Comments on Dupré, Dupuy, and Bender

David M. Kreps

THROUGHOUT THIS COMMENTARY, I maintain the linguistic convention that took hold during the conference: The humanists are labeled Earthlings; those who represent at least to some extent rational choice theory are labeled Martians. Stereotyping people in this fashion is difficult in some cases; some humanists subscribe (to some extent) to rational choice theory, and some participants who have no use for rational choice theory might not call themselves humanists. But for what I have to say, I think it will be clear what I intend when I use those terms.

CONCERNING DUPRÉ'S "RATIONAL CHOICE THEORY AND GENOMICS"

Professor Dupré's paper, "Rational Choice Theory and Genomics," is easily summarized from the perspective of rational choice. Professor Dupré doesn't like rational choice theory (RCT). He never has. He doesn't quite tell us why, except for a general aversion to ambitious general approaches to human nature, and a more specific unhappiness that RCT doesn't explain the origin of preferences. In any event, as he contemplates the sorts of very difficult choices that developments in genomics will offer individuals and societies, and he tries to think how RCT will help inform those choices, he sees no reason to change this opinion.

It goes without saying, but I'll say it anyway, that his description of what genomics does and will have to offer is extremely interesting and informative. But when I initially read his very informative and interesting description, I came to a conclusion opposite to his: The decisions genomics will force upon us as individuals and as a society will be very well illuminated, both descriptively and normatively, by rational choice theory. Of course, RCT will provide no final answers. But it will prove tremendously helpful in clarifying choices and consequences.

Or so I thought before the first day of this conference. Now I am much less certain, because, as we have learned here, the term "rational choice theory" means very different things to different people. I cannot say with precision what Dupré has in mind when he uses the term. But I

A

Kreps, David. "Comments on Dupré, Dupuy, and Bender." *Occasion: Interdisciplinary Studies in the Humanities* 1, no. 1 (October 15, 2009), http://occasion.stanford.edu/node/15.

know enough about Dupré's definition to know that, were I to adopt it, I would agree with him that RCT is pretty useless, if not dangerous.

But I do not agree with his definition. And I suspect that our disagreement arises from the use of the adjective "rational." If "rational" means conformance to some canonical set of axioms of behavior, such as those of Savage, Herstein and Milnor, or Anscombe and Aumann, then the empirical evidence is conclusive: Rational choice, so defined, is badly and systematically flawed as a descriptive model of human behavior.[1]

This doesn't mean that rational choice theory is useless as descriptive theory. Although Newton's laws of motion are badly flawed when applied to objects moving at relativistic speeds, they work quite well, as approximate laws of nature, for objects moving at down-to-earth speeds. Analogously, RCT, defined as conformance to the Savage, Herstein-Milnor, or Anscombe-Aumann axioms, has its domain of descriptive applicability, which the empirical evidence helps us identify.

RCT has a wider domain of normative applicability. But even here, the domain of applicability is narrower than most people think. For instance, if we consider gambles with objective probabilities and monetary prizes, but where the uncertainty involved resolves as different dates, these theories fail normatively. This is true even if the uncertainty resolves at a single date, if that single date is not "right now."[2]

But who mandated that conformance to Savage's postulates (or those of Herstein and Milnor, or Anscombe and Aumann) defines rational choice? When one says "rational choice theory," it sounds as if only one theory or model of choice could qualify. How could two distinct theories or models both be rational? But people behave in different ways, depending on the specific context and the more general social situation, and I see no reason to privilege one universal model of behavior with the adjective *rational.* To do so is, if not demonstrably silly, at least demonstrably misleading.

To give an example from subjects that I've studied: In providing extrinsic incentives to employees, one of the earliest results, due to Bengt Holmstrom, is that one should avoid tying rewards to random factors outside the control of the employee.[3] In some cases this is necessary, when there is noise interposed between employee action and observable consequences. But, to be very concrete, tying the compensation of an employee to the fortunes of the entire enterprise is, in most cases, the wrong thing to do.

This conclusion comes, of course, with underlying assumptions. Most importantly for current purposes, it assumes a particular type of behavior for the employee. The employee is a risk-averse expected-utility maximizer who is effort averse and whose utility depends only on his monetary reward and his level of effort.

---

[1] These are three classic, axiomatic treatments of choice under uncertainty: L. J. Savage, *The Foundations of Statistics* (New York: John Wiley and Sons, 1954; rev. and enlarged ed., New York: Dover, 1972); I. N. Herstein and J. Milnor, "An Axiomatic Approach to Measurable Utility," *Econometrica* 21 (1953): 291–97; F. J. Anscombe and R. J. Aumann, "A Definition of Subjective Probability," *Annals of Mathematical Statistics* 34 (1953): 199–205. Savage and Anscombe-Aumann deal with so-called subjective expected utility, where the decision maker's choices reveal both her preferences over prizes (utilities) and the likelihood she ascribes to various states of nature (subjective probability). Herstein-Milnor provide a powerful axiomatization of so-called von Neumann-Morgenstern expected utility, where probabilities are given objectively. A textbook presentation of these models is D. M. Kreps, *Notes on the Theory of Choice* (Boulder, CO: Westview Press, 1988), including a brief presentation, in chapter 14, of their major empirical flaws.

[2] See Kreps, *Notes on the Theory of Choice*, chap. 12.

[3] B. Holmstrom, "Moral Hazard and Observability," *Bell Journal of Economics* 10 (1979): 74–91.

As we look at this theoretical conclusion of RCT, we can see that it seems to work in some instances and fail in others. It seems a sensible conclusion in the context of, say, a company that is trying to motivate employees to work quickly at the task of installing replacement windshields in automobiles.[4] But it is much less sensible—indeed, very savvy practitioners will tell you it is flat out wrong—when it comes to motivating professional employees of high-tech start-ups. And, received wisdom, supported by less substantial data from the field, asserts that any form of extrinsic incentive will be counterproductive (not just nonproductive) when it comes to motivating scientists in a drug-development laboratory or professors at research universities.

Holmstrom's theorem, leading to an empirically testable theory about the impact of incentives, clearly doesn't work in all contexts. Since it arrives via a theorem, we know that the problem must lie in some of its assumptions; in this case, in assumptions about the behavior of workers. But it is not hard to repair the theory, by changing the model of human behavior, in the instance of high-tech start-ups to take into account endogenous processes of internalization of the welfare of others: in the instance of university professors, well-researched social-psychological processes of under- and over-sufficient justification. The details are unimportant here. What is important is to recognize that the behaviors of employees at high-tech start-ups and my university colleagues are not irrational; they just don't conform to the assumptions of behavior that most individuals (and, I suspect, Professor Dupré) associate with RCT. Indeed, their behavior is not only not irrational; it is based on social psychological processes that can have strong survival values in evolutionary models and that promote efficiency in a number of complex settings.

My point for now is simple. Human behavior is complex. By and large, it is purposeful; individuals have aims—some base, others noble—toward which they strive. Social scientists should not label one sort of behavior as rational and dismiss the rest, but instead identify circumstances in which various forms of behavior seem to predominate and, having done that, to test one's theory by deriving testable hypotheses about the consequences and testing those hypotheses. We (the Martians) should not speak of a single form of rational choice, but a multi-faceted set of purposeful choice and behavior models that, to avoid turning the exercise into a tautology, generate testable hypotheses. Those theories are likely to be more "local" in terms of social environment and context than some of the more doctrinaire Martians might like. The range of applicability, as a useful approximation, of any specific model of human behavior is unlikely to rival even the narrow range of applicability of Newton's laws in a universe of relativistic phenomena; we must accept that as a fact of life and move on. And Earthlings and other observers of Martian social science should understand that RCT does not mean a single, often unrealistic mode of behavior.

And what of genomics? Dupré has posed some pretty tough choices for us to think about, and I certainly have not invested the time or thought necessary to see what Martians might say about those choices, in terms of individual descriptive choice, individual normative choice, or the choices of social, political, and legal institutions in response to these new individual choices. Choices involving families and children are bound to be more fraught with social and social psychological factors than are choices of, say, whether to buy an apple or a banana for lunch. But if we think not of a single rational choice theory but a collection of models of purposeful choice behavior, I believe modeling individual choices as purposeful, but with a wider variety of purposes than you probably find in the average issue of the *American Economic Review*, then aggregating them to see their consequences in various institutional settings will be a useful and informative exercise.

---

[4] E. Lazear, "Performance Pay and Productivity," *American Economic Review* 90 (2000): 1346–61.

CONCERNING DUPUY'S ". . . ENLIGHTENED DOOMSAYING"

Let me say at the outset that I enjoyed Dupuy's paper very much. No, "enjoy" is not the right choice of word here; the apocalyptic visions with which he begins cannot be enjoyed. I suspect I'm more optimistic about the future than is he—perhaps this shows me to be a victim of one of the behavioral traits he laments—but regardless, "enjoy" doesn't seem quite right. So let me begin by saying that I found this a fascinating paper from a methodological point of view.

I'd like to make three fairly specific comments. Earthlings, beware: This is going to be an exchange between Martians and, therefore, a bit laden with jargon. And let me apologize in advance to Dupuy: I have not read the references to earlier work that he provides. He says very clearly that this paper is a "Classics Illustrated" version of a lot of deeper writing. So, I suspect my comments will tell him things he already knows.

The first of my three specific comments concerns Dupuy's … I don't know whether "disdain" or "dislike" is the appropriate word, so I'll use … dislike of Savage's axiomatic blending of the risk and uncertainty. As he doubtless knows, alternative models of choice can preserve the distinction. The class of empirical phenomena is known collectively as the Ellsberg paradox, in which individuals are more cautious regarding uncertainty than they are risk.[5] For the Earthlings, I summarize briefly: Ellsberg (who, before the Pentagon Papers, had a very fruitful life as a choice theorist) asks us to imagine an urn filled with balls colored red, blue, and green. He asked subjects to choose among gambles, where the payoff received by the subjects depended on what color ball was drawn from the urn; he told the subjects that precisely one-third of the balls were colored blue, while the other two-thirds were some unknown mixture of red and green. He observed—and this empirical result is as replicable as anything in social science—that many people preferred bets where their chance of winning is known to them—say $10 if the ball is blue and $0 otherwise—to bets where the chance of winning is ambiguous—$10 if the ball is green, or $10 if it is red. This sort of behavior violates one of Savage's postulates; it cannot be modeled by any choice model that conforms to Savage's vision of "rational choice." If we want to model the behavior of individuals who behave in Ellsbergian fashion, Savage's model is inadequate. But there are alternatives to Savage's model, such as those developed by David Schmeidler, that can accommodate this type of behavior.[6] (Parenthetically, I don't believe that these alternatives fully fit the empirical bill, because the full empirical bill would allow uncertainty-based caution to retreat hand in hand with uncertainty. One's caution about betting on red ought to disappear as balls are repeatedly drawn from the urn, and the decision maker learns the empirical frequency with which red occurs. At least, this is an empirical phenomenon I would expect; I don't know that anyone has run the experiment. But I do know that, if my empirical prediction is correct, existing alternatives to Savage are inadequate.)

That said, Dupuy's dislike seems to have deeper roots. A part of Savage's theory, buried more deeply than his explicit postulates, instead part and parcel of the mathematical context for his axioms, is the idea that the decision maker knows both the full set of states of the world and, for each action the decision maker might take, the consequences that will ensue. Epistemic uncertainty is rigidly limited in Savage's theory to uncertainty about what state will occur, but not what is the range of possible states or what is the consequence for the decision maker given his choice and the state. But some states and/or consequences are surely unforeseen, and my sense,

[5] D. Ellsberg, "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics* 75 (1961): 643–69.

[6] D. Schmeidler, "Subjective Probability and Expected Utility without Additivity," *Econometrica* 57 (1989): 571–87.

not grounded in any careful empirical investigations, is that Ellsberg-style caution is magnified for choices where the decision maker cannot foresee all the consequences of his choice or (formally somewhat equivalently) all the states that might prevail. Modeling this aspect of choice is at best an unfinished project for choice theorists; when it is done, I think it is likely to address some of Dupuy's dislike.

My second specific comment concerns the main point of Dupuy's paper, his metaphysics of projected time and his need for the epsilon e for some but not all futures. The issue may be hard for some to grasp at first, so I hope he will not mind if I give a caricature of his problem. A well-known problem of time travel in the literature of science fiction concerns the ability of someone traveling back in time to affect the conditions of her own existence. A time-traveler can certainly go back in time to help her mother and father meet for the first time or to regard each other romantically—their meeting and involvement confirms her existence and, therefore, her ability to go back in time, to arrange the relationship. But another time traveler, who goes back in time to shoot his father prior to the moment of the time traveler's conception, has a problem. If he kills his father, he wasn't conceived, so he couldn't have been around to travel in time and kill his father.

In roughly this fashion, Roger Guesnerie's characterization of the French planning system, as related by Dupuy, is like the time traveler who causes her parents to meet: The system creates conditions that confirm itself. But when Dupuy turns to managing prevention of catastrophes, he has the problem of the other time traveler. One must recognize that the catastrophe is coming, if one is to have the motivation to take precautions against it. But if those precautions prevent the catastrophe, we don't need to worry about it, and we are not motivated to prevent it. To deal with this, Dupuy inserts into his story prevention that isn't perfect; a small chance e the catastrophe will occur cannot be avoided. Then, if the consequences of the catastrophe are sufficiently large, prevention will be attempted and, one hopes, succeed.

A very close analog to this is found in the theory of dynamic games, in the problem of counter-theoretical actions. Suppose we are watching an individual who may have to make two choices in succession. At this moment, the individual must choose between A or B. And if the individual chooses A, he may at a later date have to choose between C and D. These choices have consequences for you, the observer, and for the decision maker. (It may help to give a graphical representation of this story. One is provided in Figure 1, where the individual who must choose between A and B is labeled RN for reasons that will become clear, and the other person in the interaction is labeled You. The numbers at the ends of sequences of moves are payoffs, the first number for RN and the second for You.)

Classic game theory tells us to analyze this situation using backward induction. Go to the end of the possible choices—to the point where the other person must choose between C and D—and ask whether C or D is better or, to use the loaded phrase, more rational for the other. This involves knowing his preferences, but suppose we know enough to suppose that D is the rational choice for him.

Now interpose between his two choices a choice for you, between X and Y. If you choose X, he must choose between C and D; if you choose Y, the interaction is over. And suppose the consequences for the two of you are arranged so that you do better with X followed by D than with Y, but Y is better for you than X followed by C. Since a paragraph ago we have you concluding that the rational thing for him to do is D, backward induction—reasoning step by step from the end of the game using RCT—tells us that you should pick X.

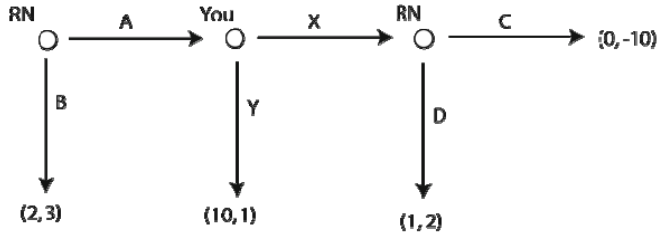

Figure 1. Is RN Crazy? RN must choose between A and B. If he chooses A, you must choose between X and Y, and if you choose X, he must choose between C and D. You both receive payoffs depending on where the game ends; his payoff is listed first at each endpoint. At his second choice, it seems rational for him to choose D, hence if you get a choice, X seems safe. But then he ought to choose B. If he chooses A, can you trust him to choose C, if you choose X? If not, maybe you should choose Y. But then A is rational for him, hence perhaps you can safely choose X, in which case...

But now go to the first stage of the interaction. Suppose that A-X-D for him is worse than B, but A-Y is best of all for him. Backward induction says he should choose B, because A from him elicits X from you, and then he is "stuck" with D. So much for analysis: You have used RCT to study the situation and predict he will choose B, ending the interaction. Then you go into the interaction and, counter to your theory, he chooses A. Given that he did something you didn't predict based on your model of how he will act, what should you now think of his final choice? And note, if you think that he might choose C at the end, because he did the "irrational" A at the start, and if you in consequence choose Y, then his choice of A is, according to his payoffs, entirely rational. So perhaps you are safe with X after all, which then makes A irrational or, at least, counter-theoretical, and not X is not safe. If X is safe, it is not; if it is not safe, then your theory of behavior suggests that it is.

Lest you think this is a whimsical construction, this is the story of Richard Nixon, where A is the "mad-man strategy," trying to induce enemies to give up the struggle by convincing them that A will be followed by the equally mad C. (If this modeling is correct, then the flaw in Nixon's choice of A is that either it didn't convince his enemies to choose Y or, what I think is more likely, he got their payoffs and purposes wrong.)

Is backward induction a useful tool of analysis? It certainly is a tool much practiced in the multi-person version of rational choice theory, noncooperative game theory. But in this situation, it seems to be less than useful and, perhaps, misleading. It only becomes useful if we have explanations for an initial choice of A by the other person that still allows you to reason, using RCT, what he will do if he does choose A. In some of the earliest work on this topic, Reinhard Selten told a story of a small chance that even a rational decision maker would tremble, choosing A when B is the rational thing to do.[7] Within this story, having seen A, you are comfortable with X; his ability to pose as a madman cannot convince you, since you have an alternate explanation.

This is just the bare beginnings of a literature that has, in its many variations, consumed huge amounts of effort among game theorists. I mention it now because of the slight formal similarity of Selten's trembling hand to Dupuy's imperfect prevention; perhaps some of the more recent developments in this literature, especially my colleague Yossi Feinberg's work on

[7] R. Selten, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 4 (1975): 25–55.

subjective epistemics, might give leads for other ways to resolve his metaphysical dilemma.[8] But I also mention it because it will return when I get to Professor Bender's paper.

My third comment on Dupuy concerns a behavioral observation that he makes almost in passing. Paraphrasing, faced with a possible but unrealized catastrophe, mankind seems unable to admit that it might happen and, therefore, take steps to prevent it. We know, for instance, of the theoretical possibility that terrorists will bring an atomic bomb into a city and detonate it. But the level of prevention of this nightmare does not seem commensurate with the scale of the tragedy that would result; we seem as if we prefer to deny the possibility and, to do that, we ignore it.

But then, the catastrophe happens. How then does our society react? Dupuy suggests that they still go about their business as best they can; the event that was previously too horrific to be taken seriously is now banal.

I don't know if this sort of dynamic behavior is part of the human condition, but if it is, it ought to be addressed. There are contingencies we cannot foresee or, in Dupuy's comment, we choose not to take seriously. Dupuy, at least, foresees that such contingencies will occur. Now it is possible within formal choice models to address ideas such as unforeseen contingencies and individuals who foresee that there are contingencies they don't foresee. I've written one such paper; there are others. These models do not operate quite in the manner of Dupuy, but my guess is that it wouldn't be hard to adapt the formal structures to his story. What is harder in the theory is to model a sequence of choices and events in which unforeseen consequences, or unimaginably horrible catastrophes, do occur, and then to see how such an event changes (or not) the level of precaution individuals take. Dupuy suggests that such an occurrence will not change the subsequent behavior of individuals and societies; I wonder if developing models of this and, within those models, exploring their consequences for societies, might not, in the style of one of his fixed points, change behavior, so we take greater precaution against the unimaginable. That would be a very interesting meta-theoretical development.

ON RATIONAL CHOICE THEORY AND LOVE

Professor Bender is clearly a very confirmed Earthling, as much an Earthling as I am a Martian. So his paper allows me to suggest why I hope for a better exchange of ideas between the two planets. I do not suggest that Martians, with their theories and models of choice, have anything useful to provide Earthlings. That's an assertion that should be made (or not) by Earthlings. But as an arch-Martian, I can and do propose that practitioners of RCT have a lot to learn from the study of planet Earth.

I have already discussed how a broader conception of choice models allows us to study things like the madman effect. I should mention here that especially powerful analyses come about when one combines (a) a situation in which an individual is taking a large sequence of actions and (b) a small probability in the minds of those who must deal with this individual that she might, maybe, be a little crazy. The form of this craziness is important; to be thought to be crazy in this fashion should confer advantages upon her in her series of actions and interactions with others. The craziness can have deterrent value, as in the story about the rich father or

[8] Y. Feinberg, "Subjective Reasoning—Dynamic Games," *Games and Economic Behavior* 52 (2005): 54–93.

Nixon; in other cases, it might promote joint interests; it might be a craziness to ignore one's own immediate interests for the greater social good, if others will and do respond in kind.[9]

These analyses confirm what we know from everyday experience: In such circumstances, she may "rationally" wish to act as if she is crazy, to keep the possibility of craziness alive in the minds of her companions. And because the crazy actions become rational, she derives the full benefit of them; her companions know that acting as if she is a bit crazy is in fact the rational thing to do and anticipate precisely this behavior.

This is all well worked out in the literature of game theory. Less well worked out are models of choice that permit the behavior Bender recounts from *Les Liaisons Dangereuses.* He wants us to choose between calculating, rational lovemaking and the more romantic form. I'd prefer models of behavior in which one can gradually and somewhat uncontrollably shift from the first into the second. In the sort of work I do concerning the theory of employment relationships, the ideas that (a) individuals can and sometimes do internalize the welfare of their coworkers both individually and collectively, and (b) the degree to which this happens is affected by the history and atmospherics of the ongoing relationship, are empirically quite important. Social psychologists know all about this stuff; economists lag to some extent. Economists' models of choice rarely capture the idea that one party can internalize the welfare of others—but "rarely" does not mean "never"—and even more rare are models in which the extent of this internalization—the weight one party puts on the welfare of others—is determined endogenously. (In this case, "even more rare" means "virtually never.") But the rarity springs from a lack of imagination about the phenomena rather than from any technical limitations in the types of choice behavior that can be modeled.[10]

Humanists, in my perhaps poor understanding, understand a lot about the human condition; how individuals act, where their passions and their reason lead them. If my Martian task is to build models of behavior that capture how individuals act, dropping the word "rational" and aiming for "purposeful," then I can learn a lot from what you of Earth have figured out about the human condition. I hope I have something in return to give you, but if there is going to be a balance of trade deficit on my side, so be it. A

[9] See, for instance, D. Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory* 27 (1982): 245–52.

[10] See D. Kreps, "Beliefs and Tastes: Confessions of an Economist," in *Models of a Man: Essays in Memory of Herbert A. Simon*, ed. M. Augier and J. March, 113–42 (Cambridge, MA: MIT Press, 2004).