How Not to Manage the News

Anon

2020-12-22

J. Dmitri Gallow (2020) has proposed an adjustment to causal decision theory to handle cases like Death in Damascus. The adjustment is ingenious, but it creates problems that are bigger than those it aimed to solve.

Gallow's theory has two main parts, the first dealing with choice between two options, and the second extending the theory to choice between more than two options.

The part of the theory dealing with binary choice is easiest to understand in terms of regret.[1] A is preferable to B iff the chooser regrets choosing B when they could have chosen A more than they regret choosing A when they could have chosen B. More formally, let I be an *improvement* function, in the following sense. ('Improvement' here is basically the converse of regret.) $I_C(A, B)$ is the weighted average of $D(AK) - D(BK)$, where K is a possible state of the world that is causally independent of the choice, D measures the desirability of choice-state pairs, and the weights are given by $\Pr(K|C)$. That is, the weights are conditional probabilities of states given choices. Very very roughly, $I_C(A, B)$ measures how much better off you would have been choosing A rather than B, assuming you did actually choose C. Gallow is primarily interested in the special case where $A = C$; only that special case will play a role in what follows. The news value N of A over B is defined as $I_A(A, B) - I_B(B, A)$. In the case where there are only two options, A is strictly preferred to B iff $N(A, B) > 0$.

[1] I am simplifying a bit here; see section 2 of Gallow's paper for a more detailed presentation.

This is not an implausible view about binary choice. To position it against more familiar views, consider the special case where the possible states are predictions about choices made by a very good predictor. How good? To make our lives easier, we shall simplify the math a lot and say that for any X, the probability that the predictor predicted X given that X is chosen is 1. If we write PX for X is predicted, the simplifying assumption is $\Pr(PX|X) = 1$ for all X. And I will make this simplifying assumption for the rest of this paper. (It could be dropped if you do not like infallible predictors; it would just make the math a tiny amount more complicated.) Now imagine a very special case of this, where $D(PA \wedge A) > D(PA \wedge B)$ and $D(PB \wedge B) > D(PB \wedge A)$. In this case, both A and B are self-ratifying. If you imagine the chooser is playing a game with the predictor, where the chooser wants to maximise D and the predictor wants to make correct predictions, then both $\langle A, PA \rangle$ and $\langle B, PB \rangle$ are equilibria of the game. (This idea of treating the predictor as a player in a game is from Harper (1986).) In this case, Gallow's view will prefer A iff it is the risk-dominant equilibria in the sense of Harsanyi and Selten (1988). And it is plausible, as Harsanyi (1995) argues, that this is the right choice in such a case.

But what is novel about Gallow's view is not what he says about binary choice, but what he says about choice among a larger class of options. There have been a flurry of papers in recent years proposing something like this as the correct rule for binary choice.[2] Gallow offers a distinctive view, drawing on sophisticated work in voting theory, about how to extend this view to the general case. On the one hand, Gallow's extension is better justified, and in many cases more plausible, than the other extensions that have been offered. On the other hand, we shall soon see that it faces some serious counterexamples. I suspect this is bad news for all of these regret based theories, but that is a story for another day.

In order to see the counterexamples I am going to develop, we do not need to understand the full details of Gallow's theory. Indeed, for two of the examples we just need to see how it applies

[2] See, for example, Wedgwood (2013), Robert (2018), Podgorski (2020), Barnett (n.d.) for similar views, and Bassett (2015) for earlier criticisms. My criticisms are distinct from Bassett's, but I think complementary; his arguments appear to be motivated by the same kind of considerations that are behind the arguments offered here.

in the case of three-way choice. A first pass thing to say is that the preference ordering over the choices includes $X > Y$ iff $N(X, Y) > 0$. But this will not quite do for a couple of reasons. One of these has to do with cases where $N(X, Y) = N(W, Z)$, but that does not particularly matter for what will follow. The bigger reason is that this 'first pass' leads to cyclic preferences. It is possible that $N(A, B)$, $N(B, C)$ and $N(C, A)$ are all positive. And Gallow, rightly, wants to avoid that. So in that case, he says we should igore the smallest of the three. If $N(A, B)$, $N(B, C)$ and $N(C, A)$ are all positive, but $N(C, A)$ is the smallest of the lot, then we accept the first pass judgment that $A > B$ and $B > C$, and use transitivity to conclude $A > C$. That is the feature of his theory that will do a bit of work in what follows.

There is another point that will be important when we turn to an example with more than three choices. If $N(X, Y)$ is positive for all $Y \neq X$, then X is the Condorcet winner of the contest and is chosen overall. This is not something that is forced into the theory - he shows how it naturally falls out of other independently motivated principles. But it is true in the theory, and it will be relevant in the final example.

Our first counterexample starts with something like Stag Hunt, and adds an option that dominates one of the choices.

	PA	PB	PC
A	200	400	0
B	0	500	11
C	210	410	10

Remember that PX just means that the predictor predicts that X will be chosen, and a standing assumption is that for all X, $\Pr(PX|X) = 1$. Given that standing assumpption, it is easy enough to calculate the N values. $N(X, Y)$ is $(D(X \wedge PX) + D(X \wedge PY)) - (D(Y \wedge PX) + D(Y \wedge PY))$. So from the table we can read off:

- $N(A, B) = (200 + 400) - (0 + 500) = 100$, and hence $N(B, A) = -100$.

- $N(B, C) = (500 + 11) - (410 + 10) = 91$, and hence $N(C, B) = -91$.
- $N(C, A) = (210 + 10) - (200 + 0) = 20$, and hence $N(A, C) = -20$.

This would generate a cycle if we just looked at which values were positive. So we ignore the smallest positive N value, and the resulting ranking of the choices is $A > B > C$.

But this is absurd. For one thing, A is strongly dominated by C. For another, both evidential decision theory and (most plausible versions of) causal decision theory would agree on B. For yet another, consider the 'gamified' version of this problem where the numbers here are Row's payouts, and Column's payouts are 1 in each pair $\langle X, PX \rangle$, and 0 otherwise. The only equilibrium of the game is $\langle B, PB \rangle$. Even more strongly, the only pair that is even rationalizable[3] is $\langle B, PB \rangle$. That pair is the only strategy pair that survives iterated deletion of dominated strategies.[4] The player who selects B will (correctly) believe that they are not just maximising expected value, but getting the best possible outcome in the game. Intuition and theory agree that this is an easy case: the right choice is B. But Gallow's theory says otherwise: he says to choose A.

Now to be fair, Gallow does note that his theory sometimes recommends dominated options like A. But the example he gives of this phenomena is a case where no option seems particularly plausible. The gamified version of the case has no pure strategy equilibria, and every choice seems regrettable.[5] In that case he argues that being dominated is bad, but since every choice in the game is bad in one way or another, it might be that the dominated choice is the least bad option. That response is not available here. There is nothing wrong with choosing B. It produces,

[3] I'm using 'rationalizable' in the sense defined by Bernheim (1984) and by Pearce (1984). Gallow uses the same term for a different notion in a part of his paper I'm not discussing.

[4] There is a small wrinkle here. If we are deleting strategies, then we first delete A. After that, we have two possible grounds to delete PA. One is that it is weakly dominated by PB, and for that matter PC. Another is that it is strongly dominated by any proper mixture of PB and PC. Some people are suspicious of appeals to weak dominance, and others are suspicious of appeals to dominance by mixed strategies. So the argument here is not completely watertight. But the fact that there are two quite different ways to get this step to work makes it more credible.

[5] The restriction to pure strategies is important here. There are ratifiable mixed strategies in the game in question, and personally I think one of them is the optimal choice. The way that contemporary decision theorists handle mixed strategies is a topic for a much longer paper.

with probability 1, the best outcome on the table, and is endorsed (for different reasons) by both evidential and causal decision theory. There would need to be a very good reason to prefer a dominated option to it, and I rather doubt such a reason is forthcoming.

Onto the second example. Gallow notes that his theory does a good job of handling 'clone' cases: adding another option that duplicates an existing option does not change anything. That is a nice result. But the theory does less well handling dominated options in cases like the following.

	PA	PB	PC
A	11	1	-500
B	1	10	1
C	0	0	0

In this game, we have

- $N(A, B) = 1$, and hence $N(B, A) = -1$.
- $N(B, C) = 11$, and hence $N(C, B) = -11$.
- $N(C, A) = 489$, and hence $N(A, C) = -489$.

Again we would have a cycle, and cycles are bad, so we ignore the smallest positive N value. And the result is that the theory says that the ordering over the options is $B > C > A$.

But again, this is implausible. A is the natural choice according to both evidential decision theory and the best versions of causal decision theory. More importantly, look at the game we get if we just treat C as something that obviously will not be chosen, and hence the predictor knows we will not choose. (Why is C obviously bad? Its best case scenario is worse than A's worst case scenario. That is as strong a form of domination as you can find.)

	PA	PB
A	11	1
B	1	10

That is very similar to Gallow's Cake in Damascus example. In that case he argues, very plausibly, that the right choice is A. But adding the absurd choice C, along with the possibility that the predictor will predict C, leads to a change in choice. That should not happen. It is at least as bad as the complaints he makes about how causal decision theory handles clone choices.

Finally, look at a case that Gallow handles, I believe, the same way as other people who share his view on binary choice. Remember that I noted that on his view, if A is a Condorcet winner, if it is pairwise preferred to all other choices, then t is the overall best choice. This part of his view is not novel - though the elegant way he derives it is. Still, it leads to bad results. Note that for A to be the Condorcet winner in a situation where the predictor is perfect, we only have to look at three things: the values in the first row, the values in the first column, and the values along the main diagonal. But intuitively, the other values on the table might be relevant to which choice is best. So consider this case.

	PA	PB	PC	PD
A	0	3	3	3
B	1	1	1000	1000
C	1	1000	1	1000
D	1	1000	1000	1

If $X \neq A, N(A, X) = 1$. So A is the Condorcet winner, and is the best option according to Gallow's theory. (And, for that matter, most theories that use something like regret to make pairwise choices.)

But again, this is very unintuitive. Evidential decision theory rejects this conclusion; it says you should be indifferent between B, C and D. And causal decision theory rejects it. There is no probability distribution over the four states that makes A utility maximising. So what is there to say in favor of A? Well, if you choose one of B, C and D, you will regret not choosing A. But you will regret the other two options even more. And if you do choose A, you will wish you had done literally anything else, since you will probably get the worst outcome on the board.

So Gallow's strategy for extending this view about binary choice to a general theory runs into a number of problems. I have not really argued for it here, but I suspect a more general conclusion can be drawn from this. This part of Gallow's theory was inventive, sophisticated and carefully motivated - if it fails, we should be a little suspicious that there is any good way of turning this theory of binary choice into a general theory. And since this theory of binary choice is getting rather popular in the last couple of years, that would be a striking result. But proving that is for a longer paper.

References

Barnett, David James. n.d. "Graded Ratifiability." http://www.davidjamesbar.net/wp-content/uploads/2018/03/Barnett-Graded-Ratifiability-March-2018.pdf.

Bassett, Robert. 2015. "A Critique of Benchmark Theory." *Synthese* 192: 241–67. https://doi.org/10.1007/s11229-014-0566-3.

Bernheim, B. Douglas. 1984. "Rationalizable Strategic Behavior." *Econometrica* 52 (4): 1007–28. https://doi.org/10.2307/1911196.

Gallow, J. Dmitri. 2020. "The Causal Decision Theorist's Gudie to Managing the News." *The Journal of Philosophy* 117 (3): 117–49.

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis* 24 (1): 25–36. https://doi.org/10.1007/BF00183199.

Harsanyi, John C. 1995. "A New Theory of Equilibrium Selection for Games with Complete Information." *Games and Economic Behavior* 8 (1): 91–122. https://doi.org/10.1016/S0899-8256(05)80018-1.

Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge, MA: MIT Press.

Pearce, David G. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection."

Econometrica 52 (4): 1029–50. https://doi.org/10.2307/1911197.

Podgorski, Aberlard. 2020. "Tournament Decision Theory." *Noûs* tbc (tbc): xx–xx. https://doi.org/10.1111/nous.12353.

Robert, David. 2018. "Expected Comparative Utility Theory: A New Theory of Rational Choice." *The Philosophical Forum* 49 (1): 19–37. https://doi.org/10.1111/phil.12178.

Wedgwood, Ralph. 2013. "A Priori Bootstrapping." In *The a Priori in Philosophy*, edited by Albert Casullo and Joshua C. Thurow, 225–46. Oxford: Oxford University Press.