Synthese (2012) 187:123–145
DOI 10.1007/s11229-011-0022-6

Regret and instability in causal decision theory

James M. Joyce

Received: 30 September 2011 / Accepted: 30 September 2011 / Published online: 27 October 2011

Abstract Andy Egan has recently produced a set of alleged counterexamples to causal decision theory (CDT) in which agents are forced to decide among *causally unratifiable* options, thereby making choices they know they will regret. I show that, far from being counterexamples, CDT gets Egan's cases exactly right. Egan thinks otherwise because he has misapplied CDT by requiring agents to make binding choices before they have processed all available information about the causal consequences of their acts. I elucidate CDT in a way that makes it clear where Egan goes wrong, and which explains why his examples pose no threat to the theory. My approach has similarities to a modification of CDT proposed by Frank Arntzenius, but it differs in the significance that it assigns to potential regrets. I maintain, contrary to Arntzenius, that an agent facing Egan's decisions can rationally choose actions that she knows she will later regret. All rationality demands of agents it that they maximize unconditional causal expected utility from an epistemic perspective that accurately reflects all the available evidence about what their acts are likely to cause. This yields correct answers even in outlandish cases in which one is sure to regret whatever one does.

Keywords Expected utility · Ratifiability · Causal decision theory · Regret · Decision instability · Reflection principle · Dynamics of deliberation

For help with this paper, I would like to thank Gordon Belot, Richard Bradley, Aaron Bronfman, Andy Egan, Dmitri Gallow, Allan Gibbard, Alan Hájek, Bill Harper, Wlodek Rabinowicz, Jan-Willem Romeijn, Teddy Seidenfeld, Dan Singer, Paul Weirich, and audiences at the London School of Economics, University of Kent, University of Waterloo, University of Missouri, and the 2009 Progic Conference.

J. M. Joyce (✉)
Department of Philosophy, University of Michigan, Ann Arbor, MI, USA
e-mail: jjoyce@umich.edu

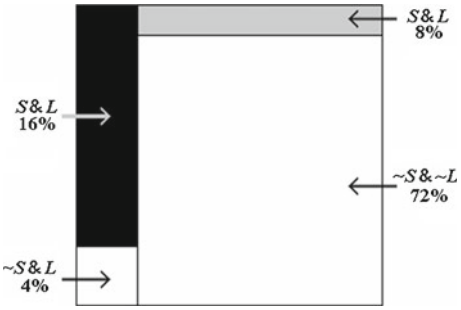

Fig. 1 In the *black region*, where you shoot (S) and have the lesion (L), you obtain the worst outcome $u\,(S, L) = -30$. In the *grey region*, where you shoot but lack the lesion, the best outcome $u\,(S, \sim L) = 10$ is achieved. In the white regions, where you cannot bring yourself to shoot, the status quo is preserved $u\,(\sim S, L) = u\,(\sim S, \sim L) = 0$

Egan (2007) has offered a series of purported counterexamples to causal decision theory (CDT) in which the choice of any act provides evidence about its own causal consequences, and this evidence undermines the act's rationale. Here is such a case:

> *Murder Lesion*. Life in your country would be better if you killed the despot Alfred. You have a gun aimed at his head and are deciding whether to shoot. You have no moral qualms about killing; your sole concern is whether shooting Alfred will leave your fellow citizens better off. Of course, not everyone has the nerve to pull the trigger, and even those who do sometimes miss. By shooting and missing you would anger Alfred and cause him to make life in your country much worse. But, if you shoot and aim true the Crown Prince will ascend to the throne and life in your country will improve. Your situation is complicated by the fact that you are a random member of a population in which 20% of people have a brain lesion that both fortifies their nerve and causes their hands to tremble when they shoot. Eight in ten people who have the lesion can bring themselves to shoot, but they invariably miss. Those who lack the lesion shoot only one time in 10, but always hit their targets. So, assuming for definiteness that the utility of killing Alfred has four times the magnitude of the disutility of shooting and missing,[1] your decision looks like Fig. 1.

Should you shoot?

The answer is not obvious. Since you know only the information given, your initial subjective probability estimates are $prob_0(S\,\&\,L) = 0.16$, $prob_0(S\,\&\sim L) = 0.04$, $prob_0(S\,\&\sim L) = 0.08$ and $prob_0(\sim S\,\&\sim L) = 0.72$. Thus, you initially see yourself as 20% likely to have the lesion and 24% likely to shoot.[2] Moreover, since $prob_0\,(S \mid L) = 0.8$ and $prob_0\,(S \mid \sim L) = 0.1$ you recognize a strong correlation between the presence/absence of the lesion and your tendencies toward/against

[1] This entails that you are entirely indifferent between the status quo and an arrangement in which a coin biased 3:1 in favor of tails is tossed and the bad/good outcome results from heads/tails. If one of these options seems better or worse to you, then you are operating with different utilities. The arguments of this paper go through just as well for different utility assignments.

[2] I assume throughout that it makes sense for agents to assign subjective probabilities to their potential actions in the course of their deliberations about what to do. There are decision theorists who disagree with this, most notably Levi (2000) and Spohn (1977). For defenses of act probabilities see Joyce (2002) and Rabinowicz (2002).