Causal Decision Theory

Johan E. Gustafsson

Consider the following principle:

> **Dominance**
> If there is a partition of states of nature such that relative to it, an act x is at least as preferred as y in each state and x is preferred to y in at least one state, then x should be chosen rather than y.

	Pass	Fail
Study	3	1
Not study	4	2

Standard unconditional expected utility

$$\mathrm{VAL}(x) = \sum_{s \in S} P(s) * U(s \wedge x)$$

	Pass (prob 0.8)	Fail (prob 0.2)
Study	3	1
Not study	4	2

$$\mathrm{VAL}(\text{Study}) = 0.8 * 3 + 0.2 * 1 = 2.6$$

$$\mathrm{VAL}(\text{Not study}) = 0.8 * 4 + 0.2 * 2 = 3.6$$

$P(x|y)$ is the agent's subjective probability for x conditioned on the evidence that y.

$$P(x|y) = \frac{P(x \wedge y)}{P(y)}$$

	Pass	Fail
Study	3 (prob 0.8)	1 (prob 0.2)
Not study	4 (prob 0.2)	2 (prob 0.8)

$$\mathrm{VAL}_{EDT}(x) = \sum_{s \in S} P(s|x) * U(s \wedge x),$$

where S is a partitioning of states of nature.

Evidential decision theory (EDT)

It is rational to decide upon an alternative x if and only if there is no other alternative with higher VAL_{EDT} than x.

$$\mathrm{VAL}_{EDT}(\text{Study}) = 0.8 * 3 + 0.2 * 1 = 2.6$$

$$\mathrm{VAL}_{EDT}(\text{Not study}) = 0.2 * 4 + 0.8 * 2 = 2.4$$

Evidential Dominance

If there is a partition of states of nature such that it is evidentially independent of the acts and relative to it, an act x weakly dominates an act y, then x should be chosen rather than y.

A partition of states of nature is evidentially independent of the acts if and only if for each act x and each state s, $P(s|x) = P(s)$.

Roughly, when the states are evidentially independent of the acts, none of the states becomes more or less credible, when you make your choice.

Newcomb's Problem

Newcomb's problem was discovered by the physicist William Newcomb in 1960 while pondering the similarly structured prisoners' dilemma. Through a mutual friend, the problem reached Robert Nozick, who first to discuss the problem in print in 1969.

"Suppose a being in whose power to predict your choices you have enormous confidence. [...] There are two boxes, (B1) and (B2). (B1) contains \$ 1000. (B2) contains either \$ 1 000 000 (\$ M), or nothing. [...]

$$\text{(B1)}\ \{\$\ 1000\} \qquad \text{(B2)}\ \left\{\begin{array}{c}\$\ M\\ \text{or}\\ \$\ 0\end{array}\right\}$$

You have a choice between two actions:

(1) taking what is in both boxes

(2) taking only what is in the second box.

Furthermore, and you know this, the being knows that you know this, and so on:

(I) If the being predicts you will take what is in both boxes, he does not put the \$ M in the second box.

(II) If the being predicts you will take only what is in the second box, he does put the \$ M in the second box.

The situation is as follows. First the being makes its prediction. Then it puts the \$ M in the second box, or does not, depending upon what it has predicted. [...] What do you do?"

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000	\$1,000
Take one box	\$1,000,000	\$0

Robert Nozick (1969, p. 115)

"*First argument:* If I take what is in both boxes, the being, almost certainly, will have predicted this and will not have put the $ M in the second box, and so I will, almost certainly, get only $ 1000. If I take only what is in the second box, and so I will, almost certainly, get $ M. Thus, if I take what is in both boxes, I, almost certainly, will get $ 1000. If I take only what is in the second box, I, almost certainly, will get $ M. Therefore I should take only what is in the second box."

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000	\$1,000
Take one box	\$1,000,000	\$0

Robert Nozick (1969, p. 115)

"*Second argument:* The being has already made his prediction, and has already either put the \$ M in the second box, or has not. The \$ M is either already sitting in the second box, or it is not, and which situation obtains is already fixed and determined. If the being has already put the \$ M in the second box, and I take what is in both boxes I get \$ M + \$ 1000, whereas if I take only what is in the second box, I get only \$ M. If the being has not put the \$ M in the second box, and I take what is in both boxes I get \$ 1000, whereas if I take only what is in the second box, I get no money. Therefore, whether the money is there or not, and which it is already fixed and determined, I get \$ 1000 more by taking what is in both boxes rather than taking only what is in the second box. So I should take what is in both boxes."

Should a rational agent take one or two boxes in Newcomb's problem?

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000	\$1,000
Take one box	\$1,000,000	\$0

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000 (prob 0.01)	\$1,000 (prob 0.99)
Take one box	\$1,000,000 (prob 0.99)	\$0 (prob 0.01)

Assume that the utility function for money is linear.

$$\mathrm{VAL}_{EDT}(\text{Take two boxes})$$
$$= 1{,}001{,}000 * 0.01 + 1{,}000 * 0.99 = 11{,}000$$

$$\mathrm{VAL}_{EDT}(\text{Take one box})$$
$$= 1{,}000{,}000 * 0.99 + 0 * 0.01 = 990{,}000$$

A subjunctive conditional $x \mathrel{\Box\!\!\rightarrow} y$ is read as 'if it were the case that x, it would be the case that y'.

For example, if

> x says that 'Kangaroos have no tails'

and

> y says that 'Kangaroos topple over'

then

> $x \mathrel{\Box\!\!\rightarrow} y$

says that

> 'If kangaroos had no tails, they would topple over.'

These counterfactual conditionals are notoriously difficult to analyse.

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000	\$1,000
Take one box	\$1,000,000	\$0

Dominance with causal independence

If the states of nature are causally independent of the acts, it is not rational to decide upon an act x if there is an option y such that there is at least one positively probable state where the outcome of y is strictly preferred to the outcome of x and no state where the outcome of y is not weakly preferred to the outcome of x.

The states of nature are causally independent of the acts if and only if for all acts x and y and all states of nature s, $P(x \mathbin{\square\!\!\rightarrow} s) = P(y \mathbin{\square\!\!\rightarrow} s)$.

$$\mathrm{VAL}_{CDT}(x) = \sum_{s \in S} P(x \mathbin{\square\!\!\rightarrow} s) * U(x \wedge s)$$

Causal decision theory (CDT)

It is rational to decide upon an alternative x if and only if there is no other alternative with higher VAL_{CDT} than x.

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000	\$1,000
Take one box	\$1,000,000	\$0

$$\mathrm{VAL}_{CDT}(x) = \sum_{s \in S} P(x \mathrel{\square\!\!\rightarrow} s) * U(x \wedge s)$$

$$\mathrm{VAL}_{CDT}(\text{Take two boxes}) =$$

$$P(\text{Take two boxes} \mathrel{\square\!\!\rightarrow} \text{Prediction one boxing}) * U(\$1{,}001{,}000)$$

$$+ P(\text{Take two boxes} \mathrel{\square\!\!\rightarrow} \text{Prediction two boxing}) * U(\$1{,}000)$$

$$\mathrm{VAL}_{CDT}(\text{Take one box}) =$$

$$P(\text{Take one box} \mathrel{\square\!\!\rightarrow} \text{Prediction one boxing}) * U(\$1{,}000{,}000)$$

$$+ P(\text{Take one box} \mathrel{\square\!\!\rightarrow} \text{Prediction two boxing}) * U(\$0)$$

Medical Newcomb Cases

Andy Egan (2007, p. 94)

> *The Smoking Lesion*
>
> *Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer. Should Susan smoke? It seems clear that she should.*

Prisoners' Dilemma as a Newcomb Problem

	I rat	I don't rat
You rat	I get 10 years in prison You get 10 years of prison	I get 25 years in prison You get 1 year in prison
You don't rat	I get 1 year in prison You get 25 years in prison	I get 2 years in prison You get 2 years in prison

David Lewis (1979)

If your choice is reliable evidence that your co-prisoner will choose the same way, prisoners' dilemma is a Newcomb problem.

	I rat	I don't rat
You rat	I get \$1,000 You get \$1,000	I get \$0 You get \$1,001,000
You don't rat	I get \$1,001,000 You get \$0	I get \$1,000,000 You get \$1,000,000

Standard unconditional expected utility

$$\mathrm{VAL}(x) = \sum_{s \in S} P(s) * U(s \wedge x)$$

Evidential decision theory (EDT)

$$\mathrm{VAL}_{EDT}(x) = \sum_{s \in S} P(s|x) * U(s \wedge x)$$

Causal decision theory (CDT)

$$\mathrm{VAL}_{CDT}(x) = \sum_{s \in S} P(x \mathrel{\square\!\!\rightarrow} s) * U(x \wedge s)$$

The Tickle Defence of Evidential Decision Theory

Brian Skyrms (1980, pp. 130–131)

The Tickle Defence accepts that it would be right to smoke in the smoking lesion, but challenges the claim that evidential decision theory gives the opposite answer.

The essence of the defence is that decision theory deals with voluntary actions, not involuntary movements; and it is plausible that voluntary actions necessarily have mental causes. Hence the genes in the smoking lesion case must act through the mind if the case is to trouble evidential decision theory.

But rational agents should know their own minds. Hence the conditional probability of a state of nature will be equal to its unconditional probability. And then evidential decision theory recommends smoking.

Replies to the Tickle Defence

Brian Skyrms (1980, p. 131)

We need not, in every decision situation, be in possession of knowledge of some convenient factor which screens off any probabilistic relationship which does not mirror a causal relationship.

David Lewis (1981, p. 10)

> *I reply that The Tickle Defence does establish that a Newcomb problem cannot arise for a fully rational agent, but that decision theory should not be limited to apply only to the fully rational agent. Not so, at least, if rationality is taken to include self-knowledge.*

Objection to taking two boxes: Most people who take one box in Newcomb's problem ends up rich, and most people who take two boxes do not get rich.

Standard reply: The world might in some cases be rigged against rational people.

The Toxin Puzzle

Gregory S. Kavka (1983, pp. 33–34)

> *[An eccentric billionaire] places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. [. . .] The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. [. . .] All you have to do is [. . .] intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin.*

Frank Arntzenius (2008)

Let Mary be an evidential decision theorist who is to bet repeatedly on the outcome of a sequence of Yankees versus Red Sox games. Mary is convinced that in the long run the Yankees will win 90% of the time. However, on each occasion just before she chooses which bet to place, a perfect predictor of her choices and of the outcomes of the games announces to her whether she will win her bet or lose it.

> A bet on the Yankees: Mary wins \$1 if the Yankees win, loses \$2 if the Red Sox win.
>
> A bet on the Red Sox: Mary wins \$2 if the Red Sox win, loses \$1 if the Yankees win.

On some occasions the predictor says "Mary, you will lose your next bet". After Mary has updated her credences on this information she calculates:

$$VAL_{EDT}(\text{bet Y}) = P(\text{Y win} \mid \text{bet Y}) * U(\text{Y win \& bet Y})$$
$$+ P(\text{Y lose} \mid \text{bet Y}) * U(\text{Y lose \& bet Y}) = -2$$

$$VAL_{EDT}(\text{bet R}) = P(\text{R win} \mid \text{bet R}) * U(\text{R win \& bet R})$$
$$+ P(\text{R lose} \mid \text{bet R}) * U(\text{R lose \& bet R}) = -1$$

So she bets on the Red Sox each time she is told she will lose her bet.

A bet on the Yankees: Mary wins \$1 if the Yankees win, loses \$2 if the Red Sox win.

A bet on the Red Sox: Mary wins \$2 if the Red Sox win, loses \$1 if the Yankees win.

On the other occasions she is told "you will win you next bet". She updates her credences and finds:

$$VAL_{EDT}(\text{bet Y}) = P(\text{Y win} \mid \text{bet Y}) * U(\text{Y win \& bet Y}) + P(\text{Y lose} \mid \text{bet Y}) * U(\text{Y lose \& bet Y}) = 1$$

$$VAL_{EDT}(\text{bet R}) = P(\text{R win} \mid \text{bet R}) * U(\text{R win \& bet R}) + P(\text{R lose} \mid \text{bet R}) * U(\text{R lose \& bet R}) = 2$$

So she also bets on the Red Sox each time she is told she will win her bet. So Mary will always bet on the Red Sox. And, if the Yankees indeed win 90% of the time, she will lose money, big time. Now, of course, she would have done much better had she just ignored the announcements, and bet on the Yankees each time. But, being an evidential decision theorist she cannot do this.

A bet on the Yankees: Mary wins \$1 if the Yankees win, loses \$2 if the Red Sox win.
A bet on the Red Sox: Mary wins \$2 if the Red Sox win, loses \$1 if the Yankees win.

Death in Damascus

Allan Gibbard and William L. Harper (1978, pp. 157–158)

A man in Damascus knows that he has an appointment with Death at midnight. He will escape Death if he manages at midnight not to be at the place of his appointment. He can be in either Damascus or Aleppo at midnight.

As the man knows, Death is a good predictor of his whereabouts. If he stays in Damascus, he thereby has evidence that Death will look for him in Damascus. However, if he goes to Aleppo he thereby has evidence that Death will look for him in Aleppo. Wherever he decides to be at midnight, he has evidence that he would be better off at the other place.

	Death in Damascus	Death in Aleppo
Go to Damascus	Dead	Alive
Go to Aleppo	Alive	Dead

$$\mathrm{VAL}_{CDT}(x) = \sum_{s \in S} P(x \mathbin{\square\!\!\rightarrow} s) * U(x \wedge s)$$

If the agent goes to Damascus, he will have evidence for that

Go to Aleppo $\mathbin{\square\!\!\rightarrow}$ Death in Damascus

So causal decision theory recommends that he goes to Aleppo. But if the agent goes to Aleppo, he will have evidence that

Go to Damascus $\mathbin{\square\!\!\rightarrow}$ Death in Aleppo

So causal decision theory recommends that he goes to Damascus. Hence neither choice is stable.

Ratifiability

	Prediction one boxing	Prediction two boxing
Take two boxes	\$1,001,000	\$1,000
Take one box	\$1,000,000	\$0

Objection to one boxing: the choice of one boxing is not ratifiable. If chooses to take one box, one gets the good news that one is probably going to be rich, since the predictor has probably put a million in the box. But one will also regret one's decision, since taking also the other box would have made one even richer.

Ratificationism

Richard Jeffrey (1983, pp. 19–20)

$$\mathrm{VAL}(x) = \sum_{s \in S} P(s) * U(s \wedge x),$$

where S is a partitioning of states of nature.

An option x is *ratifiable* if and only if there is no alternative y such that $\mathrm{VAL}(y)$ exceeds $\mathrm{VAL}(x)$ on the supposition that x is decided upon.

Ratificationism

It is rational to decide upon an option x if and only if x is the only ratifiable option.

Andy Egan (2007, p. 97)

> *The Psychopath Button*
>
> *Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying.*

Egan takes not pushing to be intuitively rational in this case.

	Psycho	Not psycho
Push	Death	Psychopath-free world
Not push	Status quo	Status quo

Given that Paul starts off with a sufficiently low credence in that he is a psychopath, casual decision theory recommends him to press the button.

$$\mathrm{VAL}_{CDT}(x) = \sum_{s \in S} P(x \mathrel{\square\!\!\rightarrow} s) * U(x \wedge s)$$

$$\begin{aligned}&\mathrm{VAL}_{CDT}(\text{Push}) = \\ &P(\text{Push} \mathrel{\square\!\!\rightarrow} \text{Psycho}) * U(\text{Death}) \\ &+ P(\text{Push} \mathrel{\square\!\!\rightarrow} \text{Not psycho}) * U(\text{Psychopath-free world})\end{aligned}$$

$$\begin{aligned}&\mathrm{VAL}_{CDT}(\text{Not push}) = \\ &P(\text{Not push} \mathrel{\square\!\!\rightarrow} \text{Psycho}) * U(\text{Status quo}) \\ &+ P(\text{Not push} \mathrel{\square\!\!\rightarrow} \text{Not psycho}) * U(\text{Status quo})\end{aligned}$$

How does evidential decision theory and ratificationism fare?

	Psychopath	Not psychopath
Push	Death (prob 0.99)	Psychopath-free world (prob 0.01)
Not push	Status quo (prob 0.01)	Status quo (prob 0.99)

$$\mathrm{VAL}_{EDT}(x) = \sum_{s \in S} P(s|x) * U(s \wedge x).$$

Evidential decision theory recommends not pushing.

Ratificationism, however, does not recommend any option.

The choice to not push is unratifiable, since you then learn that you are probably not a psychopath. And then it is best to push the button.

But the choice to push is also unratifiable. If you chose to push you learn that you probably are a psychopath. And then it is best to not push the button.

Lexical ratificationism

It is rational to decide upon an option x if, and only if,

(1) x is ratifiable and there is no other ratifiable option with higher VAL_{EDT} than x, *or*

(2) there are no ratifiable options, and no other (unratifiable) option has higher VAL_{EDT} than x.

Chooses	*Best of with*
Push	Not push
Not push	Push

Since neither pushing nor not pushing is ratifiable, lexical ratificationism recommends not pushing, which has a higher VAL_{EDT} than pushing.

Andy Egan (2007, p. 112)

> *The Three-Option Smoking Lesion*
>
> *Samantha has three options: Smoke cigars, smoke cigarettes, or refrain from smoking altogether. Call these options CIGAR, CIGARETTE, and NO SMOKE. Due to the ways that various lesions tend to be distributed, it turns out that cigar smokers tend to be worse off than they would be if they were smoking cigarettes, but better off than they would be if they refrained from smoking altogether. Similarly, cigarette smokers tend to be worse off than they would be smoking cigars, but better off than they would be refraining from smoking altogether. Finally, nonsmokers tend to be best off refraining from smoking.*

Chooses	*Best of with*		
CIGAR	CIGARETTE	$\succ$ CIGAR	$\succ$ NO SMOKE
CIGARETTE	CIGAR	$\succ$ CIGARETTE	$\succ$ NO SMOKE
NO SMOKE	NO SMOKE	$\succ$ CIGAR	$\sim$ CIGARETTE

Ratificationism and lexical ratificationism recommends NO SMOKE, since it is the only ratifiable option. But this seems wrong, since, if you find yourself choosing CIGAR or CIGARETTE, you have good reason to think that NO SMOKE is not the way to go.

References

Arntzenius, Frank (2008) 'No Regrets, or: Edith Piaf Revamps Decision Theory', *Erkenntnis* 68 (2): 277–297.

Egan, Andy (2007) 'Some Counterexamples to Causal Decision Theory', *The Philosophical Review* 116 (1): 93–114.

Gibbard, Allan and Willian L. Harper (1978) 'Counterfactuals and Two Kinds of Expected Utility', in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory*, vol. I, pp. 125–162, Dordrecht: Reidel.

Gustafsson, Johan E. (2011) 'A Note in Defence of Ratificationism', *Erkenntnis* 75 (1): 147–150.

Jeffrey, Richard C. (1983) *The Logic of Decision*, Chicago: University of Chicago Press, second edn.

Kavka, Gregory S. (1983) 'The Toxin Puzzle', *Analysis* 43 (1): 33–36.

Lewis, David (1979) 'Prisoners' Dilemma Is a Newcomb Problem', *Philosophy & Public Affairs* 8 (3): 235–240.

— (1981) 'Causal Decision Theory', *Australasian Journal of Philosophy* 59 (1): 5–30.

Nozick, Robert (1969) 'Newcomb's Problem and Two Principles of Choice', in Nicholas Rescher, ed., *Essays in Honor of Carl G. Hempel*, pp. 114–146, Dordrecht: Reidel.

Skyrms, Brian (1980) *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*, New Haven, CT: Yale University Press.