory, a probability distribution on \mathcal{A} can be interpreted as a *mixed act* — either the act of performing various pure Acts with particular chances, or the act of having given credences that one will perform various pure Acts (see (Arntzenius, 2007) for some discussion of the merits and drawbacks of these two interpretations).⁷

To give a full specification of a deliberational decision theory, one gives a *deliberational dynamics*. In discrete time, a deliberational dynamics is a mapping $f : \mathcal{P}(\mathcal{A}) \rightarrow \mathcal{P}(\mathcal{A})$ of the space $\mathcal{P}(\mathcal{A})$ onto itself. All that we are ultimately interested in is the *fixed points* of such functions or flows, i.e. the equilibria of the deliberational dynamics, because these fixed points are the mixed acts that are rationally permitted given some rationally permitted credence function. Hence, a ‘cheap’ way of specifying a deliberational decision theory, up to features we don’t care about, is just to characterise the equilibria. However, we will not do this, because, as noted above, the point of this whole exercise is to *derive* the constraints that hold at equilibrium. Instead, we formulate a dynamics which clearly embodies the idea of CECU-seeking, and we let the equilibria fall where they will. The dynamics we are interested in are those with the following (‘CECU-seeking’) features:

DCDT-1. $Cr'(a) > Cr(a)$ only if $CECU_{Cr}(a) > \sum_{a' \in \mathcal{A}} Cr(a') \cdot CECU(a')$.
 (At each timestep, the agent increases her credence only in actions whose causally expected cognitive utility is greater than that of the

⁷The interpretation of mixed acts is interestingly different in the epistemic case. I expect pure acts with nonextremal credences to end up being, in *some* sense(s) but not all, the cognitive analogues of mixed acts. [Discussion to follow ...]

‘status quo’, i.e. of the mixed act corresponding to her *current* credences in Acts.)

DCDT-2. Let $\mathcal{A}_+ \subset \mathcal{A}$ contain just those pure acts whose causal EU according to Cr is greater than that of the (‘status quo’) mixed act $Cr|_{\mathcal{A}}$. Then, $\sum_{a \in \mathcal{A}_+} Cr'(a) > \sum_{a \in \mathcal{A}_+} Cr(a)$. (At each timestep, the agent increases the sum of her credences in all acts that have greater causal expected utility than the status quo.)

As shown by Arntzenius, at equilibrium in the Psycho Johnny case, the agent has credence strictly between 0 and 1 that he will push the button; hence the decision theory recommends a non-pure act in this case.

8.1 Deliberational cognitive causal decision theory

The discussion of section 7 showed that in the Arrogance problem, there exist credence functions Cr on $\mathcal{S}_C \times \mathcal{A}$ that will lead to regret according to causal CDT, in the following sense: if one becomes certain that one will perform (or that one has performed) the act $a \in \mathcal{A}$ that has highest CECU with respect to one’s initial credences and then ‘adjusts the rest of one’s credence function accordingly’, in general a will not have highest CECU wrt one’s *new* credences. This is precisely the problem that the move to a deliberational version of one’s decision theory is supposed to address.

We seek, therefore, to build a *deliberational* causal cognitive decision theory that, *inter alia*, can give an adequate treatment of the Arrogance problem. Nothing in the principles (DCDT-1), (DCDT-2) above was specific to *practical* decision theory; so, we make the obvious move, i.e. we adopt the same principles as constraints on our deliberational dynamics for cognitive decision theory.

We now seek to characterise the equilibrium points. But this is easy: a credence function Cr is an equilibrium of an allowed deliberational dynamics iff no pure act has higher EU than the status quo. It follows that all pure acts in the support of Cr have equal EU to one another, and no pure act outside the support of EU has higher EU than those inside.

In the case of the Arrogance problem, it is pretty clear what the equilibrium points are: they are just those credence functions Cr with $Cr(P) = \frac{1}{2}$, and with $Cr(a) = 1$ where a is the act of retaining those same credences over \mathcal{D} . (This is proved in the Appendix.) Hence, there is a unique rationally permitted credence function in this scenario.

9 Cognitive decision theory and belief reflection

We turn now to our final type of problem case — one that, as far as I can see, has *no* analogue in the case of practical decision theory. The problems of this last type include those with the following structure:

- The agent is, and is to remain, certain of one particular causal state $s \in \mathcal{S}_C$.
- That causal state is such that the chance of some particular proposition $P \in \mathcal{D}$ takes a particular value x , independently of your chosen degree of belief in P .
- The chances of the remaining propositions in \mathcal{D} depend causally on your degree of belief in P : the further your final credence in P is from x , the more extremal will be the chances of the propositions in $\mathcal{D} - \{P\}$.

In situations with this structure, one has a cognitive incentive to set one's degree of belief in the special proposition P *unequal* to the known chances, because one can reap higher CECU from one's credences in the other propositions if their chances are more extremal.

For example, consider the following (admittedly somewhat science-fictional) case:

Cognitive bribe problem. You park your car in a dodgy neighbourhood.

An epistemic hooligan comes up to you, and informs you that he would like you to form the belief that your (obviously blue) car is not blue after all. He informs you that if you fail to do this, then he will randomise the colours of the next n cars down the street (just around the corner): specifically, he will paint those other cars blue with chance $1 \cdot \overline{Cr}(B_0) + \frac{1}{2} \cdot (1 - \overline{Cr}(B_0))$, and otherwise will paint them red. This is a serious threat, since, unfortunately for your epistemic interests, your brain is wired up to a display that announces to all the world your credence concerning the colour of your car; the hooligan can simply read your credence from this display, and then repaint the cars accordingly. If, on the other hand, you comply with his request, then he will paint the unseen cars blue (with chance 1).

The act that always has the highest CECU, in this cognitive bribe problem, is one that assigns final credence 0 to the proposition that your own car is blue (and credence 1 that all the others are blue). Hence, causal cognitive

decision theory will recommend that the agent perform this act. Accordingly, the agent moves to degree of belief 1 that she will perform this act. But in consequence of this move, she acquires credences 1 in B_0 , $1/2$ in B_i for $i \geq 1$, via the Principal Principle. (Note that this is a case in which the agent is and will remain certain of one particular causal State, so the ranking of acts by causal EU will not change during deliberation. Deliberational dynamics thus has nothing to add to the ordinary causal-decision-theoretic treatment of *this* case: it simply drives the agent to an equilibrium credence function with the features just stated, viz. $Cr(B_0) = 1$, $Cr(B_i) = 1/2$, and $Cr(\overline{Cr}) = 1$ for a final credence function \overline{Cr} with $\overline{Cr}(B_0) = 0$, $\overline{Cr}(B_i) = 1$.)

This equilibrium credence function is *coherent*; her current credences $Cr(B_0)$, $Cr(B_i)$ concerning car colours and the future credences $\overline{Cr}(B_0)$, $\overline{Cr}(B_i)$ that she believes (with credence 1) that she *will* have about those same car colours are different quantities, and there is no *incoherence* involved in their taking different values. It does, however, violate Belief Reflection.

10 Deriving Belief Reflection from DCCDT

There are two attitudes one might take to the fact that deliberational causal cognitive decision theory sometimes permits credence functions that violate Belief Reflection. The first is that violations of Belief Reflection (barring anticipated cognitive mishap) are obviously always irrational, and so this fact shows that DCCDT fails to capture epistemic rationality. The second is that DCCDT is to be trusted as a theory of epistemic rationality, and shows us that Belief Reflection is not always rational.

We wish to explore this second attitude. While it is not implausible that violations of Belief Reflection may be rational in some cases within the scope of the theory, it *is* implausible that Belief Reflection is not a requirement of epistemic rationality even in the most ordinary, mundane cases. We therefore take it to be a condition of adequacy on our cognitive decision theory that Belief Reflection can be derived from that theory under suitable auxiliary conditions.

In fact this can easily be done. We take the ‘ordinary’ or ‘mundane’ cases to be cases in which the agent’s epistemic acts are those of a *pure observer*, in the sense that her agent’s degrees of belief are causally and probabilistically independent of the propositions that those degrees of belief are about. In that case, the set \mathcal{S}_C of causal States can be identified with the base algebra \mathcal{D} . But in *that* case (as not in the Cognitive Bribe problem), the fact that the cognitive utility function is a proper scoring rule entails that there is a

unique EU-maximising act, and that it is the act of retaining one's current credences over \mathcal{D} . [This is easily proved.] Hence, any equilibrium credence function will assign degree of belief 1 to the proposition that the agent's final credences over \mathcal{D} will equal to her current credences over \mathcal{D} , and hence any equilibrium credence function will satisfy Belief Reflection.

11 Epistemic deontology

Recall our starting point. We found the idea of a consequentialist theory of epistemic rationality *prima facie* at least somewhat plausible. We expected that the idea would be best precisified by a cognitive decision theory, formally at least somewhat analogous to the practical decision theories that seem to do a good job of capturing self-interested and moral consequentialisms in the practical domain; and we deferred passing judgment on epistemic consequentialism itself until that precisification had been worked out. It is time now to return to the deferred task.

I take deliberational causal cognitive decision theory to be the best of the theories we have developed. The question is therefore how plausible DCCDT is as a theory of epistemic rationality, and whether any non-consequentialist account might be better.

DCCDT has some consequences that are likely to be surprising to the uninitiated. As we have seen, according to DCCDT there are some cases in which the agent should accept 'cognitive bribes': cases in which she should, for instance, set her degree of belief in some proposition P to zero, despite having information that guarantees that P is true (as in the problem of section 9).

Suppose we were to regard these consequences as unacceptable. The alternative theory of epistemic rationality is presumably some form of *epistemic deontology*: a theory according to which there are certain rules that must be followed for their own sakes, regardless of the epistemic consequences. We might, for instance, insist on some principle whose only purpose is to rule out rational acceptance of cognitive bribes, and separately insist that

Epistemic deontology is not incoherent. But, from the point of view of an epistemologist, to become a deontologist is to accept defeat: the consequentialist project is the search for a unifying explanation of various epistemic norms, and epistemic deontology is the position that there is no such unifying explanation. We should, therefore, be driven to deontology, if at all, only very reluctantly, by particularly stubborn and troublesome problems for

consequentialism. Recalling the comments of Stalnaker and Percival noted in the introduction to this paper, it seems to me that the development of cognitive decision theory turns the tables: *epistemic deontology* is an interesting idea, but until the details have been worked out — the most plausible deontological principles stated precisely, and their consequences investigated — we cannot accept a deontological theory for the simple reason that we have no such theory to accept: the idea requires further development.

In any case, it seems to me that the second half of the last paragraph is moot, since (it seems to me) no damning problems for cognitive decision theory have been found. The surprising consequences noted above are very plausibly regarded as surprising novel predictions of a true theory of the deep structure of epistemic rationality; it is eminently plausible, that is, that intuitions that were grown by experience with homely cases will deliver incorrect verdicts about epistemic rationality in the more outlandish scenarios that cognitive decision theory leads us to consider. Provided our theory recovers well-known epistemic norms *in the cases in which they are well-known to apply*, the existence of some surprising and counterintuitive predictions need not count against the theory. The discussion of sections 9 and ?? shows that, at least for the two well-known principles we have examined, indeed it does.

12 Conclusion

Previous work in cognitive decision theory (CDT) has, to my knowledge, explored only the simplest sorts of case, in which decision-theoretic ‘states’ are causally and probabilistically independent of epistemic acts, so that a Savage-style decision theory suffices. In this paper we have explored cognitive analogues of the main developments in the foundations of practical decision theory, and thus developed successively more refined CDTs: evidential, causal and deliberational cognitive decision theory. The main aim of this paper has been to develop cognitive decision theory in sufficient detail for its merits and drawbacks usefully to be discussed. The most refined of these, deliberational causal cognitive decision theory, deals well with all the puzzle cases we have considered. Further, from it we can derive, under suitable auxiliary constraints, such principles as Belief Reflection and Conditionalization; we can also explain why, in particular types of case that violate those constraints, it may *not* be epistemically rational to conform to the derivative principles in question. The deontological alternative seems to have the the status of a fall-back position, since it has less explanatory

power: deontologists would offer us no unifying explanation of the principles that cognitive decision theory promises to derive, instead taking each deontological injunction as primitive. As such, we should be driven to deontology only in the face of severe objections to the consequentialist alternative, and we have found no such severe objections. We therefore propose DCCDT as a partial theory of epistemic rationality.

A Calculations

A.1 Evidential CDT analysis of the Promotion problem

Claim. The act that maximises evidentially expected cognitive utility (EECU) in the Promotion problem is that of adopting credence $\frac{1}{2}$ in each of $P, \neg P$.

Proof. The EECU of an arbitrary constant cognitive act $\overline{Cr} \in \mathcal{A}$ is given by

$$\begin{aligned}
 EECU(\overline{Cr}) &= \sum_{s \in \mathcal{S}} Cr(s|\overline{Cr}) \cdot U(s, \overline{Cr}) \\
 &= Cr(P|\overline{Cr}) \cdot U(P, \overline{Cr}) + (1 - Cr(P|\overline{Cr})) \cdot U(\neg P, \overline{Cr}) \\
 &= (1 - x)(-x^2) + x(-x^2) \\
 &= -3x^2 + 3x - 1,
 \end{aligned}$$

where $x = \overline{Cr}(P)$.

This quantity takes its maximum value when $\frac{d}{dx} (EECU(\overline{Cr})) = 0$, i.e. when $x = \frac{1}{2}$.

A.2 Causal CDT analysis of the cognitive Newcomb problem

Claim. The act that maximises causally expected cognitive utility (CECU) in the cognitive Newcomb problem is one that involves adopting credence 1 in G .

Proof. We may take the causal States to be specified by the chance x that the files have been randomised.

We assume that the cognitive utility function is additive, in the sense that it can be written as a sum

$$U(s, \overline{Cr}) = U_G(s, \overline{Cr}) + U_{\{P_i\}}(s, \overline{Cr}), \quad (7)$$

where U_G is given by

$$U_G(s, \overline{Cr}) = -(1 - \overline{Cr}(G))^2, \quad (8)$$

and $U_{\{P_i\}}(s, \overline{Cr})$ is independent of $\overline{Cr}(G)$.

In this case, we can give a dominance argument for our claim, without specifying the precise form of $U_{\{P_i\}}$ or tangling with the details of *its* contribution to CECU.

The CECU of a cognitive act \overline{Cr} is given by

$$\begin{aligned} CECU_{Cr}(\overline{Cr}) &= \sum_{s \in \mathcal{S}_C} Cr(s) \cdot U(s, \overline{Cr}) \\ &= \sum_{s \in \mathcal{S}_C} Cr(s) \left(-(1 - \overline{Cr}(G))^2 + U_{\{P_i\}}(s, \overline{Cr}) \right) \\ &= -(1 - \overline{Cr}(G))^2 + \sum_{s \in \mathcal{S}_C} Cr(s) U_{\{P_i\}}(s, \overline{Cr}). \end{aligned}$$

But clearly, any act \overline{Cr} with $\overline{Cr}(G) < 1$ has strictly lower CECU than the corresponding act \overline{Cr}' given by

$$\begin{aligned} \overline{Cr}'(P_i) &= \overline{Cr}(P_i), \text{ for all } i \in I; \\ \overline{Cr}'(G) &= 1. \end{aligned}$$

A.3 Evidential CDT analysis of the cognitive Newcomb problem

Suppose that my credences in the P_i , in the absence of any information, are uniformly $\frac{1}{2}$. Of course, I ‘have information’ iff the files have not been randomised; conditional on the proposition that they have not been randomised, my credence in each P_i is unity, since the files purport to record that each P_i is true. Hence, my current credence function Cr is such that, for each $i = 1, \dots, n$,

$$\begin{aligned} Cr(P_i | Rand) &= \frac{1}{2}; \\ Cr(P_i | \neg Rand) &= 1. \end{aligned}$$

Hence, I also have

$$Cr(P_i | Chance(Rand) = x) = x \cdot \frac{1}{2} + (1 - x) \cdot 1 = 1 - \frac{x}{2}. \quad (9)$$

So, given my knowledge of the causal state, I have

$$Cr(P_i|\overline{Cr}) = 1 - \frac{\overline{Cr}(G)}{2}. \quad (10)$$

The evidentially expected cognitive utility of a given act \overline{Cr} is then given by

$$\begin{aligned} & EECU(\overline{Cr}) \\ = & -(1 - \overline{Cr}(G))^2 + \sum_{i=1}^n \{Cr(P_i|\overline{Cr})U(P_i, \overline{Cr}) + (1 - Cr(P_i|\overline{Cr}))U(\neg P_i, \overline{Cr})\} \\ = & -(1 - \overline{Cr}(G))^2 + \sum_{i=1}^n \left\{ \left(1 - \frac{\overline{Cr}(G)}{2}\right) \left(-\left(1 - \left(1 - \frac{\overline{Cr}(G)}{2}\right)\right)^2\right) + \frac{\overline{Cr}(G)}{2} \left(-\left(1 - \frac{\overline{Cr}(G)}{2}\right)^2\right) \right\} \\ = & -(1 - \overline{Cr}(G))^2 + \frac{n}{2} (1 - \overline{Cr}(G)) \overline{Cr}(G). \end{aligned}$$

Regarded as a function of $\overline{Cr}(G)$, this takes its maximum value at $\overline{Cr}(G) = 1 - \frac{n}{4}$, if this is in the interval $[0, 1]$ (i.e. if $n \leq 4$), and at 0 otherwise.

B Deliberational causal cognitive decision theory

[To follow. Include technical details of the theory: incl. possible dynamical rules.]

References

- Arntzenius, F. (2007). *No regrets*. (Available online at <http://philsci-archive.pitt.edu/archive/00003342/>)
- Egan, A. (2007). Some counterexamples to causal decision theory. *The Philosophical Review*, 116(1), 93-114.
- Greaves, H., & Wallace, D. (2006). *Justifying conditionalization: Conditionalization maximizes expected epistemic utility*. (Forthcoming in *Mind*, July 2006. Available online from <http://philsci-archive.pitt.edu>)
- Jeffrey, R. (1965). *The logic of decision*. Chicago: University of Chicago Press.
- Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge, UK: Cambridge University Press.
- Kelly, T. (2003, May). Epistemic rationality as instrumental rationality: A critique. *Philosophy and phenomenological research*, LXVI(3), 612-640.

- Levi, I. (1967). *Gambling with truth*. New York: Knopf.
- Lewis, D. (1981, March). Causal decision theory. *Australasian Journal of Philosophy*, 59(1), 5.
- Maher, P. (1993). *Betting on theories*. Cambridge University Press.
- Percival, P. (2002, July). Epistemic consequentialism. *Supplement to the proceedings of The Aristotelian Society*, 76(1), 121-151.
- Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.
- Stalnaker, R. (2002, July). Epistemic consequentialism. *Supplement to the proceedings of The Aristotelian Society*, 76(1), 153-168.