CAUSATION IN DECISION THEORY

Causality Study Fortnight

University of Kent, Canterbury, UK

September 9, 2008

Jim Joyce
Department of Philosophy
The University of Michigan
jjoyce@umich.edu

RATIONAL CHOICE THEORY

Rational choice theory seeks to provides a formal account of practical rationality that can (a) help agents with incomplete information 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).

> Central Aim: to identify the conditions under which a decision maker's beliefs and desires *rationalize* the choice of an action.

Main Topic Today: The debate between "causal" and "evidential" decision theorists

This debate concerns, not so much what agents should do, but whether we must make explicit reference to their beliefs about what their acts are likely to cause when characterizing their *reasons* for doing what they do.

- Causalists claim that we cannot properly capture a rational agent's rationale for acting without discussing either her beliefs about propositions with explicitly causal content or forms of belief revision that subject to explicit causal constraints.

- Evidentialists claim that, for the purposes of decision theory, we can capture all relevant beliefs about causes and effects by appealing to nothing more than the agent's ordinary subjective probabilities for non-causal propositions.

SAVAGE'S MODEL

In the influential model of Savage (1954), decision makers choose among *acts* that have 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.

	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:

- States are maximally specific descriptions of things in the world that the agent can**not** control.
- Acts are maximally specific descriptions of things in the world the agent **can** directly control.
- Each consequence is sufficiently detailed to settle *every* matter about which the agent intrinsically cares (including future contingencies).
- I will follow Jeffrey in thinking of consequences as act/state conjunctions, so that $C_{A,S} = A \,\&\, S$.

RATIONALITY AS SUBJECTIVE EXPECTED UTILITY MAXIMIZATION

The ‘standard model’: A rational agent’s preferences should conform 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* $\boldsymbol{u}$ that assigns a real number $\boldsymbol{u}(c)$ to each consequence c. $\boldsymbol{u}(c)$ measures the degree to which c would satisfy the agent’s desires and promote her aims.

- The strengths of the agent’s beliefs about states of the world can be characterized by a *subjective probability function* $\boldsymbol{P}$ whose values express her subjective degrees of confidence, or *credences*.

- The agent’s (instrumental) desires for acts can be characterized by their expected utilities computed using $\boldsymbol{P}$ and $\boldsymbol{u}$. An act A’s (unconditional) expected utility is a probability weighted average of the utilities of its consequences.

$$\boldsymbol{Exp}(A) = \Sigma_S\, \boldsymbol{P}(S)\cdot\boldsymbol{u}(A \mathbin{\&} S)$$

SEUM. Act A is rational only if it maximizes the agent’s (unconditional) *subjective expected utility*, so that $\boldsymbol{Exp}(A) \geq \boldsymbol{Exp}(B)$ for all alternative acts B.

This is *not* proposed as a decision procedure, but as a way of assessing the results of such procedures!

TROUBLES WITH DOMINANCE

Dominance. If a rational agent sees act *A* as producing a better outcome than act *B* in *every* state of the world, then the agent should choose *A* over *B* when these are her only choices.

Even though Dominance seems entirely obvious, and *SEUM* entails it, the principle cannot be employed unless relevant causal information is taken into account. For example,

The Bee Sting

	Anaphylactic shock	*No Anaphylactic shock*
Injection	mild pain & death	mild pain & survival
No Injection	no pain & death	no pain & survival

Just One More

	No Car Accident	*Accident*
Another Beer	no damage & beer	Damage & beer
Stop Now	no damage & beer	Damage & no beer

Moral: **Dominance does not apply when actions can causally influence states of the world on which their consequences depend!**

Note: Everybody agrees about this in these cases.

SAVAGE'S SOLUTION

We should (i) preserve the (unconditional) expected utility law, but (ii) only apply it to *well-formed* decision problems in which the agent judges states to be *independent* of her acts.

Savage's Claim. One can always find well-formed decisions by rewriting states as conditionals that capture dependency relationships between acts and outcomes.

	wreck if beer, wreck if no beer	wreck if beer, no wreck if no beer	no wreck if beer, wreck if no beer	no wreck if beer, no wreck if no beer
beer	damage & beer	damage & beer	no damage & beer	no damage & beer
no beer	damage & no beer	no damage & no beer	damage & no beer	no damage & no beer

Dominance no longer applies. Since the probability of the second state is much higher than that of the third, the expected utility forging the extra beer exceeds that of having another.

Big Question. How are we to understand the "if"'s that figure into Savage's states?

Two Answers:

(1) They are *indicative* conditionals, which signal evidential relationships.

(2) They are *counterfactuals*, which signal causal relationships.

JEFFREY'S SOLUTION

We should (i) jettison the (unconditional) expected utility law and replace it by a requirement to maximize *conditional* expected utilities of acts, which are obtained by weighting the utilities of each outcome by the probability of the state that brings that outcome about, where this probability is c*onditioned on the act in question being performed*.

$$Exp(A) = \sum_S P(S \text{ given } A) \cdot u(A \,\&\, S)$$

This has the benefit that (ii) the utility law can be applied to decisions involving any state partition, but (ii) dominance reasoning only applies when an agent she regards states as *independent* of the choice between acts, so that $P(S$ given $A) = P(S$ given $B)$ for all acts A over B.

Big Question. How are we to understand these conditional probabilities?

Two Answers:

(1) They are **ordinary conditional probabilities**, $P(S$ given $A) = P(S / A)$. These capture the agent's views about evidential relationships, e.g., she sees A as confirming (disconfirming) s more strongly that b does when $P(S / A)$ exceeds (is exceeded by) $P(S / b)$.

(2) They are **causal probabilities,** $P(S$ given $b) = P(S \setminus b)$. These capture the agent's views about the causal powers of her acts, e.g., she sees A as causally promoting (or inhibiting) S more (less) strongly than B does when $P(S \setminus A)$ exceeds (is exceeded by) $P(S \setminus b)$.

EVIDENTIAL DECISION THEORY

If we follow Jeffrey and let $P(S \text{ given } A) = P(S / A)$, then we ask agents to choose acts that provide them with *evidence* for thinking that desirable outcomes will ensue (even when these acts do not causally promote those outcomes). The expected utility law then has the form

$$\textbf{(EDT)} \qquad \mathcal{V}(A) = \Sigma_s P(s / A) \cdot u(C_{A,s})$$

- $\mathcal{V}(A)$ 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.
- EDT has the advantage of using ordinary conditional probabilities, which are well understood and which do not involve explicitly modal notions, e.g., counterfactual conditionals, objective chances, closeness relations among possible worlds, and so on.
- It is easy to show that news value is **partition invariant** in the sense that for any partition $\boldsymbol{E}$ of propositions formed by taking disjunctions of the states, one has

$$\mathcal{V}(A) = \Sigma_E P(E / A) \cdot \mathcal{V}(A \;\&\; E)$$

where $\mathcal{V}(A \;\&\; E) = \Sigma_S P(S / A \;\&\; E) \cdot \mathcal{V}(A \;\&\; S)$. (A similar "coarsening" property is true for acts.)

Key Point: Jeffrey hoped that all the information about causal relations needed for decision making could be captured in the conditional probabilities that figure into $\mathcal{V}$'s values.

CAUSAL DECISION THEORY

If we let $P(S \text{ given } A) = P(S \setminus A)$, then we ask agents to choose acts that *causally promote* desirable outcomes (even if they are evidence for undesirable outcomes they do not cause).

$$\textbf{(CDT)} \qquad \mathcal{U}(A) = \Sigma_S\, P(S \setminus A) \cdot u(A \mathbin{\&} S)$$

- $\mathcal{U}(A)$ 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*.
- $\mathcal{U}(A)$ and $\mathcal{V}(A)$ can coincide, because acts typically indicate outcomes they are likely to cause. But, they can also diverge when acts indicate outcomes they do not cause. This happens in "common cause" cases where one effect indicates another, via a "backtracking" argument.
- In most formulations, CDT is not partition invariant. For example, in the Skyrms-Lewis formulation

$$\mathcal{U}(A) = \sum_K P(K) \cdot u(A \mathbin{\&} K)$$

one must employ a special partition of "dependency hypotheses" that each provide a maximally complete specification of how things the agent cares about might depend on what she does.

Key Point: Causal decision theorists believe that causal information, either encoded in the partition or directly in the conditional probabilities $P(S \setminus A)$, must be explicitly included in the decision theory if we are to capture agents' rationales for what they do!

NEWCOMB PROBLEMS

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

EDT: “Make good news, even at the expense of causing bad results!”
CDT: “Cause good results, no matter how bad the news might be!”

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

	Predict One	*Predict Two*
Choose One	$1,000,000	$0
Choose Two	$1,001,000	$1,000

- Evidentialists: “Choose one,” it indicates that you will become rich (but this news costs $1,000).
- Causalists: “Choose two,” it earns you $1,000 (dominance), and only indicates that you will become poor.

PD with Twin: You and your twin are likely to choose the same way.

	Cooperate	*Defect*
Cooperate	$1,000,000	$0
Defect	$1,001,000	$1,000

- Evidentialists say “cooperate.” It indicates that your twin will cooperate (but costs you $1,000).
- Causalists say “defect” since you have no influence over his actions (dominance).

OBJECTION: “WHY AIN’T YA RICH”?

If PDT were played many times in a large population in which cooperators tended to be matched up with other cooperators and defectors tended to be matched up with other defectors, cooperators would do better, on average, than defectors. Hence, cooperation is rationally required because, *when compared with defectors, cooperators tend to do better*!

This seductive reasoning conceals three fallacies

- Gibbard and Harper (1978): The idea that the rational choice is the one with the highest overall utility is dubious in decisions set up to reward irrationality in a way that does *not* causally depend on what the agent actually does.
- Arntzenius (2007): There are cases in which CDTers get rich and EDTs end up poor.
- Joyce (1999/2008): The comparison between what cooperators get and what defectors get is either irrelevant or misleading. The objection is based on the assumption:

  ($) Defectors would likely do better if they were cooperators, but cooperators would likely do worse if they were defectors.

Unfortunately, ($) has two readings: one true but irrelevant, the other relevant but untrue.

A. ($) is true but irrelevant when read as a statement about player's "types" (= features of decision makers that explain the correlations between their acts and states, but which they cannot alter by their actions).

($$_{Type}$) Defector-types would likely do better if they were cooperator-types, but cooperative-types would likely do worse if they were defector-types.

BUT: Defector-types do worse not because they make worse choices, but because, *whatever they choose*, they tend to have worse options. Cooperative-types do not do better because they make better choices, but because, *whatever they choose*, they tend to have better options.

B. ($) is false but irrelevant when read as a statement about player's acts

($$_{Act}$) Whatever their type, those who actually defect would likely have done better had they cooperated, but actual cooperators would likely have done worse had they defected.

But, this is false. Defectors can expect worse options, but they do a better job with the options they actually get. This is what matter for assessments of decision-theoretic rationality!

Moral: When we compare apples to oranges (asking how cooperation for a cooperator stacks up against defection for a defector) it can seem as if cooperators make out. But, when we compare apples to apples (asking how cooperation compares to defection for a cooperator) and oranges to oranges (asking how defection compares for cooperation for a defector), it is clear that the defectors do better. Defection is the only rational choice in PDT.

RATIFICATIONISM (Jeffrey 1983, Eells1982)

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

a. An act A is ***ratifiable*** iff it maximizes expected utility on the assumption that it will be decided upon: $Exp(A/dA) \geq Exp(B/dA)$ for all acts B. Note: A and dA are logically distinct propositions.

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 A. (Harper)

AN EVIDENTIALIST SOLUTION TO NEWCOMB?

Jeffrery, Eells: Auspicious but inefficacious acts in Newcomb problems are not *e*-ratifiable; they fail to 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.
>
> **Screening Assumption**. Conditioning on the decision to perform any act *A* screens off any evidence that *A* might provide about states of the world, so that $P(S/A \;\&\; dA) = P(S/B \;\&\; dA) = P(S/dA)$, for all acts *B* and states *S*.
>
> > E.g., deciding to cooperate in PDT leaves you thinking that you twin is no more likely to defect if you actually do cooperate than if you actually do defect (having chosen to cooperate).

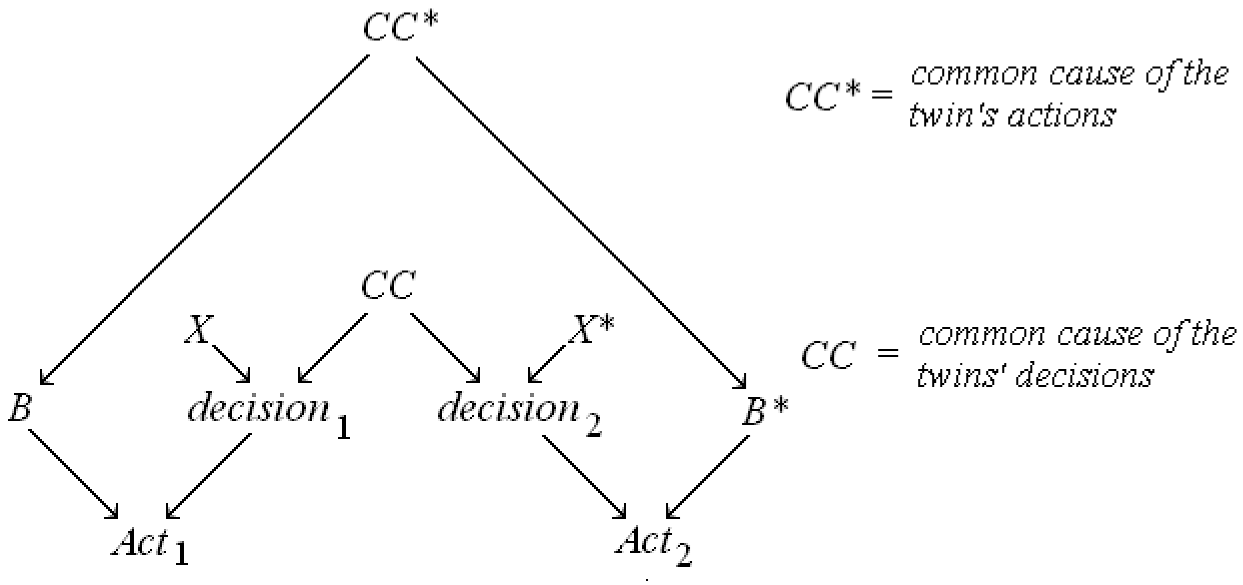
In Newcomb: One-boxing is not *e*-ratifiable because it indicates that the million dollars is in the box, and if you know the million is in the box you'd rather two-box. Two-boxing is *e*-ratifiable.

In PDT: Cooperating is not *e*-ratifiable because it indicates that your twin will cooperate, and if you know your twin will cooperate you'd rather defect. Two-boxing is *e*-ratifiable.

According to Jeffrey and Eells, ratificationist reasoning provides a *purely evidentialist* rationale for choosing efficaciously in Newcomb problems. Not so!

WORRIES ABOUT SCREENING

van Fraassen pointed out that Screening is implausible in some Newcomb problems because the act itself can be a better indicator of the state of the world than the mere decision, (Joyce 1999, p. 159).


CC*
CC* = common cause of the twin's actions
CC
CC = common cause of the twins' decisions
X
X*
B
decision 1
decision 2
B*
Act 1
Act 2

This sort of argument led Jeffrey to give up on ratificationism as a solution to Newcomb's problem

Eells (2000, p. 896) argues that van Fraassen's example should be "disqualified" because "rational deliberation loses its point" when there is a correlation among acts that cannot be screened off by decisions. Each player then believes that her action depends on factors that causally influence the other player's action without affecting his decision. Eells is right!

A LACUNA IN THE RATIFICATIONIST ARGUMENT

The auspicious *A* is unratifiable, and so cannot be rationally chosen.

The efficacious ~*A*, the only remaining option, is ratifiable.

[Hidden Premise: An act is choiceworthy iff it is ratifiable and it maximizes news value among all ratifiable acts, even if it does not maximize news value across *all* acts.]

Therefore, ~*A* should be chosen.

Is the evidentialist entitled to **HP**? Not so obviously!

- Though ~*A* is *e*-ratifiable, it is still lousy news. Though *A* is *e*-unratifiable, it is still good news.
- It could be that there is *no* rational choice when no *e*-ratifiable act maximizes news value.
- Evidentialists must explain how, in light of *A*'s unratifiability, the fact that ~*A* is *e*-ratifiable provides a reason for doing ~*A* that is sufficient to override or outweigh its lower news value. ~*A*'s *e*-ratifiablilty must somehow be portrayed as a reason *in favor* of performing it.

A JUSTIFICATION OF RATIFICATIONISM?

You should choose for the person you expect to be once you have chosen because, in virtue of knowing your decision, your "post-decision self" will be better informed at that time, and, a person with more information is better positioned to make wise choices than someone with less information.

Preference Reflection (1st pass). You should choose what your post-decision self would prefer you to choose, whenever you can determine what those preferences would be.

> Major Problem (with all simple reflection principles): One's later self might be an idiot from one's current perspective.

Preference Reflection (2nd pass). You should choose what your post-decision self would prefer you to choose, whenever you can determine what those preferences would be, provided that you are certain that

(i) You and your post-decision self have identical interests: your utilities *for basic outcomes* are the same.

(ii) Your post-decision self evaluate acts on the basis of the same sort of subjective expected utility (evidential or causal), one that you endorse.

(iii) Your post-decision self has the all the information you have, but he or she has come to *know* the truth of some propositions about which you are ignorant, e.g., at the moment of choice her or she will know what you have decided to do.

A SHARPER FORMULATION OF REFLECTION

Compare Joyce (2007) and Arntzenius (2008)

Suppose that an agent will come to know which one of a set of mutually exclusive and collectively exhaustive propositions {E_1, E_2, .. E_N} is true, and that this learning experience (i) will not disturb her basic desires, (ii) will not alter her ability to make rational decisions, and (iii) will not destroy any of the relevant knowledge she already possesses, then

(**R**) If the agent's post-learning self would choose *A* over *B* no matter which E_j he or she learns, then the agent should prefer *A* to *B* before learning which E_j is true.

This is the principle at work in ratificationist justifications of 2-boxing in Flagship Newcomb or of cooperation in PDT.

In Flagship, the partition is {decide to 1-box, decide to 2-box} and

- 2-boxing > 1-boxing if one knows one has decided to 1-box
- 2-boxing > 1-boxing if one knows one has decided to 2-box

So, Preference Reflection mandates 2-boxing.

WHAT JUSTIFIES PREFERENCE REFLECTION?

Preference Reflection can be justified on the basis of the following "expert principle" (Gaifman, 1988):

> **K** Suppose you know that *S* is a person (i*) whose interests coincide with your own, (ii*) who evaluates acts the same way you do, and (iii*) whose epistemic state is identical to yours, except that her or she as learned the truth-values of some propositions about which you remain uncertain. Under these conditions, you should regard *S* as an ***evaluative expert*** in the sense that you should prefer act *A* to act *B* if you can deduce that *S* will prefer *A* to *B*.

When combined with (i)-(iii), **K** entails that you should treat your post-decision self as an evaluative expert. This is just Preference Reflection!

The justification for Preference Reflection, and hence for **HP**, thus rests squarely on the expert principle **K**.

IS THE EXPERT PRINCIPLE TRUE?

K seems correct in many cases. E.g., J. H. Sobel's rationale for two-boxing in Flagship.

> You have a friend who splits the money with you, and who gets to look into the boxes after you have made your choice (but without knowing what choice you make). Your friend will surely prefer that you take two boxes whatever she learns about the contents of the boxes. So, according to **K**, you should take both boxes.

But **K** fails when you can causally influence what the expert learns!

Example: Joshua can study for his exam tomorrow or can spend time watching a TV show. He strongly prefers passing the exam to seeing to show, but he would like to see the show.

A fallacious inference:

I'll learn next week whether or not I passed the test.
If I learn that I passed, then it would be better to have seen the show than to have studied.
If I learn that I failed, then it would be better to have seen the show than to have studied.

Therefore, by **K** and Preference Reflection, I should prefer seeing to show to studying.

So, we need some restriction on K and Preference Reflection to handle cases in which the agent can influence that the "expert" learns.

TWO EXPERT PRINCIPLES

Consistency requires that we treat **K** the same way we treat the sure-thing principle.

- Causalists must restrict **K**'s use to decision problems whose acts do not causally influence the truth-values of any proposition that the expert learns.
- Evidentialists must go further and bar **K**'s use in any decision problem whose acts provide evidence about truth-values of the propositions that the expert learns, *and this is so even though the acts might have no causal influence over the truth-values of the propositions the expert learns*.

Accordingly, we have two different forms of Preference Reflection (assume conditions (i)-(iii))

$\boldsymbol{PR_C}$ You should choose what your post-decision self would prefer you to choose, whenever you can determine what those preferences would be, provided that your acts do not *causally influence* the truth-values of the propositions that you post-decision self will learn.

$\boldsymbol{PR_E}$ You should choose what your post-decision self would prefer you to choose, whenever you can determine what those preferences would be, provided that your acts do *provide evidence about* the truth-values of the propositions you post-decision self will learn.

Compare these with the two forms of the dominance principle!

RATIFICATIONIST SOLUTIONS TO NEWCOMB PRESUPPOSE CDT!

- Proponents of CDT must endorse PR_C for the same reason they endorse a causal independence restriction on dominance reasoning.

- Proponents of EDT must endorse PR_E for the same reason they endorse an evidential independence restriction on dominance reasoning.

But,

- Since your post-decision self learns what you decide (and no more), and since acts do not cause decisions (but are caused by them), $\boldsymbol{PR_C}$ provides ratificationist justification for choosing the efficacious act in Newcomb problems.
- In contrast, since your actions DO provide evidence about your decisions (because they are caused by them), it follows we CANNOT invoke $\boldsymbol{PR_E}$ to provide ratificationist rationale for choosing efficacious acts in Newcomb problems.

MORAL: Surprisingly, it is the causal decision theorist, not the evidentialist, who is in a position to employ ratifiability reasoning to rationalize efficacious choices. Evidentialists who seek to do so are subtly begging the question by invoking premises and principles (**HP** and $\boldsymbol{PR_C}$) that can only be justified within the confines of a causal decision theory.

Jeffrey's (2nd) Solution to Newcomb's Problem

"Causality and the Logic of Decision," *Probabilsitic Thinking*, Chapter 4

Jeffrey: Newcomb problems are not *decisions* at all. Those who face them know too much about the correlations between their behavior and states of the world to see themselves as acting *freely*.

- ♦ In a genuine decision an agent must see her acts as *causes* of outcomes.
- ♦ This does *not* involve having any explicit causal beliefs. It requires having beliefs about one's own actions that evolve in a very specific way during deliberation.

> During the course of her deliberations about what to do, an agent's subjective probabilities and news values will evolve through a series of states
>
> $$(P_0, \mathcal{V}_0) \longrightarrow (P_t, \mathcal{V}_t) \longrightarrow (P_1, \mathcal{V}_1),$$

Arntzenius's Test: An agent regards A as a *promoting cause* of the outcome A & S only if

- *Positive Correlation*. $P_0(S \mid A) - P_0(S \mid {\sim}A) > 0$.
- *Rigidity*. $P_t(S \mid A)$ and $P_t(S \mid {\sim}A)$ remain fixed as t varies.
- *Variation*. $P_t(A)$ varies during deliberation until $t = 1$, at which time it will have settled at zero or at one.

Jeffrey's View: Each action in a *genuine* decision problem must pass Arntzenius's Test.

CONSEQUENCES OF ARNTZENIUS’S TEST

- *Positive Correlation* says that learning A leaves the agent with more evidence for S than she will have should she learn S.

- Rigidity requires that belief changes induced by deliberation always proceed by Jeffrey Conditioning on $\{A, \sim A\}$.

  This entails that $\mathcal{V}_t(A)$ and $\mathcal{V}_t(\sim A)$ remain *fixed* as t varies.

  Thus, for Jeffrey, (ideal) deliberation is *not* a process in which an agent’s revises her views about the desirabilities of acts.

  Rather, she revises her views about act probabilities using a belief revision rule that, at any time t, (a) raises $P_t(A)$ just in case $\mathcal{V}_t(A) > \mathcal{V}_t(\sim A)$, and (b) then conditions on $\{A, \sim A\}$.

- Variation entails that $\mathcal{V}_t(\top) = P_t(A)\cdot\mathcal{V}_0(A) + P_t(\sim A)\cdot\mathcal{V}_0(\sim A)$, the news value of the “status quo”, will vary with time until $t = 1$, at which point a decision will be made

- When the decision is made, either $P_1(A)$ or $P_1(\sim A)$ will be 1, and $\mathcal{V}_t(\top)$ will coincide with either $\mathcal{V}_0(A)$ or $\mathcal{V}_0(\sim A)$, respectively.

MORAL: In any *genuine* decision an agent who deliberates properly is sure to ultimately settle on the act that maximizes news value *at time-0*.

JEFFREY ON WHY NEWCOMB PROBLEMS AREN'T DECISIONS

♦ In Newcomb problems the agent starts out with enough evidence about correlations between acts and states to fix definite values for all four conditional probabilities

$$P_0(S \mid A),\ P_0(S \mid {\sim}A),\ P_0(A \mid S),\ P_0(A \mid {\sim}S)$$

♦ "As decision problems are normally understood, values are fixed once given," specifically the values of $P_t(S \mid A)$, $P_t(S \mid {\sim}A)$, $P_t(A \mid S)$, $P_t(A \mid {\sim}S)$ must remain at their $t = 0$ values as t varies.

Why? Answer (I think):

(1) $P_t(S \mid A)$ and $P_t(S \mid {\sim}A)$ must remain fixed for the problem even to be a decision.

(2) $P_t(A \mid S)$ and $P_t(A \mid {\sim}S)$ remain fixed because they are justified on the basis of exogenous evidence of act/ state correlations, and the agent acquires no new evidence relevant to these correlations during deliberation.

♦ If $P_t(S \mid A)$, $P_t(S \mid {\sim}A)$, $P_t(A \mid S)$, $P_t(A \mid {\sim}S)$ remain at their $t = 0$ values, then, the unconditional probabilities of A and ${\sim}A$ (and S and ${\sim}S$) must remain fixed as well because

$$P_t(A) = \frac{P_t(A \mid S)\cdot P_t(S \mid {\sim}A)}{P_t({\sim}A \mid S)\cdot P_t(S \mid A) + P_t(A \mid S)\cdot P_t(S \mid {\sim}A)} = P_0(A)$$

Thus, according to Jeffrey, in Newcomb problems the agent's beliefs are so constrained by evidence about act/state correlations that she cannot alter her opinions about *A* and ~*A* without changing $P_t(S \mid A)$ and $P_t(S \mid {\sim}A)$.

- This means that she must either violate Rigidity or Variation, and so will be unable to see her acts as genuine causes of outcomes.

- This makes them "bogus" decision problems.

- Related Point: On Jeffrey's model of deliberation, an agent ends up increasing her degree of confidence in *A* simply in virtue of the fact that she prefers *A* to ~*A*. Why isn't this just wishful thinking if she already has evidence that justifies a specific probability for *A*?

Note: Some decision theorists (Levi, 2000) maintain that an agent cannot see herself as free to perform *A* to if she has beliefs about *A*. Jeffrey (famously) does not think this, but he does hold that *A*'s probability must be sufficiently unconstrained by the evidence so that it can be modified in deliberation *by acts of the will*.

ARE PROBABILITIES "FIXED ONCE GIVEN"?

Jeffrey's argument rests on the claim that $P_t(A \mid S)$ and $P_t(A \mid {\sim}S)$ are "fixed once given" in any genuine decision problem. But many decision theorists (both evidential and causal) have suggested that free agents are in a position to legitimately ignore evidence about their own acts.

Whatever evidence an act might provide
On facts that precede the act,
Should never be used to help one decide
On whether to choose that same act. Judea Pearl, *Causality*, p. 109

"Evidential decision theory preaches that one should never ignore genuine statistical evidence... but actions – by their very definition –render such evidence irrelevant to the decision at hand, for actions change the probabilities that acts normally obey." Pearl

"From the agent's point of view contemplated actions are always considered to be *sui generis*, uncaused by external factors. As [Ramsey] puts it, "my present action is an ultimate and the only contingency."… This amounts to the view that free actions are treated as probabilistically independent of everything except their effects." Huw Price, "Agency and Probabilistic Causality"

The agent must say 'Although I do have some opinion about what I'm likely to do in this situation, I will, for purposes of deliberation suspend judgment about what course of action I will in fact pursue." Chris Hitchcock, "Causal Decision Theory and Decision Theoretic Causation"

AN INCOMPLETE RESPONSE

For the full story see Joyce (2007)

1. A deliberating agent *may* legitimately ignore evidence about the causes of her actions, e.g., the evidence that fixes values for $P_0(A \mid S)$ and $P_0(A \mid {\sim}S)$.

   Thesis of Epistemic Freedom. Agents are free to believe what they want about their own acts.

2. A deliberating agent *may not* legitimately ignore the evidence about the effects of her actions, e.g., the evidence that fixes values for $P_0(S \mid A)$ and $P_0(S \mid {\sim}A)$.

Given (1) and (2), and Jeffrey's model of deliberation, agents will be able to revise their probabilities for A and $\sim A$ (and for S and $\sim S$) while keeping $P_t(S \mid A)$ and $P_t(S \mid {\sim}A)$ fixed at their initial values.

Thus, Newcomb problems do count as genuine decisions if we grant (1) and (2), and the EDT still gets them wrong (without adding ratifiability).

This response will unconvincing without some rationale for (1).

> *Why is it legitimate for deliberating agents to ignore evidence about their own actions when it is not legitimate for them to ignore evidence about other matters?*

A Causalist Defense of Epistemic Freedom

With debts to David Velleman's "Epistemic Freedom"

> *Evidence Proportionism*. A subject's level of confidence in a proposition *X* should be proportioned to her "external" evidence *X* 's truth. Adding the "internal" evidence that she does or will believe or desire *X* should not alter confidence.
>
> $$P_t(X / E) = P_t(X / E \,\&\, P_{t^*}(X) = p \,\&\, Exp_{t^*}(X) = d)$$

I endorse *EP* except for propositions that the agent takes to be *self-fulfilling*, i.e., those whose truth will, in the subject's view, be *causally promoted* by the fact that she believes or desires their truth.

> Example: William James's "crevasse jumper" might know that people like him often do not clear the crevasse (even when they believe they will). Yet if he is convinced that "positive thinking" helps he should be (at least a little) more confident of clearing the crevasse if he believes he will than if he does not.

My Claim: A deliberating agent who takes herself to be entirely free will respect to *A* will be convinced that her belief that she will perform *A* coupled with her preference for *A* over ~*A* is a *complete and sufficient* cause of *A*. This puts her in a position to disregard any and all prior evidence against *A* once she comes to belief she will do *A*.

> Whatever the value of $P_t(A / S)$, one has $P_t(A / S \,\&\, p_1(A) = p) = p$ where $p_1(A) = p$ is the statement that the agent's probability for *A* at the end of her deliberations will be *p*.

HOW SHOULD WE UNDERSTAND THE CAUSAL PROBABILITIES P(S \ A)?

Three Models of Causal Reasoning

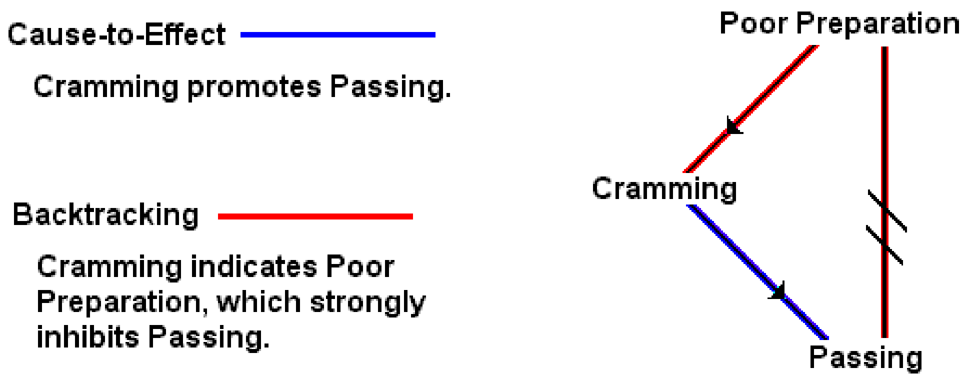
- *Probabilistic*: Believing that *C* causes *E* involves being convinced that *C*'s occurrence will raise *E*'s objective probability under conditions where all other "background" causal factors that cause *C* or *E* (but are not effects of *C*) are held fixed. Cause-to-effect reasoning involves settling on appropriate conditional probabilities for the *E* given the *C* in each maximal consistent set of such background factors

- *Interventionist*: Believing that *c* causes *e* involves believing that *e*'s properties are subject to "manipulation" by means of "interventions" that determine *c*'s occurrence. Cause-to-effect reasoning involves modifying one's beliefs to take the "setting" of *c*'s properties into account, and evaluating *e*'s probability on that basis.

- *Counterfactual*: Believing that *c* causes *e* involves believing that certain counterfactual dependency relations obtain. Causal-to-effect reasoning is a species of counterfactual reasoning in which counterfactual conditionals are interpreted in a "non-backtracking" way.

As I shall explain, for purposes of CDT these three ways of looking at things are equivalent.

BACKTRACKING (“BACK-DOOR”) INFERENCES



Cause-to-Effect

Cramming promotes Passing.

Backtracking

Cramming indicates Poor Preparation, which strongly inhibits Passing.

A cause *C*’s *total* evidential import for an effect *E* is the result of two factors:

> *Cause-to-effect Inference*. *C* is evidence for *E* in virtue of the fact that *C* promotes *E* more strongly than it inhibits *E*.
>
> *Backtracking Inference*. *C* is evidence for *E* in virtue of the fact that *C indicates* the presence of events or conditions that are causally or evidentially relevant to *E*.

The trick is to find ways to factor out the backtracking inferences.

PROBABILISTIC CAUSALITY (SKYRMS, LEWIS)

In the Skyrms-Lewis formulation of CDT

$$P(\mathrm{S} \setminus A) = \textstyle\sum_K P(K)\cdot P(S\,/\,A\ \&\ K)$$

where the *K* range over a special partition of "dependency hypotheses," each of which provides a maximally complete specification of how outcomes might depend on acts.

Issue: This leads to a formulation of CDT that is partition dependent.

This assumes that the *K* are probabilistically (and causally) independent of acts.

Issue: How should we formulate CDT when acts provide information about their own effectiveness at producing various outcomes?

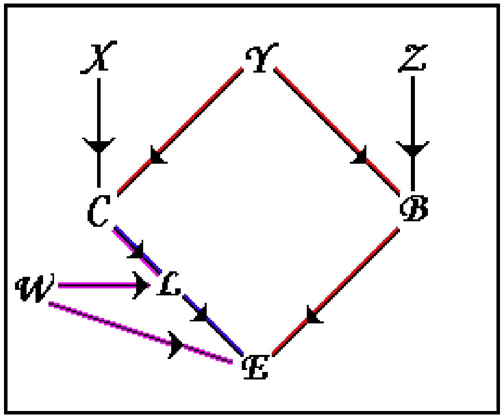
Example: Death in Damascus, Murder Lesion, Meta-Newcomb

Come to Thursday's talk!

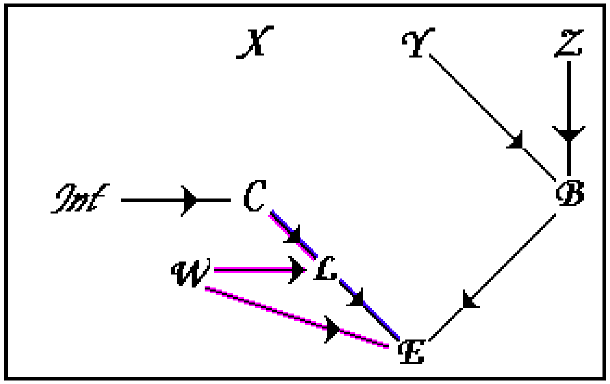
INTERVENTIONS

Interventions “set” the value of a “variable” in a way *that does nothing to disturb the patterns of causal dependency that hold between other variables in the model*.

Pre-intervention



Post-intervention



Int = {do(C) = C, do(C) = ¬C, do(C) = T}

Markov Condition: If *C* and *B* are correlated, $P(C) \neq P(C / B)$, and if no front-door path leads from *C* to *B*, then (i) *C* has causal parents, and (ii) conditioning on any set *V* of values for *C*’s parent variables screens *C* off from *B*, so that $P(C/ B \ \& \ V) = P(C/ V)$.

In a Markov graph, cause-to-effect inference is belief revision in light of an intervention.

PEARL'S "ADJUSTMENT FOR DIRECT CAUSES"

- In a Markov graph, the "causal effect of *C* on *E*" can be computed as

  $$P(E \mid do(C) = C) = \Sigma_V\, P(V)^{\cdot} P(E / C \,\&\, V)$$

  where *V* ranges over possible values of *C*'s *parent variables*.

- So, in the interventionist framework

  $$P(S \setminus A) = \Sigma_V\, P(V)^{\cdot} P(S / A \,\&\, V) = P(S \setminus \mathrm{do}(Act) = A)$$

  where *V* ranges over values of *A*'s immediate causes.

This is equivalent to the Skyrms/Lewis formulation assuming (a) each *V* is a disjunction of the K, (b) the Markov condition holds

$$\begin{aligned} P(S \setminus A) &= \Sigma_K\, P(K)^{\cdot} P(S / A \,\&\, K) \\ &= \Sigma_V\, \Sigma_{K \in V}\; P(K)^{\cdot} P(S / A \,\&\, K) \quad \text{by (a)} \\ &= \Sigma_V\, \Sigma_{K \in V}\; P(K)^{\cdot} P(S / A \,\&\, V) \quad \text{by (b)} \\ &= \Sigma_V\, P(S / A \,\&\, V)^{\cdot} [\Sigma_{K \in V}\; P(K)] \\ &= \Sigma_V\, P(S / A \,\&\, V)^{\cdot}\; P(V) \\ &= P(S \mid do(Act) = A) \end{aligned}$$

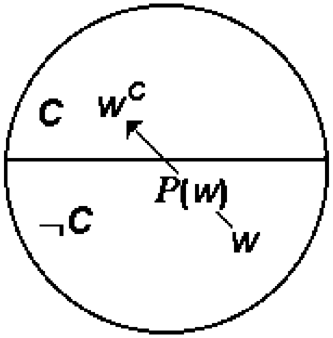
IMAGING

Counterfactualists about causation will measure the “causal effect” of A on S as $P^A(S)$ where P^A is P *imaged* on A in the sense of Lewis (1976)

$$P^A(S) = \Sigma_W\, P(W) \cdot P(S\,/\,W_A)$$

where W range over a set of possible worlds that has a “similarity relation” defined over it, and W_A is the world most similar to W in which A (it being understood that $W_A = W$ is A is true in W).

> *Limit Assumption*. For each world W and proposition C there is a world W_C that is most *like w* among the C-worlds.



All **of $P(w)$ is shifted to w^C.**

C-worlds retain their probability.

$$P^C(E) = P(C \;\square\!\!\rightarrow\; E)$$

Problem: There need be no “most similar” C-world to W, often there will be a *set* of “most similar” worlds $W[C]$.

GENERAL IMAGING

Gärdenfors (1982). The image of P on A can be written as

$$P^A(S) = \Sigma_{Z \in S} \Sigma_W \rho^A(W, Z)$$

where W and Z range over possible worlds and for each Z, $\rho^A(\bullet, Z)$ is some probability function that is uniformly zero for outside of $W[A]$. $\rho^A(W, Z)$ gives the proportion of the probability of W that is shifted onto Z. Note: $\rho^A(W, Z) = 1$ for $Z \in S$.

> Theorem (Gärdenfors). P^C is a general imaging function if and only if $P = \lambda P_1 + (1 - \lambda)P_2$ implies $P^C = \lambda P_1{}^C + (1 - \lambda)P_2{}^C$ whenever $0 \underline{<} \lambda \underline{<} 1$.
>
>> *Hidden Assumption*. P^C does not depend on P^C, i.e., the manner in which W's probability is spread over $W[C]$ does *not* depend on the prior distribution of probabilities in $W[C]$.

> P^C reflects the [believer's] judgments about similarity among possible worlds. These judgments will *not* depend on how likely she takes these worlds to be; the only place where her subjective probabilities enter the equation is through the $P(w)$ term.
>
> (Joyce, *FCDT*, p. 198)

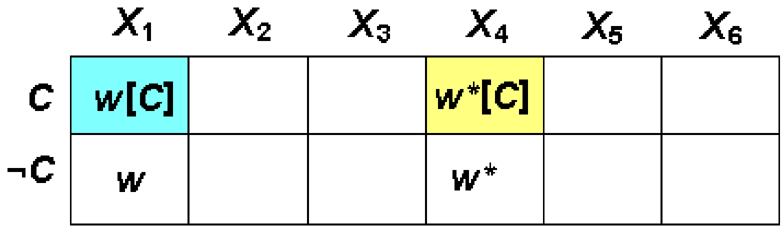
This is the *wrong* way to look at imaging!

BAYESIAN IMAGING AND STRATIFICATION

Bayesian Imaging shifts *W*’s probability to worlds in *W*[*C*] in direct proportion to their *prior* probability, so that

$$P^C(E) = \Sigma_{Z \in E} \Sigma_w P(W)\cdot P(Z / W[C]).$$

Definition. A similarity relation among worlds is *stratified* just in case, for any worlds in W_1 and W_2 and any proposition *C*, either $W_1[C]$ and $W_2[C]$ are disjoint or they are identical.



	X_1	X_2	X_3	X_4	X_5	X_6
C	*w*[C]			*w**[C]		
¬C	*w*			*w**		

The probability of each world in ¬*C* & X_i is shifted to worlds in the *C* & X_i box according to their probability.

When the stratifying partition is countable Bayesian imaging assumes the simple form

$$P^A(S) = \Sigma_{z \in S} \Sigma_X P(X)\cdot P(Z / A \& X)$$

If we chosen a similarity relation that it gets the causal relationships right, two worlds will count as equally simiar to *W* just when they agree about all the causal factors that are not ‘dowstream’ of the agent’s acts. In which case the *X*’s and *K*’s coincide. So, this formulation is equivalent to the others.

References

Arntzenius, Frank [2008] “No Regrets, or: Edith Piaf Revamps Decision Theory.*”* *Erkenntnis* **86[2]**: 277-97.

Bradley, Richard [2000]. “Conditionals and the Logic of Decision,’" *Philosophy of Science (Proceedings)* 67.
----- (1999) ”Conditional Desirability”, *Theory and Decision* 46.

Eells, Ellery [1982] *Rational Decision and Causality*. Cambridge, MA: Cambridge University Press.

Egan, Andy [2007] “Some Counterexamples to Causal Decision Theory”, *Philosophical Review* **116(1)**: 93-114.

Gaifman, Haim [1988] “A Theory of Higher Order Probability,” in *Causality, Chance and Choice*, edited by B. Skyrms and W. Harper, pp. 191-219. Dordrecht: Kluwer.

Gardenfors, Peter [1982] “Imaging and Conditionalization,” *Journal of Philosophy* 79: 747-60.

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.

Hitchcock, Christopher [1996] “Causal Decision Theory and Decision-Theoretic Causation,” *Noûs* **30**: 508 - 526.

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.

----- [2002] “Levi on Causal Decision Theory and the Possibility of Predicting One’s Own Acts,” *Philosophical Studies* **11**: 69-102.

----- [2007] “Are Newcomb Problems Really Decisions?” *Synthese***, 156[3**: 537-562.
----- [Unpublished] “Comment on Andy Egan’s ‘Some Counterexamples to Causal Decision Theory’,” FEW 2005.

Levi, Isaac [2000] “Review Essay: *The Foundations of Causal Decision Theory*,” *Journal of Philosophy* **97**: 387-402.
Lewis, David [1976] “Probabilities of Conditionals and Conditional Probabilities,” *Philosophical Review* **85**: 297-315.

----- [1979] “Counterfactuals Dependence and Time’s Arrow,” *Nous*, **13**: 455-76.

----- [2000] “Causation as Influence,” *Journal of Philosophy*, **97**: 182-97.

Pearl, Judea [2000] *Causality*. New York: Cambridge University Press, New York.

Savage, Leonard. [1954/1972]. *The Foundations of Statistics*, 2nd edition New York: Dover.

Skyrms, Brian [1990] *The Dynamics of Rational Deliberation*. Cambridge, UK: Cambridge University Press.

Is There a Role for Ratificationism in CDT?

Two forms of ratificationism:

> As an ***elimination rule***, ratificationism requires you to first rule out all unratifiable acts, thereby ceasing to regard them as genuine options, and to then choose among your 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 at the time your deliberations cease (in “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.

A formal model of deliberation (Skyrms 1990)

- An agent's mental state at time *t* is represented by a probability $prob_t$ and a (causal) expected utility $\mathcal{U}_t$.
- Deliberation maps an initial ($prob_0$, $\mathcal{U}_0$) through a sequence of temporal stages ($prob_t$, $\mathcal{U}_t$), $t \leq 1$, to a final state ($prob_1$, $\mathcal{U}_1$).
- At each stage, acts with higher expected utility become more probable, and this new information is feed back into the system and is used to recalculate expected utilities.
- At $t = 1$ the agent reaches a state of reflective equilibrium at which the process of deliberation stabilizes. This counts as the agent making up her mind. (+ recognition?)
- Deliberation usually ends with one act being assigned a probability of 1, but it can end up in a "mixed state" in which the agent remains undecided among equally desireable acts.

Andy Egan's Counterexamples to CDT

In a recent paper (*Phil Review*, this year), 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.

	Lesion. You would miss if you were to shoot.	*No Lesion: You would hit if you were to shoot.*
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 $prob_0(L/S) = 0.95$, 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 $prob_0(L/{\sim}S) = 0.06$, 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:

$prob_0(L) = 0.2 \qquad prob_0(S / L) = 0.75 \quad prob_0(S / {\sim}L) = 0.01$

So, $prob_0(S) \approx 0.158 \quad prob_0(L / S) \approx 0.95 \quad prob_0(L / {\sim}S) \approx 0.06$

- Your expected utilities when deliberation begins:

  $\mathcal{U}_0(S) = 10\ prob_0({\sim}L) - 10\ prob_0(L) = 6 > 0 = \mathcal{U}_0({\sim}S)$
  $\mathcal{V}_0(S) = 10\ prob_0({\sim}L / S) - 10\ prob_0(L / S\ ) = -9 < 0 = \mathcal{V}_0({\sim}S)$

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

  $\mathcal{U}_0(S / dS) \approx \mathcal{V}_0(S) = -9 < \mathcal{U}_0({\sim}S / dS) = 0$
  $\mathcal{U}_0(S / d{\sim}S) \approx 8.8 > \mathcal{U}_0({\sim}S / d{\sim}S) = 0$

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

  $\mathcal{U}_0(S / dC) = \mathcal{U}_0({\sim}S / dC) = \mathcal{U}_0(C / dC) = 0$

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

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.

	S_D = *Death seeks you in Damascus*	S_A = *Death seeks you in Aleppo*
D = Stay in Damascus	Die (*u* = 0)	Live (*u* = 10)
A = Flee to Aleppo	Live (*u* = 10)	Die (*u* = 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 $prob_0$ characterizes your beliefs at the *start* of your deliberations, then you are rationally obliged to perform an act that maximizes
>
> $$\mathcal{U}_0(A) = prob_0(L)\ u(L \,\&\, A) + prob_0({\sim}L)\ u({\sim}L \,\&\, A)$$

More generally,

> *Current Opinion Fixes Action*. If $\{K\}$ is a partition of "dependency hypotheses", and if $prob_t$ characterizes the agent's beliefs about the K at time t, then at t she is rationally obliged to perform an act that maximizes her time t causal expected utility
>
> $$\mathcal{U}_t(A) = \Sigma_K\ prob_t(K)\ u(K \,\&\, A)$$

If these principles are right, then CDT does unequivocally (and incorrectly) tell you to shoot because $\mathcal{U}_0(A) = prob_0(L)\cdot 0 + prob_0({\sim}L)\cdot 10 = 8 > \mathcal{U}_0(A) = 3$.

However, these principles are *wrong*!

What Causal Decision Theory is Committed To

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

$$\mathcal{U}_t(A) = prob_t(L)\ u(L \,\&\, A) + prob_t({\sim}L)\ u({\sim}L \,\&\, A)$$

- 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 (also in Joyce Unpublished): Decision theory should require agents to act only on those evaluations that issue from beliefs and desires that are in deliberational equilibrium.

In Egan's example the mixed state C = [0.38 S, 0.62 ~S] is the equilibrium.

2 (also in Joyce Unpublished): 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.

3 (Arntzenius is unclear about this): 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.

4 (a fact): This mixed act is always causally ratifiable.

Some Questions to Keep in Mind

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

  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 DCDT 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 version of 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.

- 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 $prob_t$ and a (causal) expected utility $\mathcal{U}_t$.
- Probabilities are assigned to acts as well as states.
- Deliberation maps an initial ($prob_0$, $\mathcal{U}_0$) through a sequence of temporal stages ($prob_t$, $\mathcal{U}_t$), $t \leq 1$, to a final state ($prob_1$, $\mathcal{U}_1$).
- At each stage *t*, each pure act *A* has a causal expected utility $\mathcal{U}_t(A)$, and the desirability of the agent's overall situation is given by the utility of "status quo" $\mathcal{U}_t(SQ) = \sum_A prob_t(A) \cdot \mathcal{U}_t(A)$.
- There is an update rule that takes information about the values of $prob_t(A)$ and $\mathcal{U}_t(A)$ for each *A*, and on that basis determines new act probabilities $prob_{t + \varepsilon}(A)$ 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 $prob_1(A) > 0$ for more than one act *A*, and $\mathcal{U}_1(A) = \mathcal{U}_1(B)$ for all such acts. Here the agent ends up being torn among *equally desirable* acts.

The details in Egan's example:

- $prob_t(S / L) = 0.75 \qquad prob_t(S / {\sim}L) = 0.01$ for all t.

- And, for any given value of $prob_t(S) = p_t$, we have

$$prob_t(L) = [p_t - 0.01] \,/\, 0.74$$

  since $p_t = prob_t(L)\, prob_t(S / L) + prob_t({\sim}L)\, prob_t(S / {\sim}L)$.

- Thus,

$$\mathcal{U}_t(S) = prob_t(L) \cdot -10 + prob_t({\sim}L) \cdot 10 = 10 \cdot (1 - 2 \cdot [p_t - 0.01] \,/\, 0.74)$$
$$\mathcal{U}_t({\sim}S) = 0$$
$$\mathcal{U}_t(SQ) = p_t \cdot 10 \cdot (1 - 2 \cdot [p_t - 0.01] \,/\, 0.74)$$

Note that $\mathcal{U}_t(SQ) = \mathcal{U}_t(S) = \mathcal{U}_t({\sim}S)$ when $p_t = 0.38$.

But, note too that

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 influences her views about what her actions are likely to 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 $prob_1$ beliefs.

Second Way: The equilibrium probability $prob_1(A)$ 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 $prob_1(A) = 1$ for some A.
- When more than one act has positive probability it can happen that $prob_1(A) > prob_1(B)$ 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.

- Deliberation terminates in an "equilibrium in belief" (in the sense of Aumann) that corresponds to a ratifiable (mixed) option for the decision problem. 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 the mixed act. But, if the mixed act is not available, then they can choose any act with a positive probability without inviting irrationality.

Illustration 1 – Death in Damascus:

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

  $prob_1(\text{Stay}) = prob_1(\text{Flee}) = 1/2$
  $\mathcal{U}_1(Stay) = \mathcal{U}_1(Flee) = \mathcal{U}_1(SQ) = 5$

- 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 two act were not the same.

  If $prob(\text{Stay}) = p$, 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 $prob_t(S/L) = 0.75$ and $prob_t(S/{\sim}L) = 0.01$ for all t, the equilibrium is such that

- You are much more strongly inclined toward refraining than toward shooting: $prob_1(S) = 0.38 < prob_1({\sim}S) = 0.62$.

  Note: this implies $prob_1(L) = prob_1({\sim}L) = 0.5$.

- You estimate that either act will contribute equally well toward your happiness: $\mathcal{U}_1(S) = \mathcal{U}_1({\sim}S) = \mathcal{U}_1(SQ) = 0$.

- 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. It recommends that you reason yourself into a position where you are leaning more strongly toward refraining than toward 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. You 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.

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.

References

Arntzenius, Frank [2008] "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis* **86[2]**: 277-97.

Bradley, Richard [2000]. "Conditionals and the Logic of Decision,'" *Philosophy of Science (Proceedings)* 67.
----- (1999) "Conditional Desirability", *Theory and Decision* 46.

Eells, Ellery [1982] *Rational Decision and Causality*. Cambridge, MA: Cambridge University Press.

Egan, Andy [2007] "Some Counterexamples to Causal Decision Theory", *Philosophical Review* **116(1)**: 93-114.

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.

----- [2007] "Are Newcomb Problems Really Decisions?" *Synthese***, 156[3**: 537-562.
----- [Unpublished] "Comment on Andy Egan's 'Some Counterexamples to Causal Decision Theory'," FEW 2005.

Savage, Leonard. [1954/1972]. *The Foundations of Statistics*, 2nd edition New York: Dover.

Skyrms, Brian [1990] *The Dynamics of Rational Deliberation*. Cambridge, UK: Cambridge University Press.

Two Potential Worries

In the relevant equilibrium in ML, one must have some method for going from the equilibrium state *C* = [0.38 *S*, 0.62 ~*S*] to one of the pure acts. Of course, one has no reason to prefer one to the other, since both have equal utility, so this method must be a method of “picking” between equally desirable alternatives.

Once you pick, isn’t your act unratitfiable, hence irrational to perfomr?

Yes/No. Ratifiability should apply only to equilibrium states, not to the acts that one picks in those states.

If *S* and ~*S* are equally good, how is that one should remain in a state where one believes that one will performs the two acts with unequal probability.

This calls for discussion.