Reply to Data Colada [124]
Ryan Oprea

I am grateful to Uri for giving me an opportunity to write a short, informal response to his post. I have also written a much longer and more detailed response to the comment that motivated the post (linked in the blog post), which the reader can take a look at to see some of these points expanded upon. In this short response, I want to emphasize four main points.

First, the post says that the AER paper rejects prospect theory and proposes that we replace it with "complexity." That's not quite right. Prospect theory has always been a "descriptive theory" and there has been a long-running ambivalence in the literature about what exactly it is the theory describes. What the AER paper is really doing is offering some evidence in favor of one of two classical interpretations (specifically in the context of valuation tasks). One is that prospect theory describes subjects' rationally expressed tastes for risk. The other is that it describes the consequences of a collection of cognitive shortcuts people use to value lotteries, because rationally expressing one's taste for risk is costly or difficult (the "complexity" interpretation). These are both orthodox interpretations of the theory, and both have a long history. The AER paper is providing some evidence in support of this latter classical interpretation – that complexity (the difficulty of valuation) and the shortcuts it inspires are major drivers of prospect theoretic behavior. The punchline of the paper is not that the patterns described by prospect theory are not real, but rather that they potentially extend beyond the domain of risk to other complex settings.

Second, the post (and the comment it builds on) argues that subjects in the AER experiment were confused, and in particular that they were "payoff confused" – that subjects in the mirror tasks thought they were actually in lotteries. Their main evidence for this is that subjects make errors in a series of training questions included in the experiment's instructions, and the post takes these errors as a measure of payoff confusion. The problem with this interpretation is that the point of these questions (as the paper says) is not to *measure* confusion, but to try to *train confusion away*. Subjects were given corrective feedback when they made mistakes and were forced to submit the right answer before moving on, with the goal of reinforcing their understanding of the payoff rule via this correction. And importantly there is very strong evidence that the questions worked as intended. Subjects make initial errors on these questions (particularly for the mirror task), but the rate of errors drops sharply from question to question and eventually falls to a low level that we have good reasons to believe just represents inattentive answering behavior (e.g., random clicking of answers by some subjects). Which means there are good reasons to believe that (in part because of the effectiveness of these training questions at doing what they were designed to do) subjects entered the experiment with far less confusion than the raw error rate suggests.

If the total errors in these questions (the measure discussed in the post) doesn't measure the confusion subjects retained in the experiment, what does it measure? Probably subjects' propensity for low effort behavior (e.g., low effort question-answering). These questions were tricky (particularly the mirror questions) and required cognitive effort to answer correctly the first time, but subjects were given no motivation at all to answer them correctly the first time. Which means many subjects likely had significant motivation to answer these questions randomly or inattentively at first and learn from the corrective feedback the software provided in order to learn how to think about the payoff rule. When the post shows that subjects who made more training question errors (most of which occurred in earlier questions, before subjects had had a chance to learn from feedback) make less consistent decisions (i.e., violate FOSD) what they are likely showing is not the effects of confusion but to a great extent the effects of the inattentive and noisy behavior that we should expect from relatively low effort subjects.

Third, even if we were to interpret subjects who made errors in these questions as confused, dropping them from the sample actually doesn't have much of an effect on the AER paper's main results. As I show in more detail in my formal response to the comment, even if we drop subjects who made any errors in these questions, subjects on average continue to show strong evidence of prospect theoretic behavior in mirrors. Mean valuations continue to be highly prospect-theoretic and standard statistical tests show that these departures from expected value continue to be highly statistically significant. The post emphasizes that when you drop these subjects, the effect seems to go away at the median but as I also show in my formal response, this statistic (and some of the others highlighted in the post and comment) doesn't do a very good job of representing the distribution of valuations as a whole: the distribution of deviations from expected value in both lotteries and mirrors remain highly skewed in the direction of prospect theory. What's more, subjects' overall tendencies to make prospect theoretic decisions in mirrors tend to be highly predictive of the same overall tendencies in lotteries, suggesting that the cognitive shortcuts subjects use in mirrors may also be used by them in lotteries too, at least to some degree. So, even if we remain concerned that some of the people who made errors in these questions remained confused upon entering the experiment, dropping them doesn't do much to the AER paper's primary conclusions.

Finally, towards the end, the post hypothesizes that a lot of what is driving these results is "measurement error." The idea is that people might be engaged in behaviors that aren't terribly rational but that bias valuations towards the center of the support of the lotteries they are valuing (including, for instance, making fairly random decisions). As the post says, this kind of behavior can produce the patterns of prospect theory. But this is just the kind of behavior that motivated the AER paper and was exactly the sort of behavior that mirrors were designed to measure. A long literature in and around prospect theory has speculated on a number of cognitive shortcuts -- simpler-than-optimal behaviors -- that are capable of producing prospect-theoretic patterns and that therefore might confound our efforts to measure people's taste for risk. From this perspective, this long literature has described a number of behaviors that produce "measurement error" relative to the typical goal of

these elicitation tasks, which is often to try to measure people's true valuations for lotteries (their real taste for risk). Because the whole idea of the AER paper is to use mirrors to measure this kind of behavior, what the post calls "measurement error" is in fact exactly what the AER paper's experiment is trying to measure.

When the post (and in more detail the comment) proposes to use mirrors as a way of debiasing lottery choices to recover "true" preferences, it fundamentally accepts the interpretation offered in the AER paper: that a lot of misbehavior that arises in mirrors likely also arises similarly in lotteries because both are difficult to value. By subtracting prospect-theoretic errors in mirrors from lotteries, the post proposes to get closer to a "true" measure of people's rational values for lotteries – a kind of debiasing achieved with the help of mirrors. As I discuss in my formal response, when you do this sort of differencing, what you find is that (even for the seemingly most sophisticated subjects), the prospect theoretic behavior that remains is very small relative to the behavior that appears in the original, uncorrected lottery valuations. In other words, mirrors reveal (at least if we accept the premises of this decomposition) that a lot of what we typically measure in these kinds of valuation tasks might not be people's true tastes for risk, but instead the shortcuts people use to cope with what the post agrees is a difficult task. Which means both the interpretation underlying the post's analysis and the conclusions one is compelled to draw from that analysis's results are very strongly in line with the main conclusions of the AER paper.