STRUCTURAL STABILITY IN CAUSAL DECISION THEORY

GREG LAURO AND SIMON M. HUTTEGGER

Abstract. There are decision problems in which rational deliberation fails to result in choosing a pure act. This phenomenon, known as *decision instability*, has been discussed in the literature on causal decision theory. In this paper we investigate another type of instability, called *structural instability*. Structural instability indicates that certain qualitative features of the process of rational deliberation are under-determined in a decision situation. We illustrate some of the issues arising from structural instability in terms of a recent argument against causal decision theory proposed by Caspar Hare and Brian Hedden. We show that their argument is undermined by considerations arising from decision instability and structural instability.

1. Introduction

The literature on causal and evidential decision theory is full of decision problems that involve complex informational feedback processes. The theory of rational deliberation developed in Skyrms (1990) has recently been used to investigate such situations (Arntzenius, 2008; Joyce, 2012). As demonstrated in those papers, introducing models of rational deliberation is crucial both for understanding the subtle issues involved in complex decision problems and for avoiding misleading conclusions.

The main argument in Hare and Hedden (2016) is a more recent example of the use of rational deliberation in causal decision theory. Based on a decision situation they call the "Three Crates" problem, Hare and Hedden attempt to show that evidential decision theory, and not causal decision theory, provides the correct normative account of decision making. Here, we present arguments against this conclusion. In a first step, we analyze Hare and Hedden's argument in detail (see §2). We find that it depends on a tacit assumption that, when stated explicitly, undermines another part of their argument. In a second step, we try to reconstruct Hare and Hedden's argument with the help of dynamical models of rational deliberation. The situation here turns out to be subtle. On the one hand, we find a type of *decision instability* similar to, but not quite the same as, the one discussed by Arntzenius (2008) or Joyce (2012), which has none of a decision problem's acts clearly being a rational choice (see §3). But the Three Crates problem exhibits another type of instability, known as *structural instability* in dynamical systems theory.[1] Structural instability implies that arbitrarily small perturbations of deliberational models lead to qualitative changes in the model's dynamics (see §4). In §5, we argue that both types of instability block a crucial step in Hare and Hedden's argument. We conclude that causal decision theory emerges unscathed and remains the normatively correct theory of decision making.

Date: December 15, 2019.

[1]See, e.g., Katok and Hasselblatt (1995, 68ff.).

	Crate A	Crate B	Crate C
She predicts you select A:	$\$1,000,000$	$\$1,001,000$	\$0
She predicts you select B:	\$0	\$0	$\$1,000$
She predicts you select C:	\$0	\$0	\$0

FIGURE 1. Three Crates problem.

2. Hare and Hedden's Argument

Hare and Hedden (2016) invite us to think about a modification of Newcomb's problem, the Three Crates problem. Crate A is intended to correspond to one-boxing (payoff $\$1,000,000$), Crate B corresponds to two-boxing (payoff $\$1,001,000$), and Crate C corresponds to taking just the box containing $\$1,000$. The standard setup of the Newcomb problem is slightly altered in order to frustrate the agent: if the agent is predicted to two-box (Crate B), then the predictor shifts the money away to the new Crate C; and if the agent is predicted to take Crate C, then there's no money at all (see Figure 1). Thus, the Three Crates problem is not just the Newcomb problem with another act made available, since the payoffs of the original decision are also changed.

Evidential decision theory recommends choosing Crate A, assuming that the predictor is sufficiently reliable. Hare and Hedden endorse this as the correct choice (Hare and Hedden, 2016, p. 616). We take issue with this, but postpone a discussion until §5. For now, we focus on their main argument, which purportedly shows that causal decision theory recommends choosing Crate C.

Hare and Hedden's argument is based on modeling the causal decision theorist as a *rational deliberator*. Deliberational processes differ in the details of how rational deliberation unfolds; what they have in common is that causal consequences of acts are evaluated in the light of the agent's current information, and that those evaluations may generate new information, allowing her to re-evaluate the causal consequences of acts. At the end of the deliberational process, the state of the agent reflects all the evidence about the likely consequences of her actions, putting her in a position to make a fully informed decision.

Joyce (2012) points out that this is crucial for rational choice theory, since a rational decision maker should not decide based on evaluations of acts which involve under-informed probabilities. That an agent ought to have incorporated all the available information about the outcomes her acts will causally promote is thus a central assumption of causal decision theory, even though it is not always stated explicitly.[2] Within a deliberational approach this reasoning process is made explicit. Sometimes, the level of precision achieved by a deliberational model is unnecessary. But in situations like the Three Crates problem it is important to proceed rigorously since it is not immediately clear what the correct choice may be.[3]

Hare and Hedden begin their argument by outlining four tenets of the self-aware causalist (Hare and Hedden, 2016, p. 616). The causalist is assumed to (1) respect weak dominance, (2) be sure that she respects weak dominance, (3) be sure of her

[2]For more information we refer the discussion in Joyce (2012). In a recent paper, Armendt (2019) develops a different view of rational deliberation that puts less emphasis on arriving at a state of full information. We think that Armendt provides an interesting account, but we'll work with the more idealized picture here for the sake of simplicity.

[3]The same goes for other decision problems in which acts are *unstable*.

	Red	Yellow	Green
Up	\$2, \$0	\$1, \$1	\$0, \$2
Middle	\$1, \$1	\$1, \$1	\$1, \$1
Down	\$3, \$0	\$0, \$3	\$2, \$0

FIGURE 2. Two-player, three-strategy game.

knowledge of the contents of the crates (here knowledge is expressed in terms of credences), and (4) not choose a crate that she is sure she will not choose. These assumptions provide a framework for rational deliberation. Hare and Hedden argue by elimination of cases that the causalist's deliberation will lead her into choosing Crate C.

Let's consider the argument in detail. Hare and Hedden first establish that the causal agent does not choose A. The deliberation goes as follows: Crate B weakly dominates Crate A. By assumptions (1), (2), and (3) the agent eliminates the option of choosing Crate A and is certain of this. Now comes the crucial step in the argument: *the agent concludes from this that the predictor will therefore predict that she will not choose A*. In other words, the agent is certain that the predictor is perfect. Let's call this the *perfect predictor assumption*. The perfect predictor assumption allows the agent to eliminate the state in which the predictor predicts that she selects Crate A; so she may examine the reduced decision problem in which the predictor predicts B or C and she chooses between B and C. Here, Crate C weakly dominates Crate B. Thus, again by (1), (2), and (3), the agent eliminates the option of choosing Crate B, arriving at the conclusion that she will choose C, which follows from (4).

An immediate problem for this argument is the perfect predictor assumption, which is not stated explicitly as an assumption by Hare and Hedden. Having the predictor be perfect is necessary for mimicking a procedure known in game theory as *elimination of weakly dominated strategies*, which essentially underlies Hare and Hedden's reasoning.[4] Consider the two-player, three-strategy game shown in Figure 2. In this game, the row player can choose between Up, Middle, and Down, and the Column player can choose between Red, Yellow, and Green. The left payoff in each cell of the table represents the row player's payoff and the right one represents the column player's payoff. Iterated elimination of weakly dominated strategies proceeds along the following lines. First, observe that Yellow weakly dominates Red (but not Green). Eliminating the weakly dominated strategy Red leaves a reduced game in which the column player only chooses between Yellow and Green. In this subgame, Up is weakly dominated by Middle. Eliminating Up then reduces the available strategies of the row player to Middle and Down, and so Yellow weakly dominates Green. Eliminating Green leads to a situation where Middle dominates Down. In this way iterated elimination of weakly dominated strategies leads the players to choose Middle and Yellow.[5]

[4]See Hare and Hedden (2016, ft. 17).

[5]This procedure only works under rather restrictive conditions. Both players must have knowledge of the structure of the game. For instance, the row player must know her own payoffs, but also the strategies and payoffs of her opponent; otherwise she could not conclude that the column player will not choose Red. In order to arrive at this conclusion, she also needs to know that her opponent eliminates weakly dominated strategies. Iterating the procedure also requires players to have higher-order knowledge of the structure of the game and the elimination procedures. That

Hare and Hedden's argument is not quite an instance of this procedure, since the Three Crates decision problem is not a game (the predictor is not a player). But their deliberational procedure is similar. The agent thinks that if she won't choose Crate A, then the predictor won't predict that she will choose A. More precisely, if she believes with probability one that she will not choose A, then she also believes with probability one that the predictor has predicted that she will not choose A. No value less than one will do: if she believes the predictor won't predict that she will choose A with some very high probability, she cannot eliminate the first row of the payoff table, in which case taking Crate C does not weakly dominate Crate B in the second elimination step.

The perfect predictor assumption is implausible. To start with, it is extremely unrealistic. For that reason, arguments in the causalist-evidentialist debate are usually required to be robust under having a less than perfect predictor (who might still be very reliable). In the Newcomb problem, for example, the predictor does not need to be perfect.

But the perfect predictor assumption faces a more serious problem. The assumption is, on the one hand, necessary to get weak dominance reasoning off the ground, but, on the other, it undermines the weak dominance principle. Consider the elimination of A (the argument works, mutatis mutandis, for the deletion of B once A is gone). Suppose the agent eliminates A because it is weakly dominated by B. This elimination implies that she is certain she will not choose A. Since the agent assumes the predictor to be perfect, she must also be certain that the predictor has not predicted that she will choose A. But this leaves her without any reason to eliminate A. Weak dominance reasoning only works if there is a positive probability that the dominating and the dominated options have different payoff consequences, which is not the case if the agent eliminates A. Thus the agent finds herself in an epistemic catch-22 situation: either she moves to eliminate A by virtue of weak dominance and thus undermines weak dominance reasoning because she no longer has a justification for applying it; or she prioritizes the weak dominance precondition and assigns positive probability to choosing A, thereby not eliminating the act at all. Trying to have it both ways—using weak dominance *and* eliminating A—is incoherent in light of the perfect predictor assumption.[6]

At another crucial point in the paper, Hare and Hedden are ambiguous regarding the perfect predictor assumption. After establishing, in the way described above, that a causalist will choose Crate C, Hare and Hedden claim that the causalist's choice is "not desirable even by their own lights" (Hare and Hedden, 2016, p. 615). We will discuss this argument in more detail in §5, but one of its elements is important for the present discussion. In order to argue for their conclusion, Hare and Hedden need to invoke that the causalist could have done better in relevantly similar situations, i.e. in situations where the predictor has predicted A or B. This seems to rely on treating the predictor as imperfect. If we don't, then we find no relevantly similar situations, and so there appears to be nothing wrong with

is, they need to know that the other player knows their strategies, payoffs, and the fact that they eliminate weakly dominated strategies.

[6]There is another, slightly different, problem. Assuming the agent believes with probability one that the predictor has predicted C, she has no preference among the acts. But it is also assumed that the predictor can predict with certainty that the agent will choose Crate C. What facts about the agent would such a predictor have to know? Since it cannot be anything about the preferences of the agent, we are left in the dark.

choosing Crate C—or A, or B, for that matter (see footnote 6). If the events that the predictor predicted A or B are both assigned probability zero, the corresponding payoffs have no effect on causal expected utilities.

Besides the perfect predictor assumption, the most significant problem with Hare and Hedden's argument is that they proceed as if assumptions (1) to (4) are constitutive of a self-aware causal decision theorist. The assumption that the causal decision theorists's reasoning be based on iterated elimination of weakly dominated acts is especially troubling. Weak dominance reasoning might be useful as a first approximation, but it is very limited. In particular, the elimination of weakly dominated strategies leads to radical leaps in the agent's beliefs, leaving her open to overlook more fine-grained information as to how causal expected utilities affect the relative attractiveness of acts. In addition, weak dominance reasoning is rarely used in game theory because the order in which players eliminate strategies can affect the outcome of the reasoning process; moreover, the procedure sometimes eliminates Nash equilibria.[7] For these reasons, the prototypical causalist's reasoning is best reflected by the models of rational deliberation we describe in the next two sections.

3. Rational Deliberation I: Ideal Predictor

The aim of this section is to reconstruct Hare and Hedden's argument in the setting of dynamical models of rational deliberation (Skyrms, 1990). This allows us to drop the problematic perfect predictor assumptions and strict adherence to weak dominance reasoning while, effectively, keeping the remainder of Hare and Hedden's setup.

Thus, the setting is that of an agent who endorses causal decision theory, is self-aware in the way described by Hare and Hedden, and seeks evidence on the causal payoff consequences of acts. But instead of eliminating weakly dominated strategies, she updates her probabilities for acts in the Three Crates problem by a deliberational process that, as Skyrms puts it, *seeks the good*. That is to say, the agent calculates the causal expected utility of acts given her current probabilities, but instead of immediately acting on these calculations, she updates probabilities of choosing acts incrementally by putting more probability on acts with higher causal expected utility. She then repeats this process. As a result, her choice probabilities move in the direction of improved causal expected utility.

Sometimes, this process reaches a *deliberational equilibrium*: a point at which the agent's deliberational process is at rest and updating on new information does not lead the agent to change her choice probabilities. At a deliberational equilibrium the agent can be said to have fully informed evaluations of causal expected utilities.

[7] Let us illustrate order effects in terms of the following game:

	Left	Right
Up	\$1, \$1	\$0, \$0
Middle	\$1, \$1	\$2, \$1
Down	\$0, \$0	\$2, \$1

If the row player applies weak dominance on Down, then the column player will eliminate Right by weak dominance, securing a payoff of \$1 for the row player. If instead the row player applies weak dominance on Up, then the column player will eliminate Left by weak dominance, securing a payoff of \$2 for the row player. These maneuvers fit the schematic form of weak dominance Hare and Hedden give, and so we're faced with the curiosity that different orders of deduction yield different outcomes.

Deliberational equilibria correspond to states where one can make a fully informed decision.

Before getting bogged down in generalities, let's explore the Three Crates problem in the context of some of Skyrms's deliberational dynamics.[8] We begin by considering the *Darwin map*.[9] A state of indecision in the Three Crates problem is described by a probability assignment p_i to each of the three acts (the index i refers to the act). The Darwin map takes a state of indecision, p_i, to a new state of indecision, p_i', which is given by

$$p_i' = p_i \frac{U(A_i)}{U(SQ)}$$

Here, U is the agent's causal expected utility function, and SQ denotes the status quo (the agent's present state of indecision). Accordingly, $U(A_i)$ is the causal expected utility of act A_i and $U(SQ)$ is the causal expected utility of the status quo.[10]

In calculating these causal expected utilities, we assume for now that the agent believes the predictor is *ideal*. Formally, if the agent's current choice probabilities for Crate A, Crate B, and Crate C are respectively p, q, and r, then the predictor will predict that the agent chooses Crate A with probability p, Crate B with probability q, and Crate C with probability r. The ideal predictor assumption is not quite the same as the perfect predictor assumption, which says that the predictor can say with certainty which act the agent will choose. Yet the ideal predictor assumption claims something similar, namely that the predictor has full knowledge of the agent's probability judgments throughout the process. This is problematic in its own way, and we will discuss in the next section what might happen if we drop it. For now we keep it in order to stay as close as possible to Hare and Hedden's setup: For choice probabilities p_i that adjust in a more sensitive way to new information, the ideal predictor assumption is the natural analog to the perfect predictor assumption.[11]

The Darwin map is a deliberational process that seeks the good. The status quo, $U(SQ)$, is the baseline causal expected utility with respect to which the causal consequences of acts are evaluated. An act A_i that has higher causal expected utility than $U(SQ)$ will be chosen with higher probability, $p_i' > p_i$. If A_i has lower causal expected utility than $U(SQ)$, its choice probability decreases, $p_i' < p_i$. If A_i has exactly the same causal expected utility as the status quo, its choice probability stays the same, $p_i' = p_i$. Thus, after each update the Darwin map moves a small step in the direction of improved causal expected utility.

---

[8]The models will utilize simplified payoffs for greater visibility in the diagrams. This induces no qualitative change in the dynamics. The new payoffs we'll use: $\left(\begin{smallmatrix} 10 & 11 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right)$. All calculations are carried out to a minimum of 500 digits of precision. In particular, each stage of the deliberation is calculated symbolically and then rounded to 500 digits.

[9]This dynamics is better known as the *replicator dynamics* in game theory. Skyrms (1990) shows that the Darwin map is a simple instance of a Bayesian learning dynamics.

[10]See Skyrms (1990, p. 37). If $U(A_i)$ is the causal expected utility of A_i and if p_i denote the status quo choice probabilities, then $U(SQ) = \sum_i p_i U(A_i)$.

[11]Let P_A denote the proposition that the predictor predicted A and C_A the proposition that the agent chooses A, and let $\mathbb{P}$ be the agent's probability. Then $\mathbb{P}(P_A) = \mathbb{P}(C_A)$ is equivalent to $\mathbb{P}(P_A|C_A) = \mathbb{P}(C_A|P_A)$ (whenever the relevant unconditional probabilities are non-zero). Thus the ideal predictor assumption amounts to saying that the propositions C_A and P_A depend on one another in a strictly symmetric way. This need not be the case, but it is also not implausible.

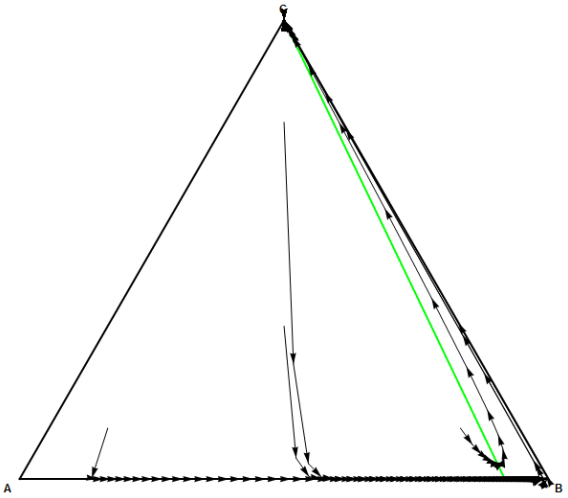

FIGURE 3. Darwin (Replicator) dynamics (100 iterations)

In Figure 3 we show some sample simulations of the Darwin map. The triangle represents the simplex of all choice probabilities over the three acts.[12] If the deliberational process starts in the upper part of the simplex (i.e., close to Crate C), the agent decreases her probability of choosing Crate C, then she prioritizes Crate B over Crate A, and finally begins moving toward Crate C over Crate B. This is broadly consonant with the process Hare and Hedden (2016) describe: the agent is first concerned in comparing Crates A and B, and eliminates Crate A as a contender. With Crate A excluded, the agent then compares Crates B and C in favor of Crate C.

Before giving a more detailed analysis of the deliberational dynamics, we consider two more dynamical processes. The point of this exercise is to show that the Darwin map is typical for dynamics that seek the good in the Three Crates problem.

The first dynamics is the *Nash map*. Let the *covetability* of an act A given status quo SQ be given by

$$cov(A) = \max(U(A) - U(SQ), 0).$$

We can interpret $cov(A)$ as a measure of how much better A looks compared to SQ. The *Nash map* updates the agent's choice probabilities, p_i, to

$$p_i' = \frac{kp_i + cov(A_i)}{k + \sum_i cov(A_i)},$$

where A_i is the act corresponding to p_i and $k > 0$ is the agent's *index of caution* (higher k gives slower movement).[13] Clearly, the Nash map also seeks the good and moves in the direction of improved causal expected utility.

Two sets of simulations of the Nash map for the Three Crates problem are shown in Figure 4. One is with a high and the other with a low level of caution. The qualitative features are similar to the dynamics of the Darwin map. There

[12]Cf. Figure 8 of Hare and Hedden (2016, p. 616) for their simpler assessment of the dynamics. Formally, the simplex is the set $\{\langle p, q, r\rangle \in \mathbb{R}^3 : p, q, r \geq 0, p + q + r = 1\}$.
[13]See Skyrms (1990, pp. 30-31).

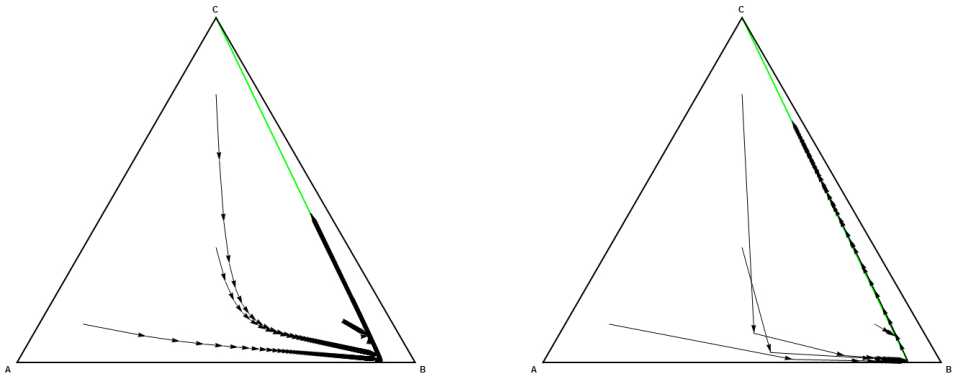

FIGURE 4. Nash dynamics, k=5 (250 iterations) and k=$\frac{1}{5}$ (25 iterations)

is a large portion of the simplex where the process moves away from the act of choosing Crate C before it moves towards choosing Crate B and ultimately back to Crate C. There is one difference to the Darwin map, however. Due to the small payoff difference between choosing Crate C and choosing it with very high probability, the deliberation moves toward Crate C extremely slowly. Moreover, the deliberation stagnates in the region near B. If we were to add considerations of resource-boundedness or level-k thinking, the agent might end up selecting Crate B with high probability.[14]

Finally, we look at the best response map. A pure strategy A_i is a *best response* in a state of indecision if the expected utility of A_i is greater than or equal to the expected utility of A_j for $i \neq j$. We can express this in our problem as

$$\sum_{k=1}^{n} p_k U(A_i, \text{predictor guesses } k) \geq \sum_{k=1}^{n} p_k U(A_j, \text{predictor guesses } k) \ (i \neq j),$$

where p_k is the choice probability of act k. We let $BR(SQ)$ denote a best response at the status quo, SQ.[15] The *best response map* takes the choice probability p_i to a new choice probability p_i' according to the following rule:

$$p_i' = \frac{kp_i + BR(SQ)}{k+1}$$

Here, $k > 0$ again is the agent's *index of caution*.[16] The best response map also seeks the good, but it does so in a more radical manner than the Darwin or the Nash map. While the latter two move in the direction of all acts that have higher than above average causal expected utility, the best response map moves only in the direction of the act with maximum causal expected utility. This can be seen in Figure 5. For most of the space in the simplex, choosing Crate B is the best

[14]See Camerer et al. (2004).

[15]The status quo, SQ, is a probability distribution over acts. According to the ideal predictor assumption the probability distribution carries over to states. The best response to SQ is an act that maximizes expected utility with respect to that distribution. It's possible that a best response may not be unique. For the purpose of calculating the dynamics with multiple best responses, we let $BR(SQ)$ be the uniform mixed act of those strategies. Nothing of substance depends on that assumption.

[16]See Hofbauer and Sigmund (1998, p. 94) for details.

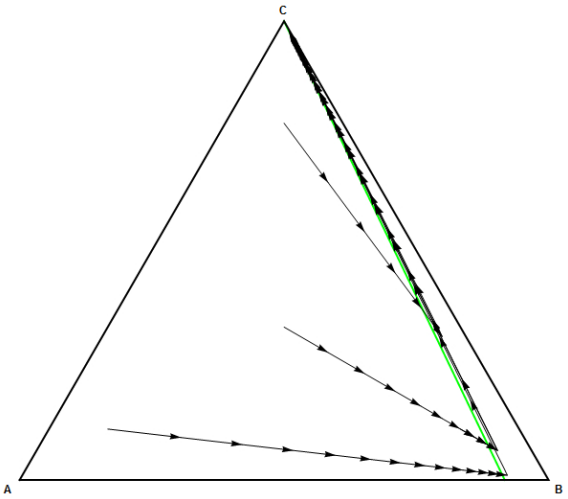

FIGURE 5. Best-response dynamics, k=5 (25 iterations)

response to the status quo. As the deliberational process comes close to B, however, the best response switches to Crate C. Overall, however, the picture is similar to the ones we get from the other two dynamics.

Among the three deliberational dynamics considered here, the best response dynamics is most faithful to Hare and Hedden's depiction of the reasoning process. The main difference is that the best response process allows agents to make up their minds gradually, while weak dominance reasoning eliminates consideration of an act in one step. As a result, the best response dynamics, along with the other two processes, exhibits a more fine-grained view as to what the current evidence says about the causal expected utility of acts than is possible in the Hare-Hedden setup.

This has important consequences. To facilitate our discussion, we denote by $\langle p_A, p_B, p_C\rangle$ the vector of choice probabilities. In the Darwin and Nash map, the states $A = \langle 1, 0, 0\rangle$, $B = \langle 0, 1, 0\rangle$, and $C = \langle 0, 0, 1\rangle$ are deliberational equilibria (meaning that if deliberation starts in such a state, it remains there forever). The states A and B are clearly unstable. At A the direction of improved causal expected utility points towards B. At B the direction of improved causal expected utility points to C. For this reason, A and B are not even equilibria in the best response dynamics. The state C, on the other hand, is a deliberational equilibrium in all three dynamics because, at C, the causal expected utilities of all three acts are the same. Hence, *at* C causal expected utility *does not* point away from C. This is what distinguishes C from A and B.

The state C is nonetheless *dynamically unstable*. Consider Figure 6, which shows the phase diagram of the Nash dynamics (the phase diagrams of the other dynamics are qualitatively the same). The straight line, L, starting at C and extending to a point close to B represents those states at which the acts C and B have the same causal expected utility. To the west of L, trajectories of the deliberational dynamics point away from C. More formally, consider *neighborhoods* N of C (i.e., open sets that include C) that are close to C. For any such N there exist deliberational

trajectories that leave N. Since the neighborhood N can be arbitrarily close to C, the equilibrium C is dynamically unstable.

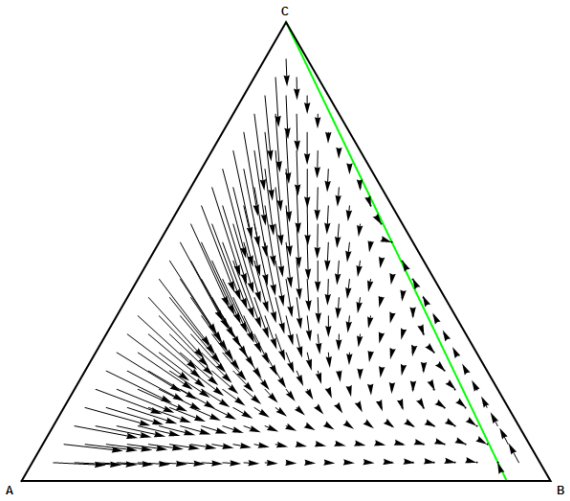

FIGURE 6. Phase diagram under Nash deliberation, k=10

This result is true for a large class of deliberational dynamics, as the following qualitative argument shows. Consider the causal expected utilities on the boundary of the deliberational space, shown in Figure 7. If the decision is only between Crate A and Crate B, then the direction of improved causal expected utility points toward B. If the decision is between Crate B and Crate C, improved causal expected utility is directed toward C. In both cases, this is because one act weakly dominates the other one. However, the same is true if the decision is only between Crate A and Crate C. In this case, the direction of improved causal expected utility points toward A because it weakly dominates C in the absence of B. Thus, along the boundary of the triangle, causal expected utility moves in a cycle. Now, every deliberational dynamics that seeks the good and is continuous must respect the direction of improved causal expected utility on the boundary. It follows that neither A, nor B, nor C can be stable under *any* such dynamics.

What happens to the east of line L? Here things are subtle. There is something special about state C. As pointed out above, at C acts A and B have the same causal expected utility as C does. This three-way utility-tie implies that the dynamics considered so far are *structurally unstable*. Structural stability must be distinguished carefully from dynamic stability. Dynamic stability concerns the stability properties of an eqilibrium (or some other region of state space). Dynamic stability says that an equilibrium is robust under small perturbations—the process equilibrates in case the system undergoes a small disturbance from equlibrium. A dynamically unstable equilibrium should thus not be expected to persist: there are always small perturbations, and an arbitrarily small one is enough to disrupt an unstable equilibrium.

Structural stability, by contrast, refers to the *entire* dynamical process. It says that small perturbations of the process itself have no significant effect on its qualitative properties: if the process is formalized in a slightly different way, the location

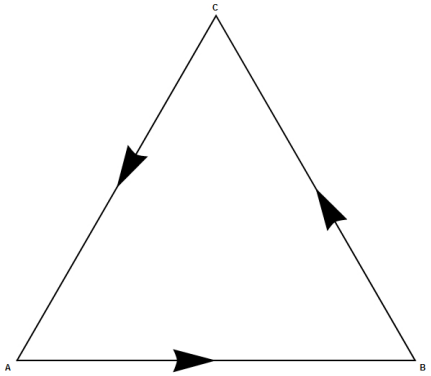

FIGURE 7. The dynamic on the boundary of the deliberational space

of equilibria (and similar characteristics) changes only slightly and their stability properties are preserved. A structurally stable process can thus be specified in slightly different ways without radically changing the original process. A structurally unstable process, then, fails to be robust in a very basic sense.

Structural stability is important because no dynamical model includes all potentially relevant factors. This is clearly true for dynamical models of physical processes. Any such model abstracts away from details of the actual physical situation. Structural stability indicates that those details really do not matter for certain qualitative properties of the dynamical model. Structural instability, however, indicates that the dynamical properties of the system are an artifact of the model.

What has just been said about models of physical processes also holds for deliberational dynamics. There are many ways in which one might model a process of rational deliberation, and none of them will capture everything that could go into such a model. Structural instability shows that small modifications lead to a radically different dynamics.[17]

Back to the dynamics of the Three Crates problem. The source of structural instability here is the utility-ties associated with Crate C. The dynamics considered thus far are purely payoff-driven; so, in the absence of payoff differences these dynamics are under-determined and, hence, will typically display structural instability.[18] One consequence of this is that it is difficult to determine what happens close to C when the process is to the east of line L. If the situation was structurally stable, the dynamical behavior around C would only depend on the linear terms associated with a deliberational dynamics. Structural instability entails that the dynamical behavior depends on higher-order (nonlinear) terms. These terms

[17]Structural stability can, of course, be defined more precisely. We limit ourselves to a restricted treatment for the Three Crates problem here and leave a more thorough analysis to future work. For mathematical details on structural stability see, e.g., Guggenheimer and Holmes (1983).

[18]More formally, the reason has to do with the eigenvalues of a deliberational equilibrium. If they take on specific values (for discrete-time dynamics 1, for continuous-time dynamics 0), then linear terms don't determine the stability properties of the equilibrium. Small changes to the dynamics typically change the eigenvalues, leading to discontinuous changes in the stability properties of an equilibrium.

are not easy to determine, making it difficult to say precisely what happens in the vicinity of C.

From the figures in this section there is some evidence, though, that there may indeed be convergence to C if the deliberational process is to the east of L. Let's grant this for the sake of argument; otherwise the models of this section don't provide any support for Hare and Hedden's point that causal decision theory recommends choosing C. How exactly does this fit in with Hare and Hedden's argument? One could read convergence to C as a vindication of their conclusion. Even if deliberation starts out to the west of L, and thus evolves away from C, the path followed by the deliberational process ultimately ends up at C.[19] Thus, one could say that ultimately causal deliberation ends up at C—even though C is dynamically unstable.

There is, however, a big caveat attached to what has just been said. Recall the objective of Hare and Hedden's main argument: to demonstrate that *causal decision theory recommends choosing act C in the Three Crates problem.* This conclusion only follows if it can be established that *any process of causal rational deliberation unequivocally converges to C.* The dynamic instability of C stands in the way of drawing this conclusion. Even if processes such as the Darwin map or the best response map ultimately converge to C, causal expected utility increases as we move away from C to the west of line L. This is certainly something a moderately sophisticated deliberator would realize; after all, the goal of rational deliberation is to seek the good—that is, to follow the lead of increasing causal expected utility. Consequently, convergence to C is not shared by *all* legitimate processes of causal rational deliberation. Thus one cannot conclude that causal decision theory recommends choosing C. Recommendations depend on the underlying deliberational dynamics.

This point is related to the central concern of structural stability. As explained above, structural instability implies that small perturbations result in radically different dynamical behavior. This means that different ways of evaluating utility ties—represented by distinct ways of perturbing the initial dynamics—lead to different judgments as to the relative attractiveness of acts. In the next section we discuss one such model.

4. Rational Deliberation II: Perturbations

Structural instability implies that arbitrarily small perturbations lead to qualitatively different dynamics. We don't wish to consider arbitrary perturbations, however. In order to be compelling, a perturbation should modify salient features of the models explored in the preceding section, such as the ideal predictor assumption or the assumption that the agent is capable of perfectly implementing her intention to choose an act.[20] In this section we will argue that including these

[19]There are some details one needs to consider in a careful analysis. For instance, for the Darwin map the process must start in the interior of state space.

[20]This feature leads to the idea of the *trembling hand*. Selten (1975) argues that "a satisfactory interpretation of equilibrium points in extensive games seems to require that the possibility of mistakes is not completely excluded" (p. 15). A rational agent can never be sure to implement her intentions perfectly. Her hand might "tremble", causing her to actually choose an act other than the one she intended to choose. Or there might be an external shock, like an earthquake, that interferes with her plan. In the Three Crates problem, this would mean that Hare and Hedden's

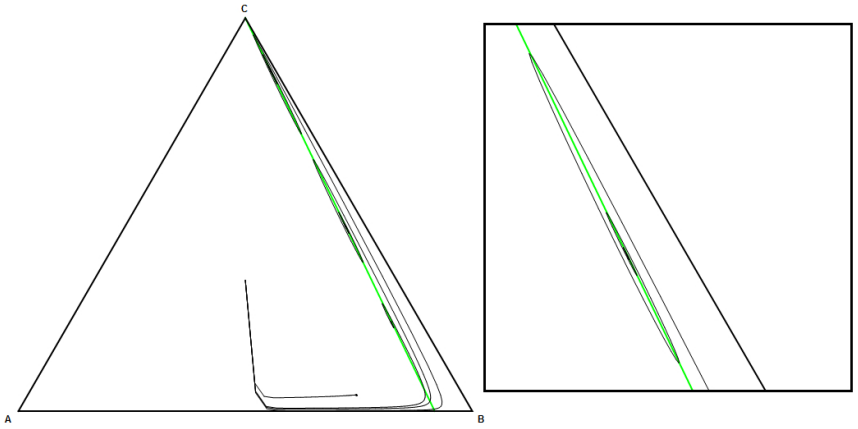

FIGURE 8. Replicator dynamics seeking the center, k=10, 50, 75, 300 (300 iterations) with magnified (3×) display of spiral behavior (k=50)

sources of uncertainty into the deliberational process also fails to yield an unequivocal recommendation in the Three Crates problem.

The following model is a perturbation of the Darwin map. The *Darwin map with point-seeking perturbation of index of caution* k is given by a triple $\langle \mathbf{p}, \mathbf{q}, k \rangle$, where $\mathbf{p}$ denotes the initial choice probabilities and $\mathbf{q}$ the locus of mutation. This map outputs a new state $\mathbf{p'}$, with each component p'_i given by

$$p'_i = \frac{kp_i \frac{U(A_i)}{U(SQ)} + q_i}{k+1}.$$

Thus, the Darwin map with point-seeking perturbation is a weighted combination of $\mathbf{q}$ and the Darwin map on $\mathbf{p}$.

If we view $\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \rangle$ as the state of maximal indecision, we can interpret the mutation with that locus as modeling *hesitation*. As k increases, the 'pull' toward the locus has lower influence, significant only when an act is assigned sufficiently low probability. Thus the Darwin map with point-seeking perturbation is a reasonable and tractable model for rational deliberation that involves uncertainty about the predictor's reliability or one's ability to implement decisions, since the important qualitative feature of these types of uncertainty is that they push deliberation away from extremal values.

As shown in Figure 8, the perturbed Darwin map alters the dynamics seen in the previous section significantly. Previously, there was an unstable rest point at Crate C (cf. Figure 3). Now there is an asymptotically stable equilibrum in the interior of state space, close to the edge between B and C. Instead of converging to

---

fourth assumption—that the agent will not choose a crate that she is certain she will not choose—can be easily violated by a rational decision maker. We don't think that being certain that one will be able to implement one's intention is a requirement of rationality. In fact, the opposite is true: believing with certainty that one will be able to implement one's intentions is highly implausible. Perturbations of the trembling-hands sort are not only compatible with the agent's epistemic and pragmatic rationality, they are required once we take into account the possibility of mistakes or external shocks.

one of the acts A, B or C, the deliberational process spirals around the equilibrium and reaches it in the limit. As a result, each act is assigned positive probability at the stable deliberational equilibrium.

We have explored other types of perturbations that push the dynamics away from extremal values, not just for the Darwin map, but also for the Nash and the best response map. While the resulting deliberational processes are not exactly the same, they are qualitatively similar to the Darwin map with point-seeking perturbation. In particular, there exists a dynamically stable deliberational equilibrium that assigns some positive probability to each of the acts. Furthermore, the deliberational process converges to that deliberational equilibrium from nearly all initial conditions (that is, convergence happens basically regardless of the agent's initial beliefs).

The reason for this robustness lies again in the payoff structure of the Three Crates problem. Recall that along the boundary of state space any deliberational dynamics that seeks the good moves from A to B, from B to C, and from C back to A (cf. Figure 7). Perturbing the deliberational process pushes the dynamics slightly away from the boundary; otherwise, however, the perturbed process respects the underlying payoff structure. In the Three Crates problem, this results in a cycling process that will often equilibrate at some interior state.

Convergence to an interior deliberational equilibrium is similar to a common type of decision instability found in decision situations such as Death in Damascus (Gibbard and Harper, 1978; Arntzenius, 2008; Joyce, 2012; Armendt, 2019). The conclusion we think one should draw here is that rational deliberation does not tell us which act should be chosen. In equilibrium, the choice probability of each act is positive, so choosing any act is admissible.

This shows that the possible convergence to C in the models of the preceding section provides no support for Hare and Hedden's conclusion that *causal decision theory wholly recommends choosing* C. Some causal models of rational deliberation might suggest choosing C; many other models that are at least as plausible do not converge to C. Instead, they converge to a state of indecision. As a result, there are many versions of causal decision theory that consider any pure act in the Three Crates problem to be admissible.

We don't wish to create the impression that the type of perturbation we have discussed in this section is the only reasonable way to perturb the models of the previous section. Structural instability allows for a wide variety of perturbations that lead to completely different dynamical behavior. Some perturbations may even sharpen convergence to C.[21] What we want to stress is that structural instability usually implies that there is no unique choice that causal decision theory recommends as the rational one. The recommendation depends on minute details of the agent's process of rational deliberation. Thus, in the Three Crates problem *causal decision theory does not recommend the choice of any particular act.*

5. Discussion

The problem for the causalist, according to Hare and Hedden, is not just that she ends up with nothing, but that she is guaranteed to end up with nothing because she is certain to choose C:

[21]This could, for instance, be achieved by perturbing the payoffs such that there is a small positive payoff for C in case the predictor predicted C.

> By taking Crate C, the self-aware causalist winds up with the princely sum of \$0. Worse, by her own lights, taking Crate C guarantees her a return of \$0. That is, given that she was certain she would take Crate C, she was certain that Crate C contained \$0. Of course, had she thought that she might take Crate B, then she would not have been certain that Crate C was empty. But she didn't think she might take Crate B, and so she was certain that her choice of Crate C would yield \$0.[22]

The problem pointed out here is that the causalist does not have a reason to choose C:

> Of course that is not to say that we cannot explain why she took Crate C—if she had taken any other box then she would not have been self-aware and rational. It is just to say that in explaining why she took Crate C we do not attribute to her any motivating reason to take Crate C. She did not take herself to have any reason to take Crate C, because, having convinced herself that the demon predicted she would take Crate C, she was certain that all the crates contained the same amount of money, and money, by hypothesis, is all she cared about.[23]

So the causal agent might as well have chosen A and B, since she is convinced that she gets nothing anyway. But the assumptions under which her deliberational process operates—epistemic rationality and self-awareness—supposedly show that she will choose C anyway. Hare and Hedden (2016, p. 617) conclude that "the claim that a practically and epistemically rational person will take Crate C in these circumstances is strongly counterintuitive and that this bears against the claim that causal decision theory is the correct theory of practical rationality."

The crucial step in this line of reasoning is the claim that a causal decision theorist will choose Crate C. Our analysis in the preceding sections shows that this is definitely not the case. The rational, self aware causal deliberator is *not* certain to choose C.[24] In the deliberational models of §3, the choice of C is an *unstable deliberational equilibrium*. Instability implies at least that the agent is not certain that she will choose C, the reason being that close to C there is relevant information about causal expected utilities which would lead the deliberational process away from C. In addition, the models are structurally unstable, meaning that there are many alternative legitimate ways of modeling rational deliberation that will show a radically different dynamical behavior. The models discussed in §4 show that deliberation often carries the agent to an equilibrium in which she does assign positive probability to choosing B (as well as A and C). Thus, the principles of causal decision theory do not imply the belief that crate C is empty.

---

[22]Hare and Hedden (2016, p. 617).

[23]Hare and Hedden (2016, p. 617).

[24]There is an issue here besides the stability of an equilibrium. One could also challenge this claim by questioning premise 4 from Hare-Hedden's original argument, namely that if a rational agent must pick an act that she is certain she will pick. If rational deliberation gets an agent to a belief state rather than an act selection, then why should arriving at full probability on a single act imply that an agent will choose that act? This inference involves a substantive step that is in need of justification, as illustrated by trembling hand considerations.

So the causal decision theorist does not choose C for certain. What act does she choose, then? While we don't have a comprehensive answer, we think it is plausible that rationality does not single out a unique act in decision situations such as the Three Crates problem. It depends on the details of an agent's deliberational process. The models of §4 show that even with a fully specified process, rational deliberation may lead to a state according to which any of the acts may be chosen. This does not mean that she will choose the *mixed act* that corresponds to the state of indecision.[25] The mixed act corresponding to the agent's state of indecision is a random device that chooses A, B and C with probabilities p_A, p_B and p_C. Such an act is not part of the models discussed here (where choice probabilities just represent an agent's beliefs). For the point we would like to make, this is enough; we don't need to use a mixed act interpretation of the agent's state of uncertainty. What is important is that for the causalist every act in the Three Crates problem can be rationalized.

This also sheds light on Hare and Hedden's remarks on how, in their view, the Three Crates problem gets around the usual weaknesses of *why aincha rich arguments*. Paraphrasing Hare and Hedden (2016, pp. 619–622): In cases like the Newcomb problem, an evidentialist could say that since she's richer the causalist cannot be so rational. The causalist may reply that this is only so because the problem is set up in such a way as to punish causalists. Hare and Hedden think that this results in a stalemate: the evidentialist and the causalist consider two different classes of cases in a decision problem as relevantly similar, and each agent is choosing optimally within her class of cases (more on that below). In the Three Crates problem, the relevantly similar cases for the evidentialist are those in which she chooses the predicted act. Since she chooses A and the causalist chooses C, the evidentialist always does better. The causalist considers the three predictions as the relevantly similar cases. For prediction A, the evidentialist gets a much higher payoff. For prediction C, both the causalist and evidentialist get the same payoff. The case of prediction B does not arise according to Hare and Hedden because "both causalists and evidentialists know that they are not in such a case."[26]

From our results on deliberational dynamics it follows again that this is not correct. Causalists typically do consider the state of the world in which the predictor predicted B to be a genuine possibility; the same is true for the other states. Thus the so-called stalemate between evidential and causal decision theory is not resolved by the Three Crates problem, as neither interlocutor finds herself in a situation where she truly fares worse in the cases she considers relevantly similar.

But is there a stalemate in the first place? We don't think so. Causal decision theorists deny that the relevantly similar cases of the evidentialist are the correct ones. The evidentialist compares those cases that correspond to the correlations in her subjective beliefs. But this is exactly what causal decision theorists think is fallacious about the evidential approach, namely that it confuses correlation with causation. In the Three Crates problem, the evidentialist should choose B because the predictor already has made her prediction of A (if indeed she has done so). There is no stalemate (at least not from the causalist's perspective) because the evidentialist compares the *wrong* cases.

[25]See Harper (1986) on mixed acts.
[26]Hare and Hedden (2016, p. 622).

6. Conclusion

We have analyzed the Three Crates problem in terms of a range of models of rational deliberation. In doing so, we have illustrated two types of instability, which should both factor into a causalist's deliberational process. An unstable deliberational equilibrium reveals that improved causal expected utility is ever near; and structural instability reveals the fragility of the qualitative properties of the decision situation. Alongside the internal problems of their argument, these factors undermine the conclusion of Hare and Hedden: a causalist is not committed to taking Crate C, and accordingly she is not doomed to a self-defeating conflict of priorities.

At a more general level, we have shown that structural stability is an important element in the theory of rational deliberation. A failure of structural stability often indicates that the underlying features of a decision situation are not sufficiently determinate to identify a single unequivocally rational choice.

Acknowledgements

We would like to thank Gerard Rothfus, Hannah Rubin, Brian Hedden, Jim Joyce, Brian Skyrms, and Brad Armendt, as well as two anonymous reviewers, for helpful comments on an earlier version of this paper. We would also like to thank the participants of the 2018 meeting of the Society for Exact Philosophy for their comments on the presentation of this article.

References

Armendt, B. (2019). Causal decision theory and decision instability. Forthcoming in *Journal of Philosophy*.

Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis*, 68:277–297.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, pages 861–898.

Gibbard, A. and Harper, W. (1978). Counterfactuals and two kinds of expected utility. In Hooker, A., Leach, J. J., and McClennen, E. F., editors, *Foundations and Applications of Decision Theory*, pages 125–162. D. Reidel.

Guggenheimer, J. and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer, New York.

Hare, C. and Hedden, B. (2016). Self-reinforcing and self-frustrating decisions. *Noûs*, 50(3):604–628.

Harper, W. D. (1986). Mixed strategies and ratifiability in causal decision theory. *Erkenntnis*, 24:25–36.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Joyce, J. M. (2012). Regret and instability in causal decision theory. *Synthese*, 187:123–145.

Katok, A. and Hasselblatt, B. (1995). Cambridge University Press.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

University of California, Irvine