Metatickles and Ratificationism

Jordan Howard Sobel

University of Toronto

Responses to Newcomb-like challenges to evidential decision theories such as Jeffrey's "logic of decision" range from allegations of incoherence and irrelevance; through stonewalling - "Just one box for me, thank you."; to arguments that maintain that when *properly* applied by an *ideal* agent such theories get the right answers and, for example, prescribe the taking of both boxes, not just one; on to conservative revisions of evidential decision theories that are held to get these supposedly right answers while remaining true to its evidential, Humean spirit; and finally to responses of radical revisionists who claim that only causal decision theories can get all Newcomb-like problem cases right. I discuss responses of the first sort in another paper - "Newcomb-like Problems for Jeffrey's Logic of Decision." The present paper is concerned with middle terms in the above response-progression - with a certain non-stonewalling defense of evidential decision theories unrevised, and with a certain conservative revision. Throughout I assume that Jeffrey's "logic of decision" can for purposes at hand stand for all evidential decision theories - that for present purposes there are no relevant differences among these theories. In Jeffrey's theory the Desirability of an action is a weighted average of the Desirabilities of its possible outcomes, the weight for the Desirability of a particular outcome being the probability of that outcome conditional on this action. In this theory conditional probabilities are defined as quotients of unconditional probabilities - $P(q/p) = P(p \& q)/P(p)$ - and measure possible evidential bearings of propositions (which of course need not coincide with their perceived possible causal bearings). "The Bayesian principle [of this theory] is to *perform an act which has maximum desirability*." (Jeffrey 1965, p. 1).

1. Metatickles: A Defense

Eells writes that "if a decision maker appropriately 'monitors' the relevant aspects of his deliberation (e.g., he knows what his beliefs and desires are), then evidential decision theory will deliver correct prescriptions." (Eells 1985). Eells' argument, in one of its versions, begins with the idea that if a fully rational agent applied evidential decision theory he would calculate Desirabilities, incline tentatively towards one or another decision, recalculate Desirabilities taking this

PSA 1986, Volume 1, pp. 342-351

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193135 Published online by Cambridge University Press

tentative inclination into account, adjust his inclinations to decisions, and so on, until his Desirabilities, his credences for actions and circumstances, and his inclinations to action had attained a steady state. Eells maintains that this process of deliberation with feedbacks would eventually lead to credences that made things probabilistically independent of actions of which they were believed to be causally independent. *Suppose* that Eells is right about this. (For his reasons see Eells 1984). Whether or not he is is a good question, but not a presently important one. For even granted this claim of eventual probabilistic independence, Eells' defense of evidential decision theory fails. The process of ideal deliberation described by Eells would not always settle on the right action, even if it *would* always result in the elimination of evidential bearings that were not based on perceived causal bearings.

I have not given a full statement of Eells' argument. Left out has been its *conclusion* - that once acts had, in the way indicated, been rendered probabilistically irrelevant to circumstances, evidential and causal decision theory would agree, and the rational act would excel in Desirability. Now this conclusion follows for problems in which the rational action is a *dominant* action, but not all Newcomb-like problem cases are like that. Here is one that is not like that:

> "*The Popcorn Problem*... .I want very much to have some popcorn," but "I am nearly sure that the popcorn vendor has sold out... .And so...I have decided not to go for popcorn... .Still, I am also nearly sure that in *this* theatre, when and only when there *is* popcorn...the signal - POPCORN!!! - is flashed on the screen, though at a speed that permits only subliminal, unconscious awareness... .[And, since] I consider myself to be a highly *suggestible* person... .I am nearly sure that I will change my mind and *go* for popcorn if and only if I am influenced by this subliminal signal to do so... .[And so] while I think it is very unlikely that I will go for popcorn...I think it is *much more* unlikely that I will go for popcorn though there is none...to be bought... .In short...my going for popcorn would provide me with a near certain sign of...there being popcorn.... .And my *not* going would provide me with a near certain sign that there is *not* popcorn... ." (J. H. Sobel forthcoming).

This case is to have the following qualitative Desirability matrix:

	Popcorn	No Popcorn
Go	very good	very bad
Don't Go	bad	so-so

And the case has the following probability matrix in which `[1]' stands for `nearly 1' and `[0]' for `nearly 0'. (How near to *x* is *[x]* to be? Near enough to make true the equalities and inequalities I endorse! This covers the present case and indeed all cases discussed in this paper. Greater specificity, though of course possible given numerical assumptions concerning Desirabilities, would make things more complicated and less clear. I note that from P(popcorn/go) = [1] and P(no popcorn/don't go) = [1] it does not follow that P(popcorn/go) = P(no popcorn/don't go).)

	[0]	[1]
	Popcorn	No Popcorn
Go	[1]	[0]
Don't Go	[0]	[1]

Numbers above the columns are my *unconditional* probabilities for there being and not being popcorn. Numbers in the cells are my probabilities for these circumstances *conditional on* my actions. These numbers, these conditional probabilities, are in the story founded on beliefs I have concerning possible causes of my actions, but they could be provided with very different kinds of bases. For example, I could be supposed to think that the manager was a very reliable predictor who acted on his predictions in order to sell as much popcorn as possible, and waste as little as possible.

Suppose that in the Popcorn Problem calculations and recalculations of Desirability *would* lead eventually to my actions being probabilistically independent of whether or not POPCORN!!! is flashing. Since neither going nor not going are dominant in this case, this supposition does *not* entail that not going, the *rational* action in the case, would eventually, and then steadily, excel in Desirability. Since I am nearly sure that there is no popcorn to be had (see the number [1] above the column for no popcorn), not going for popcorn is the rational thing to do. But not going does not excel in Desirability *in the beginning*, and *would* not excel eventually in deliberation that consisted from the beginning in the application and reapplication of evidential decision theory. *Going* for popcorn would excel in Desirability *initially*, and recalculations of Desirabilities that took into account that given the results of earlier calculations it was more likely that going for popcorn was what I would do, could only *confirm* the results of earlier calculations, and reinforce my conviction, committed Desirability maximizer that I take myself to be, that I was indeed going for popcorn. If this process would *also* reduce nearly to nil evidential bearings of my actions on the problem's states, it would lead to the following probabilities.

	[1]	[0]
	Popcorn	No Popcorn
Go	[1]	[0]
Don't Go	[1]	[0]

Eells' Continual CEU-Maximization, his CEU-maximization deliberation with "feedback loops", would in the beginning, all along, and in the end

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193135 Published online by Cambridge University Press

find *going* for popcorn to be my most Desirable act, whereas *not* going is what would be rational, and what would be prescribed by any causal decision theory. It seems plain that *no* version of Eells' defense can work in this case. Possibly he would have given up on "tickles" long ago had he not concentrated on problems like Newcomb's Problem itself not *only* in that in them evidential and causal decision theories diverge, but like it *also* in that in them the rational action is a *dominant* action.

2. Ratificationism: A Conservative Revision

Tickle defenses do not always work. There are cases, including in particular non-dominance cases, in which they would fail to align prescriptions of causal and evidential decision theories for ideally rational and sophisticated agents. And tickle defenses can seem unsatisfactory for another reason. It can seem that even if one worked, its success would be limited - that it could show only that an evidential decision theory can get right answers when applied by strongly rational agents, whereas, to quote Eells, "it is desirable for a decision theory to be applicable to less sophisticated agents." (Eells 1985). This, Eells suggests, may be part of what motivated Richard Jeffrey's passing endorsement of Ratificationism - it may, I suggest, be part of what motivated Jeffrey's move from Eells' tickle defense to what can be viewed as a "tickle revision".

"Ratificationism," Jeffrey writes, "is the doctrine that the choiceworthy acts - relative to your beliefs and desires - are the ratifiable ones." (Jeffrey 1983, p. 19). An option is ratifiable, he tells us, if and only if "on the hypothesis that *that* option will finally be chosen" its Desirability is at least as great as that of any other option (p. 19).

For precision let probability function P_q come from a function P by *conditionalization* on *q*, $P(q) > 0$, if and only if $P_q(p) = P(p/q)$. Let the *Conditional Desirability* of *p* on *q*, $Des(p/q)$, be the weighted average of the values of *p*'s worlds in which the weight for the value of a world *w* is $P_q(w/p) = P_q(w \& p)/P_q(p)$. Let *a** be a *decision* to do *a*. And, for an agent with probability function P, let an action *a* such that $P(a^*) > 0$ be *ratifiable* if and only if, for every alternative, *a*′, $Des(a/a^*) \geq Des(a'/a^*)$.

Ratificationism is a revision of Jeffrey's "logic of decision", not a defense. It replaces the rule "Maximize Desirability" with the rule "Make ratifiable decisions", or, "romantically: `Choose for the person you expect to be when you have chosen.'" (p. 16). Ratificationism is a relatively *conservative* revision of "the logic of decision". It avoids the primitives that Jeffrey most wanted to avoid - causal primitives, including kinds of causal conditionals - though it does have one new primitive, namely, *decisions* as things distinct from and prior to actions. The question is whether Ratificationism is not only a revision attractive to Humean-minded theorists, but a *successful* revision that solves Newcomb-like problems. It is not. Jeffrey has come himself to this view. He thought otherwise in 1981, when he wrote "The Logic of Decision Defended" but had changed his mind by 1983, when he was putting finishing touches to the second edition of *The Logic of Decision*. I will comment on the reason that turned Jeffrey against Ratificationism, before detailing my own objection to it.

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193135 Published online by Cambridge University Press

The rule, "Make ratifiable decisions" agrees with the rule that it would replace, "Maximize Desirability", "in ordinary cases, but gives what I take to be the right solutions in many (alas! [Jeffrey writes] not all) of the bothersome cases where agents see their decisions as merely symptomatic of states of affairs that they would promote or prevent if they could." (p. xii. Starting here all page references are to Jeffrey 1983. The aside "(alas! not all)" was, I believe, something like a "Stop the press!" insertion.) We consider first a "bothersome" prisoner's dilemma in which the new rule improves on the old rule and gives the right answer.

"[T]he Bayesian principle [the old rule] advises each prisoner <u>not</u> to confess, if each sensibly sees his own choice as a strong clue to the other's and therefore assigns high subjective probabilities...to the other prisoner's doing whatever it is...that he himself chooses... ." (p. 16). Confessing is dominant in the following Desirability matrix (in which Desirabilities are negative inverses of prison-terms, and primes mark the other prisoner's actions):

	C′	-C′
C	-4	0
-C	-10	-1

Notwithstanding this dominance, however, on plausible assumptions, conditional probabilities for the other's acts on mine can be as in,

	C′	-C′
C	[1]	[0]
-C	[0]	[1]

On these assumptions, Des(-C) = [-1] > Des(C) = [-4], and the rule "Maximize Desirability" prescribes not confessing - bad advice, Jeffrey thinks. However, if we add to our story that "on each hypothesis about his <u>decision</u>, each prisoner...sees his own <u>performance</u> as predictively irrelevant to the other's," then the new rule "Make ratifiable decisions" prescribes confessing which is what Jeffrey wants prescribed (p. 17). For this version of the prisoner's dilemma we have the following matrices for conditional-on-my-decisions conditional probabilities for the other's actions on my own actions:

P_C*

	C′	-C′
C	[1]	[0]
-C	[1]	[0]

P(-C)*

	C′	-C′
C	[0]	[1]
-C	[0]	[1]

"As probabilities are independent of acts in each matrix, my dominant performance (confess) will have the higher estimated desirability on each hypothesis about my final decision." (p. 17). That is, given that C dominates -C in the Desirability matrix and the other's actions are probabilistically independent of mine in both of these matrices, Des(C/C*) > Des(-C/C*) (in numbers, [-4] > [-10]) and C is ratifiable, though in contrast Des[C/(-C)*] > Des[-C/(-C)*] (in numbers, [0] > [-1]) and -C is not ratifiable. In this prisoner's dilemma the new rule gets the right answer and improves on the old rule.

The new rule does not, however, improve on the old one in every bothersome prisoner's dilemma. It does not, for example, in what Jeffrey describes as a "more plausible variant...(suggested by Bas van Fraassen)... .[in which] the sorts of extraneous influences that would prevent me from confessing when I had decided to are likely to work on him, too, and to the same effect." (p. 20). Similarly, presumably, for the sorts of extraneous influences that would move me to confess notwithstanding a final decision not to confess. For this version of the case we have, instead of the above conditional probability matrices, the following ones.

$$P_{C^*} \quad \begin{array}{c|c|c|} & C' & -C' \\ \hline C & [1] & [0] \\ \hline -C & [0] & [1] \\ \hline \end{array} \qquad\qquad P_{(-C)^*} \quad \begin{array}{c|c|c|} & C' & -C' \\ \hline C & [1] & [0] \\ \hline -C & [0] & [1] \\ \hline \end{array}$$

Confessing is not ratifiable in this case - Des(-C/C*) = [-1] > Des(C/C*) = [-4]. Though confessing is, Jeffrey is sure, still the choiceworthy act, not confessing is the ratifiable one - Des[-C/(-C)*] = [-1] > Des[C/(-C)*] = [-4].

Whether or not van Fraassen's variant is more plausible is not important. What matters (though Jeffrey might disagree about this) is that this variant is a possible one, and that in it Ratificationism gives the wrong answer. For the record, however, and though this is not important, it is very doubtful that van Fraassen's variant is more plausible. It is part of this case that "I am nearly certain that my performance will accurately indicate my final choice," that I will do what I decide to do (p. 16). Consistent with that, what sort of extraneous factors might I believe could get in the way and "prevent me from confessing when I had decided to" - what might "news" (not necessarily true!) that I was not going to confess after all tell me? There are of course many things that might prevent me from carrying out my final decision to confess, including several things mentioned by Jeffrey in another connection: "death or a nonfatal cerebral hemorrhage might prevent me from carrying out a decision to confess, and a surreptitiously administered dose of sodium pentothal might set me to confessing in spite of my decision not to. But changing my decision does not count as a way of not managing to perform my chosen action. [Ratifiability involves suppositions of final decisions.]" (p. 18) It is, however, clearly not plausible that I should think that things such

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193135 Published online by Cambridge University Press

as *these* would happen to me if and only if they happened to the other prisoner. And it is not easy to see how the stipulation could be made plausible, that "the sorts of extraneous influences that would prevent me from confessing...are likely to work on him...to the same effect", given that these must be influences that I think might explain my failing to act on a final decision to confess, and thus cannot be things that would work by moving me to change my mind. For a more decisive objection to Ratificationism, based on a more plausible case, we consider popcorn problems again.

Suppose I am nearly sure that I will do whatever I decide finally to do. Then, were I to become sure that I will decide to go for popcorn, I would be nearly sure that there was popcorn there, and the improbable but still possible news that I was after all *not* going, might very well no longer provide me with any evidence at all that there was popcorn there. It would be natural to suppose that I am sure that whatever would account for my failing to act on a decision to *go* for popcorn - a failure I am, and would be, nearly sure would not happen - would have nothing to do with whether or not there was popcorn there in the lobby. What might I think could cause a slip between a decision to go for popcorn, the action pursuant to it of going for popcorn? Remember that I am nearly sure that I will do what I decide to do, and that what is at issue is evidential bearings of my actions given a certainty on my part of a *final* decision to go for popcorn. Well, as Jeffrey has noted, there is, depressingly, always the possibility of being hit by a truck or struck by a nonfatal cerebral hemorrhage. It would, be possible, but it would not be easy to tell a story in which I found *these* things to be evidentially linked to the presence or absence of popcorn.

So much for the effect on my credences, and on evidential bearings of my actions, of "news" that my decision will be to go for popcorn. This effect could well be to quite neutralize these bearings. What, however, about "news" of a final decision *not* to go? It is natural to suppose that, given "news" that my decision will be *not* to go - given a subjective *certainty* of this, and not just the present *near*-certainty - *subsequent* "news" that I *was* going after all, would *remain*, as it presently is, good evidence that there was popcorn to be had *after all*. It is natural to suppose that such *further* "news" would be for me evidence that, under the influence of signals on the screen my final decision was destined to come undone - that I was, notwithstanding a final decision not to go for popcorn, going for popcorn after all *under the influence of these signals* - that I was going to go for popcorn intentionally, knowing what I am doing and not being carried away to the lobby, even though not deliberately and consequent to a decision since by hypothesis I am certain of a *final* decision not to go. (Decisions are things distinct from and prior to all deliberate actions, but not, I think, all voluntary and intentional ones.) It is natural to suppose that "news" that, notwithstanding a final decision not to go, I was going after all, would be evidence that in one or another of many possible ways the *signal* was going to make me do it.

Matrices for my "*conditionalized* on decisions $\underline{g}^*$ and $(-\underline{g})^*$" conditional probabilities for circumstances given actions, could in a popcorn problem be - for definiteness we assume that they are - as follows:

P_{G*}

	P	-P
G	[1]	[0]
-G	[1]	[0]

P(-G)*

	P	-P
G	[1]	[0]
-G	[0]	[1]

These probabilities can obtain given stipulations all of which are natural and plausible. Supposing, however, these probabilities obtain, then, given the Desirabilities in a popcorn problem,

	Popcorn	No Popcorn
Go	very good	very bad
Don't Go	bad	so-so

though going for popcorn is ratifiable, not going for popcorn is not: Des(G/G*) > Des(-G/G*), but Des[G/(-G)*] > Des[-G/(-G)*]. And yet, since I am nearly sure that there is no popcorn there to be had, it is plain that not going for popcorn is my only sensible course. Ratificationism can give the wrong answer in a popcorn problem, just as it does in certain "bothersome" prisoner's dilemmas, given certain somewhat implausible stipulations. In partial contrast, however, Ratificationism gives wrong answers in popcorn problems even when, indeed precisely when, they are elaborated by natural and plausible stipulations.

3. Conclusions

Robert Nozick in his paper, "Newcomb's Problem and Two Principles of Choice" (Nozick 1969) portrayed Newcomb's Problem as an occasion for conflict between Dominance and Utility Maximization principles of choice. Soon, however, the Problem was recast. It came to be viewed as a scene for conflict between alternative maximization principles, and between two conceptions of utility, an evidential conception and a causal one. First formulated, it seems, by Robert C. Stalnaker in a letter to David Lewis (Stalnaker 1972), this perspective was developed and promoted by Allan Gibbard and William Harper in their paper, "Counterfactuals and Two Kinds of Expected Utility " (Gibbard and Harper 1978). From this perspective, the dominance of the two-box option is incidental. It is not important for central issues that the action prescribed by causal decision theory in Newcomb's Problem is a dominant action. What is important is only that it is different from the action prescribed by evidential decision theory, and yet plainly rational which it can be without being dominant.

Had the unimportance of dominance been noticed and stressed when Newcomb's Problem was recast as an occasion of conflict between two conceptions of utility - had other problems that lacked the dominance feature been promptly constructed and advertised in order to stress this new perspective - it is unlikely that tickle defenses of Jeffrey's "logic of decision" would have ever been tried, or that ratification-revisions, if even _contemplated_, would have ever been proposed. But then it is perhaps fortunate that the state of play has for some time not been entirely clear. For truth is patient, and even if, as I think, tickle defenses and Ratificationism are non-starters as general solutions to Newcomb-like problems, we are in debt to their authors for drawing attention to subjects, in particular, that of deliberation-dynamics, and to concepts, in particular, that of ideally stable and ratifiable decisions, which promise to be of lasting interest and value.

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193135 Published online by Cambridge University Press

References

Eells, Ellery. (1981). "Causality, Utility, and Decision." Synthese 48: 295-329.

-------------. (1984). "Metatickles and the Dynamics of Deliberation." Theory and Decision 17: 71-95.

-------------. (1985). "Causal Decision Theory." In PSA 1984, Volume 2. Edited by P.D. Asquith and P. Kitcher. East Lansing: Philosophy of Science Association.

Gibbard, Allan and Harper, William L. (1978). "Counterfactuals and Two Kinds of Expected Utility." In Foundations and Applications of Decision Theory, Volume 1. (The University of Western Ontario Series in Philosophy of Science, Volume 13.) Edited by C.A. Hooker, J.J. Leach, and E.F. McClennen. Dordrecht: Reidel. Pages 125-162.)

Jeffrey, Richard C. (1965). The Logic of Decision. New York: McGraw-Hill.

------------------. (1981). "The Logic of Decision Defended." Synthese 48: 473-492.

------------------. (1983). The Logic of Decision. 2nd ed. Chicago: University of Chicago Press.

Nozick, Robert. (1969). "Newcomb's Problem and Two Principles of Choice." In Essays in Honor of Carl G. Hempel. Edited by N. Rescher et al. Dordrecht: Reidel. Pages 114-146.

Sobel, Jordan Howard. (forthcoming). "Notes on Decision Theory: Old Wine in New Bottles." Australasian Journal of Philosophy.

Stalnaker, Robert C. (1972). "Letter to David Lewis: May 21, 1972." In Ifs: Conditionals, Belief, Decision, Chance, and Time. Edited by W.L. Harper, R. Stalnaker, and G. Pearce. Dordrecht: Reidel, 1981. Pages 153-190.

https://doi.org/10.1086/psaprocbienmeetp.1986.1.193135 Published online by Cambridge University Press