

DECISION THEORY

Decisions, Games & Logic '08

Institute for Logic, Language and Computation

University of Amsterdam

June 30, 2008

Jim Joyce

Department of Philosophy

The University of Michigan

jjoyce@umich.edu

THE GOAL OF *RATIONAL CHOICE THEORY*

Rational choice theory seeks to provide a formal account of practical rationality that will identify the conditions under which a decision maker's beliefs and desires *rationalize* the choice of actions. The hope is that this will (a) help agents make choices that provide the best means for achieving their desired ends, and (b) help people assess the overall rationality of actions (whether performed by themselves or others).

Roadmap for Lecture I:

- Decision Problems
- The Expected Utility Model
- Discussion of Some Axioms
- Preference to Probability
- Representation Theorems
- Decision Making with Imprecise Beliefs

Roadmap for Lecture II:

- Evidential and Causal Decision Theory
- Ratifiability
- Egan's "Counterexamples" to CDT
- Deliberational Dynamics and Equilibrium
- Act Probabilities and the Principle of Preference Reflection

PROSPECTS AND ACTS

We think of the agent as considering a variety of *prospects* about which he or she has preferences. It might lie within the agent's power to bring certain prospects about. These are her *acts*.

In the influential model of Savage, one thinks of prospects as having different *consequences* in various *states of the world*, so that each combination of act A and state S fixes a consequence $c_{A,S}$, which describes the result of doing A in S . In general, we can think of a prospect as a bundle of counterfactuals $A = \&_S [c_{A,S} \text{ if } S]$.

	S_1	S_2	S_n
A_1	c_{11}	$c_{12} \dots$	c_{1n}
A_2	c_{21}	$c_{22} \dots$	c_{2n}
⋮	⋮	⋮	⋮
A_m	c_{m1}	$c_{m2} \dots$	c_{mn}

General Assumptions:

- *Value/Belief Separation.* The agent's basic desires affect her evaluations of prospects only via her desires for consequences. Her beliefs figure into her evaluations only by way of her uncertainty about which state obtains.
- When the agent can control what prospects obtain, she will use her beliefs about states to select an act that provides the best available means for securing desirable consequences.

Consequences

- For Savage, the value of each consequence is independent of the act and state that bring it about, and different act/state pairs can produce different consequences.
- Each consequence is sufficiently detailed to settle *every* matter about which the agent intrinsically cares (including future contingencies).
- Prospects and states have no intrinsic value except as means for producing consequences. So, whatever might be valuable about a certain state or act should be written into the consequence. Jeffrey's Solution: consequences as prospect/state conjunctions, obviating the need for "state-dependent" preferences.

States

- States are maximally specific descriptions of things in the world that the agent cannot control.
- Disjunctions of states ("events") are the objects of subjective probability.
- The agent does not believe that she has the ability to causally influence which state obtains.

Acts

- If the agent can affect and consequence by taking future actions, these options should be explicitly represented among the acts. So, one must replace *sequences of decisions* over time with one-time, up-front choices of *strategies* or *contingency plans* than specify how the person will act when faced with whatever future decisions might confront her. (Extensive/Normal Form equivalence.)

PREFERENCES

The agent has a preference ranking over prospects. Here I think of prospects as including both acts and consequences, denoted here as f, g, \dots . (Note: Savage framed his theory in terms of act preferences. Desires for consequences are reflected in preferences for “constant acts”.)

- $f > g$ means that the agent *strictly prefers* f and g in the sense that she sees the state of affairs in which f is realized as better serving her interests, all things considered, than the state of affairs in which g is realized.
- $f = g$ means that the agent is *indifferent* between f and g in the sense that she sees f and g as serving her interests equally well.
- $f ? g$ means that the agent is *has no determinate preference* between f and g .

IMPORTANT NOTE: Preferences are not the input to decision making, they are the output! An agent's preferences represent her evaluations of prospects *after* she completes her deliberations and is in a state of reflective equilibrium.

RATIONALITY AS SUBJECTIVE EXPECTED UTILITY MAXIMIZATION

The ‘standard model’: A rational agent’s preferences should be represented as conforming to the principle of expected utility maximization. This means that

- The intensity of the agent’s (intrinsic) desires for consequences can be characterized by a *utility function* u that assigns a real number $u(c)$ to each consequence c . $u(c)$ measures the degree to which c would satisfy the agent’s desires and promote her aims. In general, c is preferred to c^* only if $u(c) > u(c^*)$.
- The strengths of the agent’s beliefs about states of the world can be characterized by a *subjective probability function* P whose values express her subjective degrees of confidence, or *credences*, at least to the extent that one event E is taken to be at least as likely as another F only if $P(E) \geq P(F)$.
- The agent’s (instrumental) desires for acts (and other prospects) can be characterized by their expected utilities computed using P and u . An act A ’s expected utility is a probability weighted average of the utilities of its consequences.

$$Exp_{P,u}(A) = \sum_S P(S) \cdot u(c_{A,S})$$

SEU says that the choice of an act is rational only if it maximizes the chooser’s *subjective expected utility*, so that $Exp_{P,u}(A) \geq Exp_{P,u}(B)$ for all alternative acts B .

Note: This is *not* proposed as a decision procedure, but as a way of assessing the results of such procedures. Rational decision makers merely act *as if* they maximize subjective expected utility.

THE RAMSEY/SAVAGE PROGRAM

We can justify *SEU* by (i) imposing axiomatic constraints on preferences, (ii) arguing that these constraints express requirements of rationality, and then (iii) proving a *representation theorem* which shows that any preference ranking that satisfies the constraints can be associated with a probability P and a utility u relative to which the decision maker maximizes expected utility.

Usually, P is assumed to be unique, and u is unique once the choice of a unit and a zero for measuring utilities are fixed. But, see below.

These representation theorems are thought of as having a number of functions, including these:

- Providing a behavioral justification for thinking that people have degrees of belief and desire.
- Establishing that rational beliefs should obey the laws of probability and that rational desires should conform to expected utility maximization.
- Clarifying the content of the requirement to maximize expected utility.

Note: I don't think they have the first function at all!

FRAME INVARIANCE

Frame Invariance. The evaluation of an act should not depend on how its consequences happen to be described. Logically equivalent redescriptions of decision problems should not alter preferences.

People often violate this constraint. Consider the following two decisions:

- You receive \$300 up front, and are then given a choice between (a) getting another \$100 for sure or (b) getting \$200 or \$0 depending on the toss of a fair coin.
- You receive \$500, but are then must choose between (a*) returning \$100 for sure or (b*) returning \$200 or \$0 depending on the toss of a fair coin.

Most people prefer the “safe” choice (a) in the first case but make “risky” choice (b*) in the second case. But, *both decisions offer a sure \$400 or a fifty-fifty chance of \$300 or \$500.*

Cognitive psychologists explain this as the result of two *irrational* tendencies of decision makers.

- *Divergence from Status Quo.* People are more concerned with *gains* and *losses*, seen as changes in the status quo, than with total well-being or overall happiness.
- *Asymmetrical Risk Aversion.* People eschew risk when pursuing *gains*, but to seek risk when avoiding losses.

ORDERING

Completeness: $f > g$ or $g > f$ or $g = f$. So, $f ? g$ is not a possibility.

- Prohibits “incommensurable” goods, which cannot be compared with respect to value.
- Prohibits judgmental “indecision” in which the agent lacks the information to judge which option is best.

Completeness is not a requirement of rationality, nor is it psychologically realistic.

Transitivity: If $f > g$ and $g \geq h$ then $f > h$; if $f \geq g$ and $g > h$ then $f > h$; if $f = g$ and $g = h$ then $f = h$.

Most objections to Transitivity go wrong by relying on less than “all-things-considered” preferences.

Hard case: “discrimination effects.” Imagine 100 bottles of wine, in which bottle- n always tastes the same as bottle- $(n+1)$, but bottle-100 tastes better than bottle-1. If taste is all that matters, then it seems as if you should be indifferent between bottle- n and bottle- $n+1$ for every n , but that you should prefer bottle-100 to bottle-1.

Response-1: Preferences are *vague*. So, for some chains of indifference $f_1 = f_2 = f_3 = \dots = f_k$ one can have $f_1 ? f_k$ (so Transitivity fails), but one will not have $f_1 > f_k$ or $f_k > f_1$. Likewise, one could have $f_1 > f_n$ even though for some k neither $f_1 > f_k$ or $f_k > f_1$. If preferences are vague, Transitivity becomes

Weak Transitivity: If $f \geq g$ and $g \geq h$, then *not* $h > f$. And, if either of these preferences is strict then *not* $h = f$.

Response-2: One can also deny that these preferences are rational. There are, arguably, reasons to prefer bottle- $n + 1$ to bottle- n : viz., that bottle $n+1$ lies higher in a sequence of wines of improving taste!

DOMINANCE

Dominance. As long as the choice of an act does not affect the probability of any state, if act A 's outcomes are at least as desirable as act B 's in every state of the world, so that $c_{A,S} \geq c_{B,S}$ for all states S , then $A \geq B$. If, in addition, $c_{A,S} > c_{B,S}$ for some S that is not certainly false (= null), then $A > B$.

The italicize independence requirement is essential. Consider

	I drive home safely.	I get in a car accident.
I have another beer	11	1
I stop drinking now	10	0

Dominance clearly does not apply here! My choice about whether or not to keep drinking clearly influences my chances of being able to drive home safely.

Question: How is this sort of independence to be characterized, and how is it reflected in an agent's preferences? We will take up these questions in the second lecture.

INDEPENDENCE AND THE SURE-THING PRINCIPLE

Let $A_E \cup B_{\sim E}$ be the composite act that yields the consequences of A when E obtains and those of B when $\sim E$ obtains.

Independence. Preference among acts that have exactly the same consequences when E is false should depend exclusively on what happens when E is true. If $A_E \cup C_{\sim E}$ is preferred to $B_E \cup C_{\sim E}$ for some acts A, B and C , then $A_E \cup D_{\sim E}$ is preferred to $B_E \cup D_{\sim E}$ for all acts D .

	S_1	S_2	S_3
A_x	c	d	x
B_x	c^*	d^*	X

Independence says that an agent's preference between A_x and B_x should not depend on what consequence goes in for x . More generally, it requires agents to have well-defined *conditional* preferences: A is preferred to B in the event of E just in case $A_E \cup C_{\sim E} > B_E \cup C_{\sim E}$ for some (hence any) C .

Sure-Thing Principle: Let E_1, E_2, \dots, E_n be mutually exclusive, collectively exhaustive events. If A is weakly preferred to B conditional on each E_i , then A is weakly preferred to B simpliciter. And, if A is strictly preferred to B conditional on some event that is not judged certainly false, then A is strictly preferred to B .

"Separability" or "Complementarity" Rationale: Rational agents should be able to form coherent preferences *conditional on E* that do not depend on what transpires when $\sim E$ since counterfactual possibilities should make no difference to the agent's enjoyment of actual outcomes.

ALLAIS' PARADOX

One chooses between A and A^* and then between B and B^* .

	0.10	0.01	0.89
A	\$1,000,000	\$1,000,000	\$1,000,000
A^*	\$5,000,000	\$0	\$1,000,000
B	\$1,000,000	\$1,000,000	\$0
B^*	\$5,000,000	\$0	\$0

Empirical studies show that people systematically violate Independence here. They ‘play it safe’ and select A over A^* in the first choice, but favor the ‘riskier’ option B^* over B in the second. The standard rationale for these choices is based on two claims (i) that there is more risk involved in choosing A^* over A than there is in choosing B^* over B , and (ii) that it is rational minimize this risk even when doing so violates Independence.

The Allais Paradox suggests that Independence rules out certain rational attitudes toward risk. Specifically, it is supposed to show that (a) risk is a component of the value of actions (as all agree), but that (b) risk is not a *separable* quantity as Independence requires value to be.

The paradox suggest that an agent need not have any fixed preference between A_x and B_x because x 's value might provide information about the relative *risks* of options, and this might legitimately affect her preferences.

ELLSBERG'S PARADOX

A ball will be drawn at random from an urn that holds 30 red balls, and 60 white or blue balls in unknown proportion. One chooses between A and A^* and then between B and B^* .

	Red $p = 1/3$	White $0 \leq p \leq 2/3$	Blue $0 \leq p \leq 2/3$
A	\$100	\$0	\$0
A^*	\$0	\$100	\$0
B	\$100	\$0	\$100
B^*	\$0	\$100	\$100

Here most people prefer A to A^* and B^* to B (thus violating Independence).

	Red $p = 1/3$	White $0 \leq p \leq 2/3$	Blue $0 \leq p \leq 2/3$
A	− \$100	\$0	\$0
A^*	\$0	− \$100	\$0
B	− \$100	\$0	− \$100
B^*	\$0	− \$100	− \$100

Here most people prefer A^* to A and B to B^* (still violating Independence).

This is because people prefer risk to uncertainty when they have something to gain, but prefer uncertainty to risk when they have something to lose. Those who regard the Ellsberg paradox as a counterexample to Independence claim that such non-separable preferences for risk over uncertainty or uncertainty over risk are rational.

THE REDESCRIPTION RESPONSE

Some, e.g., Broome, have argued the Allais and Ellsberg paradoxes are *underdescribed*.

Preferences in the Allais paradox can be rationalized by noting that, when the 0.01 event occurs, agents who choose A^* over A may feel regret (because they passed up a sure thing), while those who choose B^* over B will not feel much regret (because they probably would have ended up with \$0 anyhow). For such agents, the decision matrix really looks like this:

	0.10	0.01	0.89
A	\$1,000,000	\$1,000,000	\$1,000,000
A^*	\$5,000,000	\$0 with regret	\$1,000,000
B	\$1,000,000	\$1,000,000	\$0
B^*	\$5,000,000	\$0 with no regret	\$0

Likewise, if an agent in Ellsberg paradox feels uneasy when potential gains are riding on uncertain prospects (or losses are riding on risky prospects), then the correct description of her problem is this:

	Red	White	Blue
A	\$100	\$0	\$0
A^*	\$0 + uneasiness	\$100 + uneasiness	\$0 + uneasiness
B	\$100 + uneasiness	\$0 + uneasiness	\$100 + uneasiness
B^*	\$0	\$100	\$100

If these tables are descriptively accurate, then neither the Allais or Ellsberg paradoxes provide a counterexample to Independence.

A WORRY ABOUT REDESCRIPTION (PARTLY FOUNDED)

Some object to redescription on the grounds that it trivializes the theory. One can respond to *any* violation of Independence (or any other axiom) by claiming that consequences are underdescribed.

E.g., consider a man who pays to trade an apple for an orange, pays more to trade the orange for a pear, pays even more to trade the pear for the original apple, and then pays to trade the apple for the orange, and so on. This can be reconciled with Transitivity by redescribing his options as “having an apple,” “having an orange that was obtained by trading an apple,” “having a pear that was obtained by trading an orange that was obtained by trading an apple,” “having an apple that was obtained by trading a pear that was obtained by trading an orange that was obtained by trading an apple,”... If the value of the fruit increases the more times it is traded these preferences can be transitive. But, of course, this is entirely *ad hoc*.

Response: Of course, if the redescriptions do not accurately capture what the agent cares about, then they are wrong. But, the possibility of giving incorrect descriptions does not trivialize the theory.

Compare Newtonian Mechanics: Any set of observable motions can be explained if one is willing to postulate the right forces, but the theory is not thereby trivialized.

A deeper problem is that redescription ties our ability to respond to intuitive counterexamples like Allais and Ellsberg to substantive empirical claims about people’s psychology. (For what it’s worth, I don’t find either redescription plausible).

It’s better to show that (a) risk is separable (see Rothschild & Stiglitz reference), and (b) identify some general tendency to see it non-seperable in certain cases (the endowment effect?).

COMPARATIVE PROBABILITY

A *wager* on event E is an act $c_E \cup d_{\sim E}$ that will produce the desirable consequence c in every state consistent with E and the undesirable consequence d in every state consistent with $\sim E$. Intuitively, a person should prefer such a wager more strongly the more likely she takes E to be.

Comparative Probability. Assuming $c > d$, if the agent prefers $c_E \cup d_{\sim E}$ to $c_F \cup d_{\sim F}$, she must also prefer $c^*_E \cup d^*_{\sim E}$ to $c^*_F \cup d^*_{\sim F}$ for any consequences with $c^* > d^*$.

CP can seem implausible when values of consequences seem to vary with the world's state. Suppose that c and d are monetary fortunes that one might have in ten years, $c = \$500,000$ and $d = \$300,000$. Let E and F be hypotheses about the cumulative rate of inflation over the decade: E puts the figure at 60%, while F puts it at 10%. Even if one regards E as the more probable hypothesis, one might still prefer to wager on F since one's fortune will be worth more in the event that F is true.

Savage: In all such cases, outcomes are underdescribed.

THE ROAD FROM PREFERENCE TO PROBABILITY

Step 1: Require rational preference rankings to satisfy Framing, Completeness, Transitivity, Dominance, Independence and Comparative Probability.

Step 2: Postulate a connection between degrees of belief and preferences for wagers.

Comparative (Savage, de Finetti). The agent is *more confident* in E than in F , written $E .>. F$, when $c_E \cup d_{\sim E} > c_F \cup d_{\sim F}$ for some, hence, any outcomes with $c > d$.

Quantitative (Ramsey, de Finetti). If we are able to assign utilities to outcomes a , b and c , and if the agent is indifferent between having c for certain or having [a if E , b if $\sim E$] with $u(a) > u(b)$ then agent's degree of belief in E is $b(E) = [u(c) - u(b)]/[u(a) - u(b)]$.

BIG QUESTION: What is the status of these principles? Are they definitions of comparative probabilities or degrees of belief? NO! Are they descriptive principles that explain how people actually use their beliefs in the formulation of preferences? NO! Are they normative principles that relate belief to preferences in rational agents? YES!

Step 3. Use the theory of rational preference to show that \succeq or b obey the laws of probability.

Qualitative Approach (de Finetti):

- $E \vee \sim E \succeq E \succeq E \& \sim E$ (from Dominance)
- \succeq and \succ are transitive (from Transitivity)
- If G is contrary to both E and F , then $E \succeq F$ iff $(E \vee G) \succeq (F \vee G)$. (from Independence)
- $E \succeq F$ or $E = F$ or $E \preceq F$

This requires the additional assumptions that (i) for each E and F there is a feasible wagers of the form $[a \text{ if } E, b \text{ if } \sim E]$ and $[a \text{ if } F, b \text{ if } \sim F]$ that can be used to compare them, and (ii) that the agent has a determinate preference among these wagers. This is implausible (Aumann, letter to Savage).

This is not sufficient to guarantee that the agent's beliefs can be represented by a unique probability P such that $E \succeq F$ iff $P(E) \geq P(F)$. Indeed, there need be no probability with this feature. To get the mere existence of such a probability one needs to add substantive "richness" assumptions that require the agent's preferences to be defined over a very rich set of prospects.

Quantitative Approach.

- $b(E \vee \sim E) = 1 \geq b(E) \geq 0 = b(E \& \sim E)$ (from Dominance)
- *Additivity.* When E and F are contraries $b(E \vee F) = b(E) + b(F)$ (from Independence)

So, degrees of belief obey the law of probability.

REPRESENTATION THEOREM

Existence of Subjective Expected Utility Representations. For any preference ranking that obeys the axioms there will exist at least one probability/utility pair (\mathbf{P}, \mathbf{u}) such that

- \mathbf{P} represents the agent's beliefs: $E \succeq F$ only if $\mathbf{P}(E) \geq \mathbf{P}(F)$.
- \mathbf{u} represents the agent's (intrinsic) desires for consequences: for any consequences c, d , act A and event E , $c_E \cup A_{\sim E}$ is weakly preferred to $d_E \cup A_{\sim E}$ only if $\mathbf{u}(C) \geq \mathbf{u}(C^*)$.
- $Exp_{\mathbf{P}, \mathbf{u}}$ accurately represents the agent's (instrumental) desires for prospects: A is weakly preferred to A^* only if $Exp_{\mathbf{P}, \mathbf{u}}(A) \geq Exp_{\mathbf{P}, \mathbf{u}}(A^*)$.

Uniqueness. If the set of prospects is sufficiently rich, and if the agent's beliefs and desires are sufficiently determinate to dictate preferences among all these prospects, then \mathbf{P} and \mathbf{u} are unique and "only if" can be replaced by "if and only if" in each bullet point above.

Two Issues:

1. Do representation theorems like this actually show that rational degrees of belief must obey the laws of probability?
2. What does decision theory look like if uniqueness fails?

ZYNDA'S (OFT REPEATED) OBJECTION

Zynda, and others, have argued that a representation theorem *cannot* show that rational degrees of belief must obey the laws of probability. For suppose that $Exp_{P,u}$ represents an agent's preferences, then so does

$$\sum_S \mathbf{P}^*(S) \cdot \mathbf{u}^*(c_{A,S})$$

where $\mathbf{P}^*(E) = [\mathbf{P}(E)]^2$ and $\mathbf{u}^*(c_{A,S}) = \mathbf{u}(c_{A,S})/\mathbf{P}(S)$. But, the function \mathbf{P}^* does not satisfy the laws of probability! It does not obey Additivity, but rather this series of laws (when E, F, G are contraries):

$$\mathbf{P}^*(E \vee F) = \mathbf{P}^*(E) + \mathbf{P}^*(F) + 2(\mathbf{P}^*(E)\mathbf{P}^*(F))^{1/2}$$

$$\mathbf{P}^*(E \vee F \vee G) = \mathbf{P}^*(E) + \mathbf{P}^*(F) + \mathbf{P}^*(G) + 2(\mathbf{P}^*(E)\mathbf{P}^*(F))^{1/2} + 2(\mathbf{P}^*(E)\mathbf{P}^*(G))^{1/2} + 2(\mathbf{P}^*(F)\mathbf{P}^*(G))^{1/2}$$

So, given that there is no reason to in the theory of preference to favor the one representation over the other, it follows that the representation theorem does not establish that the probability \mathbf{P} or the non-probability \mathbf{P}^* really measures the strength of the agent's beliefs.

Note, more generally, that this argument works equally well with any $\mathbf{P}^*(E) = f(\mathbf{P}(E))$ and $\mathbf{u}^*(c_{A,S}) = \mathbf{P}(S) \cdot \mathbf{u}(c_{A,S})/\mathbf{P}^*(S)$ for f any increasing function from $[0,1]$ into \mathfrak{R} .

Some of these utilities might be a little weird. But, some seem fine, e.g., when f is a positive linear transformation (Zynda's actual case) or when $\mathbf{u}^*(c_{A,S}) = \mathbf{u}(c_{A,S})/\mathbf{P}(S)$.

WHY THE OBJECTION FAILS

The objection does not appreciate that the form of the additivity law depends on our measurement conventions. Consider three such conventions, one in which we use probabilities, one in which we use a positive linear transform of probability $P^* = m \cdot P + b$ with $m > 0$, and one in which we use a squared probability $P^* = P^2$.

Preferences	Probability	Linear Transform	Squares Probability
$c_E \cup d_{\sim E} = c_E \cup c_{\sim E}$	$P(E) = 1, P(\sim E) = 0$	$P(E) = m + b, P(\sim E) = b$	$P(E) = 1, P(\sim E) = 0$
$c_E \cup d_{\sim E} = c_{\sim E} \cup d_E$	$P(E) = P(\sim E) = 1/2$	$P(E) = P(\sim E) = 1/2(b + m)$	$P(E) = P(\sim E) = 1/4$

- In the P representation, additivity says $P(E \vee F) = P(E) + P(F)$ for $E \perp F$.
- In the linear transform representation, additivity says $P^*(E \vee F) = P^*(E) + P^*(F) - P^*(E \& F)$. *This is still additivity*, just without the assumption that $P^*(E \& F)$ is zero when $E \perp F$.
- If the squared P representation, additivity says $P^*(E \vee F) = P^*(E) + P^*(F) + 2(P^*(E)P^*(F))^{1/2}$. Again, this is still *additivity*, but the version appropriate for a squared probability.

General Point: The representation theorem establishes that preferences can be represented, in the form $\sum_S P^*(S) \cdot u^*(c_{A,S})$ by any one of a family of (P^*, u^*) pairs. But, the pairs in this family are not arbitrary. Each P^* is a increasing function f of a *probability* function, and each obeys a law that is equivalent to additivity once the effects its particular f are taken into account.

JETTISONING COMPLETENESS, LOSING UNIQUENESS

Completeness is entirely implausible. If we jettison it, we need to represent beliefs and desires by *sets* of probability/utility pairs (P, u) .

- The person is determinately more confident in E than in E^* iff $P(E) > P(E^*)$ for *all* probabilities that appear in her representing set; she determinately prefers c to c^* iff $u(c) > u(c^*)$ for *all* utilities in her representing set; she determinately prefers A to A^* iff $Exp_{P,u}(A) \geq Exp_{P,u}(A^*)$ for *all* (P, u) pairs in her representing set.

Issue: How does the decision theory work? Suppose $\mathcal{P} = \{0.8 \geq P(\text{Rain}) \geq 0.1\}$.

	Rain	No Rain		Rain	No Rain
A	15	- 10		B	- 10
$\sim A$	0	0		$\sim B$	15
				0	0

Tempting reasoning: Since it is consistent with my beliefs that $Exp(A)$ is negative ($P(\text{Rain}) < 0.4$) I am permitted to choose $\sim A$. Since it is consistent with \mathcal{P} that $Exp(B)$ is negative ($P(\text{Rain}) < 0.6$) I am permitted to choose $\sim B$. So, given my imprecise beliefs, I am permitted to choose $\sim A$ & $\sim B$. *But, in doing this I make book against myself since A & B is a sure 5!*

The problem here is the decision theory, not the imprecise probabilities (as some claim)!

NEEDED: A THEORY OF PERMISSIBLE CHOICE FOR IMPRECISE PREFERENCES!

Terminology: A is *admissible* if A maximizes expected utility relative to some (\mathbf{P}, \mathbf{u}) pair.

So, relative to \mathcal{P} , all of A , $\sim A$, B and $\sim B$ are admissible.

But, Admissible \neq Permissible: If an act is admissible, it need *not* be rationally permitted for the agent to choose it. Why? Because admissibility is a categorical property, but permissibility is gradational in the realm of imprecise probabilities: acts are not permissible/impermissible per se, but only more or less permissible.

Some of what we need:

1. A is *determinately permissible* just in case it is determinately weakly preferred to any alternative, so that $Exp_{\mathbf{P},\mathbf{u}}(A) \geq Exp_{\mathbf{P},\mathbf{u}}(A^*)$ for A^* and *all* (\mathbf{P}, \mathbf{u}) pairs in her representing set. (Notice that I did *not* say “the act is not determinately dispreferred to any alternative”.)

2. Even though these questions lack determinate answers,

- a. Is it permissible for an agent to choose $\sim A$ over A ?
- b. Is permissible for her to choose $\sim B$ over B ?

it *is* determinately *impermissible* for her to choose both $\sim A$ and $\sim B$. Thus, any rule for going from preferences that makes $\sim A$ permissible should make $\sim B$ *impermissible*, and conversely. Likewise for Ellsberg. Some proposed rules fail, e.g., Levi's.

3. We have to give up on the idea that choices reveal probabilities. Good riddance!

History: Two Kinds of Decision Theory

Evidential Decision Theory (Jeffrey): Choose actions that provide you with *evidence* for thinking that desirable results will obtain (even when these acts do not causally promote those results).

$$\mathcal{V}(A) = P_A(E) \cdot u(A \ \& \ E) + P_A(\sim E) \cdot u(A \ \& \ \sim E)$$

- $P_A(E) = P(E \ \& \ A) / P(A)$ is the usual conditional probability of E given A . It is the degree of confidence you will invest in E if you come to *learn* that you will do A (and nothing else).
- $\mathcal{V}(A)$ is A 's “news value” or *auspiciousness*. \mathcal{V} -maximizers treat information about their acts as they would information about any other aspect of the world.

Causal Decision Theory: Choose actions that *causally promote* desirable outcomes (even if these acts provides you with evidence for undesirable outcomes they do not promote).

$$\mathcal{U}(A) = P^A(E) \cdot u(A \ \& \ E) + P^A(\sim E) \cdot u(A \ \& \ \sim E)$$

- $P^A(E)$ and $P^{\sim A}(E)$ are *causal probabilities* that capture your views about the causal powers of A and $\sim A$ vis-à-vis E . A is a promoting cause of E when $P^A(E) > P^{\sim A}(E)$. These are **not** ordinary conditional probabilities: $P^A \neq P_A$.
- $\mathcal{U}(A)$ is A 's “efficacy value”. \mathcal{U} -maximizers treat information about their acts as irrelevant (for purposes of decision making) except insofar as is it indicates what the acts are likely to cause.

Two Kinds of Dominance

	E	$\sim E$
A	Good	Worst
$\sim A$	Best	Bad

$\sim A$ dominates A

As we saw above, some sort of independence restriction on Dominance is required. In well-formulated decision problems states are independent of acts. But, independent in what sense?

Evidentialist Answer. If E is *evidentially independent* of A , so that learning A 's truth-value will not affect your estimate of E 's probability, then you should prefer A to $\sim A$ unconditionally.

(Note: Jeffery proposed this as a way of making decision theory sensitive to causal knowledge.)

Causalist Answer. If E is *causally independent* of A , so that you regard A 's truth or falsity as causally irrelevant to E , then you should prefer A to $\sim A$ unconditionally.

A FORMULATION OF CDT

$$u(A) = \sum_K P(K) \cdot u(A \& K)$$

The K form a partition of “dependency hypotheses” (assumed *causally* independent of the agent’s acts). Each K provides is a maximally complete specification of how the things the agent cares about might depend on what she does.

- Think of K as a bundle of counterfactual conditionals “If I were to do A_1 then c_1 would occur, and if I were to do A_2 then c_2 would occur, ...,” where the various A_i list the agent’s possible acts and the various c_j are potential outcomes. (Note, these counterfactuals need to be interpreted in a way that allows them to reflect causal connections, no “backtracking”.)
- Note: Other formulations allow one to replace the dependency hypotheses by “causal conditional probabilities” that directly give the probabilities with which probabilities will cause outcomes.

$$u(A) = \sum_S P(A \text{ []} \rightarrow S) \cdot u(A \& S) \quad (\text{Gibbard/Harper})$$

$$u(A) = \sum_S P^{\text{image-}A}(S) \cdot u(A \& S) \quad (\text{Sobel/Joyce})$$

NEWCOMB PROBLEMS

EDT and CDT conflict in *Newcomb problems*, which involve acts that are both highly auspicious but causally inefficacious. Such acts have a high news value, $\mathcal{V}(A)$ – they indicate that desirable results are likely to occur – but they do not promote desirable outcomes and so have a low value for $\mathcal{U}(A)$.

EDT: “Make good news!”

CDT: “Cause good results, no matter how bad the news might be!”

Flagship Newcomb: Predictions of a highly reliable predictor, which were made yesterday.

	<i>Predict One</i>	<i>Predict Two</i>
Choose One	\$1,000,000	\$0
Choose Two	\$1,001,000	\$1,000

Evidentialists: “Choose One” because it indicates that you will be rich (even though it costs \$1,000).

Causalist: “Choose Two” because it earns you \$1,000, and only indicates that you will become poor (without doing anything to cause it).

RATIFICATIONISM (Jeffrey 1983, Eells1982)

Maxim of Ratifiability: Choose for the person you expect to be once you have chosen.

- a. For each act A , evaluate the expected utility of each alternative act on the assumption dA that you will ultimately decide to perform A . A is *ratifiable* iff it maximizes expected utility on the assumption that it will be decided upon: $utility(A/dA) \geq utility(B/dA)$ for all acts B .

Note: I follow Jeffrey is assuming that A and dA are logically distinct propositions, so that dA does not entail A , or conversely.

- b. Choose only ratifiable acts. If A is unratifiable, then it is ruled out as a rational choice.

Comment: Unratifiable acts do seem defective. If you cannot choose to perform A without thereby giving yourself a compelling reason not to perform A , then you should not choose A .

One can endorse ratificationism from either a causalist or evidentialist perspective.

A is *e*-ratifiable iff $\mathcal{V}(A/dA) \geq \mathcal{V}(B/dA)$ for all acts B . (Jeffrey)

A is *c*-ratifiable iff $\mathcal{U}(A/dA) \geq \mathcal{U}(B/dA)$ for all acts B . (Harper)

AN EVIDENTIALIST SOLUTION TO NEWCOMB?

Jeffrey, Eells: Auspicious but inefficacious acts in Newcomb problems are not *e*-ratifiable, i.e., they do not maximize news value on the supposition that they are decided upon.

Basic Idea: A rational agent's *ability to anticipate her own decisions* nullifies any purely evidential correlations that might exist between states and acts.

In Newcomb: One-boxing is not *e*-ratifiable because it indicates that the million dollars is in the box, and if you know the million is in the box you'd rather two-box. Two-boxing is *e*-ratifiable.

According to Jeffrey and Eells, this sort of ratificationist reasoning provides a general, and *purely evidentialist* rationale for choosing efficaciously in Newcomb problems.

Not so! (But a story for another day.)

TWO FORMS OF RATIFICATIONISM

As an *elimination rule*, ratificationism requires you to first reject all unratifiable acts, and to then choose among the ratifiable alternatives.

As an *equilibrium rule*, ratificationism requires you to choose an act that is ratifiable relative to the beliefs and desires you will have when your deliberations cease (“reflective equilibrium”).

These differ because your beliefs about your own acts and about what your acts might cause can change as a result of rational deliberation.

ANDY EGAN’S “COUNTEREXAMPLES” TO CDT

In a recent paper (*Phil Review*, 2007), Andy Egan provides examples in which an act provides evidence about its own causal consequences. Egan believes that these examples pose challenges for causal decision theory. (Nearly identical examples were proposed in the early 1980s)

	<i>Lesion. You would miss if you were to shoot.</i>	<i>No Lesion: You would hit if you were to shoot.</i>
Shoot	Real Bad ($u = -10$)	Real Good ($u = 10$)
Don't Shoot	Status Quo ($u = 0$)	Status Quo ($u = 0$)

Story: Things would be better if you killed Alfred. You have a gun aimed at his head and just need to pull the trigger. Unfortunately, you know you are a random member of a population in which 1 in 5 have a brain lesion that causes poor aim. If you have the lesion and shoot, you will miss. If you lack the lesion and shoot, you will hit your target. The lesion also causes homicidal tendencies in 0.75 of those who have it, whereas only 0.01 of people without the lesion have such tendencies. With these numbers, being inclined to shoot is good evidence for thinking that you have the lesion $P_0(L/S) = 0.95$, which is good evidence for thinking that you would miss if you were to shoot. In contrast, being inclined to refrain is good evidence for thinking that you do not have the lesion $P_0(L/\sim S) = 0.06$, which is good evidence for thinking that you would not miss if you were to shoot.

Should you shoot?

Your evidential situation when deliberation begins is this:

$$P_0(L) = 0.2 \quad P_0(S / L) = 0.75 \quad P_0(S / \sim L) = 0.01$$

$$\text{So, } P_0(S) \approx 0.158 \quad P_0(L / S) \approx 0.95 \quad P_0(L / \sim S) \approx 0.06$$

- Your expected utilities when deliberation begins:

$$u_0(S) = 10 P_0(\sim L) - 10 P_0(L) = 6 > 0 = u_0(\sim S)$$

$$v_0(S) = 10 P_0(\sim L / S) - 10 P_0(L / S) = -9 < 0 = v_0(\sim S)$$

- Neither pure act is causally (or evidentially) ratifiable:

$$u_0(S / dS) \approx v_0(S) = -9 < u_0(\sim S / dS) = 0$$

$$u_0(S / d\sim S) \approx 8.8 > u_0(\sim S / d\sim S) = 0$$

- The **mixed act** $C = [0.38 S, 0.62 \sim S]$ is causally ratifiable

$$u_0(S / dC) = u_0(\sim S / dC) = u_0(C / dC) = 0$$

Note. The probabilities in this mixed act are probabilities for your own actions. What could such probabilities be? Do they even make sense?

Egan's Claims

- a. Causal decision theory recommends shooting.
- b. It would be irrational for you to decide to shoot.
- c. It will not help the causal theorist to go ratificationist because, while this does rule out shooting, it also rules out refraining.
- d. Refraining is the unique rational choice.

According to Egan, (d) distinguishes Murder Lesion from Death in Damascus (see below) where there is no rational choice. In ML there *is* a rational choice, says Egan: You should not shoot! Moreover, this choice is not one that causal decision theory can recommend even when augmented by ratifiability.

My Claims:

- a*. Causal decision theory does not recommend shooting.
- b*. It would be irrational for you to decide to shoot, and it would be irrational for you to decide not to shoot. But, you might be able to shoot, or to refrain, quite rationally if you conclude your deliberations in the right frame of mind.
- c*. If ratificationism is properly understood as an equilibrium rule, then *every* reasonable decision theory must go ratificationist.
- d*. Refraining is *not* the unique rational choice, but the intuition that refraining has more going for it than shooting is a sound one, and it can be explained within CDT.

A WELL-KNOWN EXAMPLE

At first, Egan’s example struck me as a variant of the “Death in Damascus” case discussed in Allan Gibbard and Bill Harper’s famous paper on causal decision theory.

	<i>S_D = Death seeks you in Damascus</i>	<i>S_A = Death seeks you in Aleppo</i>
<i>D = Stay in Damascus</i>	Die (<i>u</i> = 0)	Live (<i>u</i> = 10)
<i>A = Flee to Aleppo</i>	Live (<i>u</i> = 10)	Die (<i>u</i> = 0)

The Grim Reaper is coming for you tomorrow either in Damascus or Aleppo. You are certain to be in one place or the other, but it is up to you to choose which. The Reaper has already made a prediction about which city you will select, and has booked a flight there. He cannot change his itinerary (non-refundable ticket). So, his location is causally independent of your choice. However, you know that Death is reliable predictor of your actions, and so your subjective probability for his being where you choose to be is very high: you now think that you are likely to die in Damascus if you stay in Damascus, and that you are likely to die in Aleppo if you flee to Aleppo. But, crucially, you also confident that if you do stay in Damascus then you would have lived had you fled to Aleppo, and that if you do flee to Aleppo then you would have lived had you stayed in Damascus. This leaves you in a pickle because neither choice is *causally ratifiable*.

- As is Egan’s example, learning what you decide is evidence about what your acts will *cause*!
- Standard Diagnosis (CDT + Ratifiability): Death in Damascus is a pathological decision in which there is *no* rational choice. Reflection on Egan’s examples will show us that this is the wrong conclusion to draw.

DOES CDT ADVOCATE SHOOTING?

Egan: CDT “enjoins us to *do whatever has the best expected outcome, holding fixed our initial views about the likely causal structure of the world*”. Egan thus sees CDT as committed to:

Initial Opinion Fixes Action. If P_0 characterizes your beliefs at the *start* of your deliberations, then you are rationally obliged to perform an act that maximizes

$$\mathcal{U}_0(A) = P_0(L) u(L \& A) + P_0(\sim L) u(\sim L \& A)$$

More generally,

Current Opinion Fixes Action. If $\{K\}$ is a partition of “dependency hypotheses”, and if P_t characterizes your beliefs about the K at time t , then at t you are is rationally obliged to perform an act that maximizes your time t causal expected utility

$$\mathcal{U}_t(A) = \sum_K P_t(K) u(K \& A)$$

If these principles are right, then CDT does unequivocally (and incorrectly) tell you to shoot because $\mathcal{U}_0(A) = P_0(L) \cdot 0 + P_0(\sim L) \cdot 10 = 8 > \mathcal{U}_0(A) = 3$.

However, these principles are *wrong!*

WHAT CAUSAL DECISION THEORY IS COMMITTED TO

Current Opinion Fixes Evaluation. At any time t , if P_t gives your beliefs at t , then you are rationally obliged to *evaluate* each act by its causal expected utility at t :

$$U_t(A) = P_t(L) u(L \& A) + P_t(\sim L) u(\sim L \& A)$$

- This says nothing about what you should *do*; it pertains only to how you should evaluate acts given your beliefs and desires at t .
- It is entirely consistent with this that evaluations of acts should not be acted upon until they meet some further condition.

Key Question: *When should time- t evaluations of expected utility guide actions, i.e., under what conditions should an agent perform the act that maximizes expected utility relative to her beliefs at that time?*

ARNTZENIUS'S SOLUTION: MODIFY CAUSAL DECISION THEORY

“Causal decision theory is incoherent in the following sense: there are situations such that, as soon as you have made up your mind to do something, that decision looks bad... situations in which there are no stable decision states if one adheres to causal decision theory. What to do?”

Answer: Replace ordinary causal decision theory by “**deliberational** causal decision theory” which “allows that the end result of a rational deliberation will be that one has nontrivial degrees of belief in one’s possible acts.”

Basic Ideas:

1 Decision theory should require agents to act only on those evaluations that issue from beliefs and desires that are in deliberational equilibrium. (I agree)

In Egan’s example the mixed state $C = [0.38 S, 0.62 \sim S]$ is the equilibrium.

2 When no “pure” acts are causally ratifiable the equilibrium is a state in which more than one action has a positive probability of being performed. (I agree)

3 In such cases, the only rational act is the “mixed act” in which the agent lets her action be determined by a chance device that selects each given act with the probability that it has in equilibrium. (I disagree)

4 This mixed act is always causally ratifiable.

SOME QUESTIONS TO KEEP IN MIND

- Does the mere fact that there is no stable pure act in Egan's example really refute CDT?

NO. Not every decision problem should have a pure act solution. This is especially true for those decisions in which information about acts have causal import. (Note, however, that Egan thinks that there should be a unique pure act solution in his example.)

- Is CDT really a new type of decision theory, i.e., does adding an equilibrium requirement CDT really change what CDT says?

NO. As we shall see below, the equilibrium requirement falls out naturally from CDT and a principle – USE ALL YOUR EVIDENCE – that is implicit in every decision theory.

- Is there any reason to think that acts chosen on the basis of evaluations made in deliberational equilibrium are somehow more rational than acts chosen on the basis of initial, non-equilibrium beliefs and desires?

YES. See below.

- Is it really correct that only the mixed-act is recommended in equilibrium?

UNCLEAR. See below.

ANSWER TO THE KEY QUESTION

- If an agent's beliefs at t do not reflect all the information available to her at t , and if some of this information is relevant to questions about what her acts will cause, then it is a mistake for her to use her time- t beliefs as a basis for maximizing expected utility.
 - Performing an act because it maximizes expected utility *relative to the beliefs that the agent has at t* is only rational if these beliefs incorporate all of the agent's relevant information concerning what her acts are likely to cause. (CDT has us ignore information about what our acts merely indicate, but it insists that we pay attention to what our acts cause.)
- So, you should *not* act on the basis of the beliefs you hold when you start deliberations at $t = 0$ because these beliefs do not incorporate the evidence that you prefer shooting at $t = 0$.

Since you will be better informed about the effects of your acts later on in your deliberations, you should leave the decision to your “future self” who will be in a better epistemic position with respect to the question of what you should do.

Moral: Even though CDT ranks shooting above refraining at $t = 0$, it does *not* advise you to *act* on the basis of this evaluation.

No additional “equilibrium requirement” is needed to attain this result!

A FORMAL MODEL OF DELIBERATION (SKYRMS 1990)

- An agent's mental state at time t is represented by a probability P_t and a (causal) expected utility U_t .
- Probabilities are assigned to acts as well as states.
- Deliberation maps an initial (P_0, U_0) through a sequence of temporal stages (P_t, U_t) , $t \leq 1$, to a final state (P_1, U_1) .
- At each stage t , each pure act A has a causal expected utility $U_t(A)$, and the desirability of the agent's overall situation is given by the utility of "status quo" $U_t(SQ) = \sum_A P_t(A) \cdot U_t(A)$.
- There is an update rule that takes information about the values of $P_t(A)$ and $U_t(A)$ for each A , and on that basis determines new act probabilities $P_{t+\varepsilon}(A)$ at $t + \varepsilon$.
- The details of the update rule matter very little provided that acts have their probabilities raised (lowered) at $t + \varepsilon$ if and only if their time- t utilities exceed (are exceeded by) the time- t utilities of the status quo, i.e., acts that look good at t become more probable at $t + \varepsilon$.
- At $t = 1$ the agent reaches a state of equilibrium. This counts as making up her mind. (Perhaps she also needs to recognize that she is in this state.)
- Deliberation usually ends with one act being assigned probability 1 at $t = 1$, but it can end in a "mixed state" in which $P_1(A) > 0$ for more than one act A , and $U_1(A) = U_1(B)$ for all such acts. Here the agent ends up being torn among *equally desirable* acts.

The details in Egan's example:

➤ $P_t(S / L) = 0.75$ $P_t(S / \sim L) = 0.01$ for all t .

➤ And, for any given value of $P_t(S) = p_t$, we have

$$P_t(L) = [p_t - 0.01] / 0.74$$

since $p_t = P_t(L) P_t(S / L) + P_t(\sim L) P_t(S / \sim L)$.

➤ Thus,

$$u_t(S) = P_t(L) \cdot -10 + P_t(\sim L) \cdot 10 = 10 \cdot (1 - 2 \cdot [p_t - 0.01] / 0.74)$$

$$u_t(\sim S) = 0$$

$$u_t(SQ) = p_t \cdot 10 \cdot (1 - 2 \cdot [p_t - 0.01] / 0.74)$$

Note that $u_t(SQ) = u_t(S) = u_t(\sim S)$ when $p_t = 0.38$.

REFLECTIVE EQUILIBRIUM

- An agent should use her time- t beliefs as a basis for maximizing expected utility only when her beliefs (and preferences) are in a state of *reflective equilibrium*.
- Why? Because this is the only state in which all the relevant evidence at her disposal has been taken into account.
 - In particular, information about what preferences the agent holds and what probabilities she assigns to her acts should be taken into account in equilibrium. This data is usually irrelevant, but it can matter in cases (like Egan's) in which the probability that the agent is going to behave in various ways alters her views about what her acts will cause.
- So, you should act to maximize your time- t expected utility only when learning your time- t probability for any state or your time- t utility for any act will not alter your evaluations of acts.

General Point: At bottom, decision theory is about the relationships that hold between an agent's beliefs and desires and her actions *when she has attained a state of deliberational equilibrium*.

In the first instance, decision theory evaluates states of mind, not actions!

TWO WAYS OF ASSESSING ACTS IN EQUILIBRIUM

First Way: The equilibrium expected utility $\mathcal{U}_1(A)$ reflects the agent's evaluation of A (as a cause of desirable outcomes) given her P_1 beliefs.

Second Way: The equilibrium probability $P_1(A)$ reflects the degree to which the prospect of performing A contributes to the value of the status quo. It measures the extent to which she is "leaning toward" A .

- There is no difference between these modes of assessment when only one act survives deliberation, so that $P_1(A) = 1$ for some A .
- When more than one act has positive probability it can happen that $P_1(A) > P_1(B)$ even though $\mathcal{U}_1(A) = \mathcal{U}_1(B)$.

In such a situation the agent (i) sees no advantage in performing A over B , yet (ii) is perfectly happy to be in a position where she is more likely to do A than to do B . Indeed, she would be less happy if A 's probability were any lower, or higher, than it is.

- When deliberation terminates in an equilibrium, the agent's epistemic state, with respect to her pure acts, will be exactly the same as it would be if she were to decide to perform a mixed act. But, she can choose any act with a positive probability without inviting irrationality. (To be explained further below.)

Illustration 1 – Death in Damascus:

- Here the agent winds up torn equally between staying in Damascus and fleeing to Aleppo in an equilibrium where

$$\begin{aligned}P_1(\text{Stay}) &= P_1(\text{Flee}) = 1/2 \\ \mathcal{U}_1(\text{Stay}) &= \mathcal{U}_1(\text{Flee}) = \mathcal{U}_1(SQ) = 5\end{aligned}$$

- Both her acts are equivalent relative to both modes of assessment.
- She would be less happy with her total situation if the probabilities of the acts were not the same.

If $P(\text{Stay}) = p$, then $\mathcal{U}(SQ) \approx 0.198 + 19.208(p - p^2)$, which assumes its maximum value of 5 when $p = 1/2$.

- So, the agent is rational iff she acts to maximize (causal) expected utility in a state of mind that ranks both options as equally likely and equally desirable.
- This is *all* decision theory can say in this case; to think it says more is to pretend that the agent has reasons she does not have.
- This is the right answer: as far as considerations of rationality go, the agent who faces Death in Damascus *cannot go wrong* because it does not matter which way she goes. (Contrast Gibbard and Harper, who regard this as a pathological situation in which no choice can be rational.)

Illustration 2 – Murder Lesion:

Since $P_t(S/L) = 0.75$ and $P_t(S/\sim L) = 0.01$ for all t , the equilibrium is such that

- You are much more strongly inclined toward refraining than toward shooting: $P_1(S) = 0.38 < P_1(\sim S) = 0.62$.

Note: this implies $P_1(L) = P_1(\sim L) = 0.5$.

- You estimate that either act will contribute equally well toward your happiness: $u_1(S) = u_1(\sim S) = u_1(SQ) = 0$.
- The equilibrium is the unique self-ratifying state.

EXPLAINING THE INTUITION THAT REFRAINING IS BETTER

- Causal decision theorists should say that, contra Egan, it is *not* true that refraining is rational while shooting is irrational.

If you decide to refrain outright, if you “lean” all the way in that direction, then you are making an irrational choice, just as you would if you leaned all the way toward shooting.

- Still, there is a crucial difference between the two acts since you will, if rational, end up leaning more strongly toward refraining than toward shooting once you have digested all her information.
- So, CDT does not recommend *either* shooting or refraining in ML. Rather, you should reason yourself into a position where you are leaning more strongly toward refraining than shooting.
- Once you are in this position you can perform *either* action without risking irrationality.
- Again, we should resist the temptation to think that decision theory can or should deliver more than this. Your reasons are sufficient to incline you more strongly toward refraining than toward shooting, but not strong enough to make it reasonable for you to refrain outright.
- If all this is right, then causal decision theory has nothing to fear from Egan’s examples.

A FLY IN THE OINTMENT?

I have argued that someone who reasons her way to the (0.38, 0.62) equilibrium can rationally shoot or not shoot. Suppose she picks not shooting. Why is it *now* legitimate for her to ignore the information $\sim S$ conveys about her prospects of success when this would have been illegitimate at any earlier point in her deliberations?

Some relevant considerations:

- The agent is “picking” here. Nothing in her choice reveals anything about her reasons for action, which are perfectly in balance since $u_1(S) = u_1(\sim S) = u_1(SQ) = 0$.
- The agent has to believe, in equilibrium, that she will pick S with probability 0.38 and $\sim S$ with probability 0.68. A Puzzle: If S and $\sim S$ are equally good, how is it that she should remain in a state where she believes that she will pick the two acts with unequal probability. This calls for discussion.
- If the causal consequences of the agent’s choice are tied not just to her deliberative decision making but also to her *a rational* mechanism for picking, *whatever it is*, then the decision problem is genuinely pathological.
- Otherwise, the agent is in a position to disregard the evidence that her performing the act will provide about its own success.
- The agent may well regret her action, but from her current perspective this regret is irrelevant.

WEAK PREFERENCE REFLECTION

WPR (Arntzenius): One should not desire to do something that one can foresee one will regret, at least not when one knows that one's future desires will differ from one's current desires *only* to the extent that one has acquired relevant evidence in the interim.

But, in ML and DinD you might regret whatever you do, since coming to know what you do may make the other act look better, i.e., $u_1(S) = u_1(\sim S)$, but $u_1(S / S) < u_1(S / \sim S)$ and $u_1(\sim S / S) > u_1(\sim S / \sim S)$.

Is this a problem? No!

Preference Reflection should only hold when one currently believes that one's future self will be better situated with respect to the relevant evidence. Usually, the potential acquisition of new data is taken to improve one's epistemic situation, but not here.

Why? Usually, when a person imagines her future self having additional information E , she believes that her future self's utility assessments will coincide to her own conditional on E .

But, in this case $E = \text{"I pick } \sim S\text{"}$ And, when the agent conditions on this she bumps herself out of her deliberational equilibrium since she now needs to consider whether she *should* pick S , and the evidence $u_1(\sim S / S) > u_1(\sim S / \sim S)$ again needs to be taken into account.

This will lead to a whole new deliberative spiral, the result of which will be her current epistemic state, not the state that her future self will be in if she learns E ! (Some puzzles here.)

GENERAL CONCLUSIONS

- Decision theory (causal or not) must be thought of as a normative theory that concerns the decision maker's state of mind at the point when she makes her choice.
- Acts can be rationally performed just when they maximize expected utility relative to the beliefs and desires that the decision maker will have when her deliberations are complete and she has taken all the information at her disposal into account.
- Sometimes this equilibrium state will pick out a specific action as the unique right choice, but in cases like Egan's it will not.
- Here we need to distinguish two ways of assessing values of actions, and when we do the paradoxical character of the examples disappear.
- When no act is causally ratifiable, a rational agent can rationally perform any act that has a positive probability in equilibrium, but she might end up regretting whatever she does.

References

- Arntzenius, Frank [2008] "No Regrets, or: Edith Piaf Revamps Decision Theory," *Erkenntnis* **68**: 277-297.
- Bradley, Richard [2000]. "Conditionals and the Logic of Decision," *Philosophy of Science (Proceedings)* 67.
- Eells, Ellery [1982] *Rational Decision and Causality*. Cambridge, MA: Cambridge University Press.
- Egan, Andy [2007] "Some Counterexamples to Causal Decision Theory", *Philosophical Review*
- Gibbard, Allan and William Harper [1978] "Counterfactuals and Two Kinds of Expected Utility," in *Foundations and Applications of Decision Theory*, edited by C. Hooker, J. Leach, and E. McClennen, pp. 125-62. Dordrecht: Reidel.
- Jeffrey, Richard [1983] *The Logic of Decision*, 2nd edition, Chicago: The University of Chicago Press.
- Joyce, James M. [1999] *The Foundations of Causal Decision Theory*. Cambridge, UK: Cambridge University Press.
- Rothschild, Michael and Stiglitz, Joseph [1970]. "Increasing Risk: I. A Definition," *Journal of Economic Theory* **2** (1970), pp. 225-243.
- Savage, L. J. [1972] *The Foundations of Statistics* (2nd ed.). New York: Dover publications.
- Skyrms, Brian [1990] *The Dynamics of Rational Deliberation*. Cambridge, UK: Cambridge University Press.
- Zynda, Lyle [2000] "Representation Theorems and Realism About Degrees of Belief," *Philosophy of Science* **67**: pp. 45-69.