DE GRUYTER doi 10.1515/jci-2013-0027 — Journal of Causal Inference 2013; 1(2): 255–257

Causal, Casual and Curious

Judea Pearl*

The Curse of Free-Will and the Paradox of Inevitable Regret

Abstract: The paradox described below aims to clarify the principles by which empirical data are harnessed to guide decision making. It is motivated by the practical question of whether empirical assessments of the effect of treatment on the treated (ETT) can be useful for either policy evaluation or personal decisions.

Keywords: counterfactuals, treatment on the treated, regret, decision making

***Corresponding author: Judea Pearl,** Department of Computer Science, University of California – Los Angeles, Los Angeles, CA 90095-1596, USA, E-mail: judea@cs.ucla.edu

The experiment

A study was conducted to determine which of two schools, *A* or *B*, has a more effective educational program. A total of 200 randomly selected students underwent a randomized trial and were randomly assigned to the two schools, 100 to each. Another group of 200 (randomly selected) students were allowed to choose schools on their own; 100 selected *A* and 100 selected *B*. After a year of study, students were tested in a uniform, state run exam, and data showed the following:

100% of the *A*-choosing students failed the state exam
100% of the *B*-choosing students failed the state exam
50% of the *A*-randomized students failed the state exam
50% of the *B*-randomized students failed the state exam

It appears that, when given a choice, students tend to pick the school that is worse for them, which is strange but explainable. Suppose school *A* deemphasized math and *B* deemphasized history, while the state exam demands proficiency in both math and history. If students choose schools by the area of their strength then free choice amounts to a license to neglect one of the required subjects, namely a ticket to failure. Random assignment would force at least 50% of the students to study an area of weakness, which may explain the 50% success rate in the randomized groups.

But this is only one explanation; there could be others. The main point is that, somehow, students perception of how well each school would prepare them for the exam did not match reality; it was in fact uniformly misleading, as revealed by the randomized trial.

Thus far, we have been dealing with frequency data which, although strange, is not impossible and certainly not paradoxical. The paradox begins when we project the conclusion onto its counterfactual implications and how they are reflected in personal decision-making situations.

The paradox

Let us focus on an individual student named Joe, who just chose school *A* voluntarily, and ask ourselves what his chances are of passing the exam. The answer obviously is zero, based on data about the 100 *A*-choosing students, barring sampling variations and assuming of course that Joe is a typical student, perfectly exchangeable with the others who entered the database and chose school *A*.

Now let us ask ourselves a hypothetical question: What Joe chances would be had he chosen school *B* instead of *A*. Here comes a mini-surprise in a form of an unequivocal answer: 100%. The reasoning goes as follows: Examine the randomized data. Among those who were randomized into *B*, half were going to choose *B* anyhow and half were forced into *B* by randomization. The former half were bound to fail, as revealed by the data on the *B*-choosing group, so the latter group, the ones who were about to choose *A* and were made to take *B* by randomization, must have been the ones who passed the exam. The conclusion is that those who were about to choose *A* and were made to take *B* are guaranteed success (again, barring sampling variation).

Thus, despite the fact that the question we ask is purely hypothetical, "Would Joe be better off had he chosen *B*," it received a definitive affirmation from the data at hand. Assuming only the Joe, who never personally set foot at school *B* is similar to those subjects who were made to go to *B* by randomization.[1]

This transition from frequency data to a hypothetico-retrospective assertion derives its validity from a well-known theorem in counterfactual analysis, stating that, for *X* binary, the conditional probability of the counterfactual "*Y* would be true had *X* been different" can be consistently estimated from a combination of experimental and observational studies, regardless of the mechanism that gives rise to the data [1, 379].[2] But the paradox does not end here. Assume that the results of the data are made available to all students contemplating a choice of school. Each one of the students should now reason as follows: "I was about to choose school *A* but, given the data, I know I would be much better off choosing *B*. So, let me choose *B*. Alas, this puts me in the category of *B*-choosing students which, according to the data, would be much better off choosing *A*. And I am back where I started; doomed if I choose *A* and doomed if I choose *B*, I might as well flip a coin."

We are thus facing a situation where free-will becomes a curse, producing regret both ways, and the only way to escape its wrath is to surrender our will to the mercy of a random coin. Not a healthy scientific attitude.[3]

One way to resolve the paradox is to argue that once students are made aware of the data, their choices are no longer governed by the same perception as those who were counted in the data. For example, those who decided merely by comparing the school's program to their goals and personalities are not the same people as those who in addition to programmatic considerations also invoke the study data. They could not possibly be the same, so the argument goes, because that data reveal a major disparity between one's prior perception of school compatibility and the actual compatibility as tested in the exam.

This, I believe is a step toward a resolution, but it still suffers from one twist, what if Joe dismisses the data and says: "I am special, I know how to maneuver around exams." Moreover, what if all students have this attitude toward the data? Certainly this now makes Joe exchangeable with the other subjects in the study which, in turns, renders him subject to the fatalistic prediction of the data: doomed if he chooses *A* and doomed if he chooses *B*.

[1] Formally, this assumption is captured by a rule known as *consistency* [2] which is a theorem in both the structural and possible-world logics of counterfactuals [1, 3].

[2] This holds only for binary treatments. A general, graph-based analysis of the effect of treatment on the treated (ETT), including complete identification conditions under experimental and observational studies is given in Shpitser and Pearl [4]. See also the role of ETT in detecting latent heterogeneity [5].

[3] The benefit of randomization in this example differs fundamentally from the role it plays in game theory. There, a stochastic strategy becomes superior to any deterministic strategy because it guarantees a higher expected payoff against an adversary opponent; our adversary (Nature) is neutral and follows a pre-set stochastic strategy by which the exam scores are determined.

And here lies the solution. This fatalistic verdict is a predicament for Joe, true, but it is no longer a paradox of choice. It is now WE who are telling Joe of his inevitable sad fate (with certainty 100%), it is no longer Joe who is tormented by: doomed if I choose *A* and doomed if I choose *B*. Such a dilemma can only torment those who accept the invincible predictive power of the data, and then he is wrong, because by this very acceptance, he proves himself non-exchangeable with the other students in the study.

The two morals of the story are:

1. Once you are tormented you shouldn't be, but if you are not, you are doomed.
2. Torment can only save you from torment, not from harsh reality.

Bottom line

Question: Should Joe switch to school *B* once he was about to choose *A*?
Answer: Yes.
Question: Should Joe switch to school *B* after realizing how futile his choices are?
Answer: We do not know, because this very realization may make Joe special, and then the data does not apply to him. What we do know, and what Joe should seriously consider is whether his prior assessment of the school programs is valid in his case, data show him wrong.

Acknowledgment: This research was supported in parts by NSF grants #IIS-1302448 and #IIS-1249822 and ONR grant #N00014-10-1-0933 and #N00014-13-1-0153.

References

1. Pearl J. Causality: models, reasoning, and inference, 2nd ed. New York: Cambridge University Press, 2009.
2. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. Math Model 1986;7:1393–512.
3. Pearl J. On the consistency rule in causal inference: an axiom, definition, assumption, or a theorem? Epidemiology 2010;21:872–5.
4. Shpitser I, Pearl J. Effects of treatment on the treated: identification and generalization. In: Bilmes J, Ng A, editors. Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. Corvallis, OR: AUAI Press, 2009:514–21.
5. Pearl J. Detecting latent heterogeneity. Technical Report R-406. Department of Computer Science, University of California, Los Angeles, CA, 2012. Available at: http://ftp.cs.ucla.edu/pub/stat_ser/r406.pdf.