The Paradox of Inevitable Regret
(Extracted from an earlier draft of *The Book of Why* by Judea Pearl and Dana Mackenzie (2018))

To close a chapter devoted to causal paradoxes, I can't resist contributing one of my own. This one has a slightly different flavor, for it takes us on an adventurous climb from rung 2 to rung 3 of the Ladder of Causation and thus it illuminates the very foundation of personal decision making. I call it the Paradox of Inevitable Regret.

Suppose that a certain school district has two schools; school A provides strong instruction in math but poor instruction in history, while school B does the reverse. A random sample of two hundred students were selected for a randomized controlled trial, in which they would be assigned a school by a coin flip. The other students were allowed to choose their own school. At the end of the year, they took a state proficiency exam that covered both math and history, and these were the results:

- 100 percent of the A-choosing students failed the state exam.
- 100 percent of the B-choosing students failed the state exam.
- 50 percent of the students randomly assigned to A failed the state exam.
- 50 percent of the students randomly assigned to B failed the state exam.

One interpretation of these results is that students chose according to the area of their strength; the students who were good at math chose the school that was strong at math, and they were inadequately prepared for history. But I won't insist on this interpretation. The main point is that the students' perception of which school would prepare them better was uniformly mistaken. So far there is no paradox; we have only a slightly amusing story. The paradox occurs when we ask how students should use this information in making their personal decisions. First, let's ask what one student, Joe (who chose school A), should have done if a guidance counselor (like Monty Hall) had offered him the chance to change his mind. What would have been his chances of passing the state exam if he switched?

The answer may surprise you: in retrospect, his chances of passing would have been 100 percent. The reasoning goes as follows: Among those students who were randomized to go to school B, half were going to choose B anyway, and these were doomed to fail the state exams because the data tell us that 100 percent of the students who choose B and attend B fail. But we know that 50 percent of the randomized group actually passed! Therefore, the 50 percent who

passed must have been the other half of the students: the ones who were randomized to B but would have chosen A. So the passing rate was 100 percent among students who would have chosen A but were forced to attend school B. Assuming Joe is no different from them, so his chances of passing also would have been 100 percent, too. Thus, the factual data allow us to answer the counterfactual question, "Would Joe be better off if he had switched to B?"

This is a remarkable result. Not only because it assigns 100 percent certitude to a hypothetical outcome of one particular individual, but also because we were able to do it from data taken over a population of other individuals. A moment's reflection will hit us with an even more astonishing revelation: we have apparently answered a query from rung 3 of the Ladder of Causation, using only information gathered from rungs 1 and 2. This is not to be taken lightly; in fact you should even be suspicious, because I said in Chapter 1 that rung 3 is inaccessible from rungs 1 and 2. What enabled us to make this remarkable ascent, and can it also be used to equip a machine with a Homo sapiens-type imagination?

These questions are at the forefront of current research in machine learning, and I would only mention that this result is a consequence of a general theorem in counterfactual logic: if X is a binary (yes/no) variable, the conditional probability of the counterfactual "Would Y be true if X had been different¿' can always be estimated from a combination of experimental and observational data, which is what we have in this example. Evidently, it is the innocent restriction that X be binary, a restriction applied at the individual level, that allows us to climb to rung 3. My colleagues Elias Bareinboim and Andrew Forney have exploited this opening to rung 3 to great advantage in sequential decision problems. The idea is that a rational agent should not attempt to maximize the probability of success as estimated in the randomized trial, but rather the probability of success given the agent's current intention. This information—about Joe's own intention—is what gives Joe the ability to raise his chance of success from 50 percent to 100 percent.

Now let's continue with the paradox. Suppose that the results of the study are made available to all of the following year's students, including Joe's brother, Jack. He reasons as follows: "I was about to choose school A, but Judea tells me that I would be much better off choosing B.

Alas, this puts me in the category of B-choosing students who, according to the data, would be much better off choosing A. I'm doomed if I choose A and I'm doomed if I choose B! I might as well flip a coin." Alas, free will has become a curse! As long as Jack is ignorant of the underlying causal process, the message of the study is "Whichever school you choose voluntarily will be wrong, but if you pick a school at random you have a 50 percent chance of being right." So the best option, in his mind, is to abandon his free will to the tyranny of a coin flip. Not exactly a ringing endorsement of rational thinking!

Is there a way out of this paradox? How is it possible to be "doomed if you do, doomed if you don't"? The data show that for every student there is a school that will enable them to pass the exams, so nobody should be doomed.

One way to resolve the paradox is to argue that this year's students, including Jack, are not comparable to last year's students. Remember that a critical part of our argument was the assumption that Joe is no different from the other students who chose school A. Maybe that was true for Joe, but not for Jack: he knows the results of the study. Like the contestants on Monty Hall's game show, he can use that information to make a better choice. So our first conclusion is: People who have information and act on it are not doomed.

But what if Jack disregards the new information? What if, like young people everywhere, he says, "The data don't apply to me, because I know better. I know how to maneuver around exams." And what if every other student thinks the same way as Jack? Then, in fact, nothing has changed from the previous year, and the students are all doomed. So our second conclusion is: People who think they are not doomed are doomed. No doubt about it, we have paradoxes at every turn.

But in the meantime, what advice do we have for poor Jack?

I would suggest that he should approach the decision in two parts. First, he should banish the study from his head and figure out the choice he would make if he had never heard of the study. And then he should choose the other school. The data from the previous year prove that the choices the students make are always wrong, provided they are ignorant of the results of the study. The best thing that Jack can do is to mentally put himself in the position of those students. In other words, he has to answer his own counterfactual question: If I had never heard of this study, what school would I have decided on? If he can accomplish this feat (a big if, I know) then the data collected last year should apply to him. Alternatively, Jack could accept the empirical fact that students have demonstrated very poor competence at evaluating the merits of the schools, and leave the decision in somebody else's hands. But what teen-ager ever listened to a parent's advice?

It is also interesting to look at this paradox from the perspective of the school superintendent. Suppose that you have to advise students on how they should choose their school. Your objective is, of course, to maximize the number of students who pass the exam.

On the basis of the randomized study alone, the Fisherian tradition and every statistics textbook would tell you that the schools are equally meritorious and you would not have any better advice for the students than to flip a coin. But once you find out the results of the observational studies, your recommendation would change completely. Now, all students registered for School A would be encouraged to switch to school B, and vice versa. This strategy guarantees a 100 percent success rate for the community.

From the viewpoint of statistics textbooks, this result is itself a paradox. Our statistics students are taught to think of randomized trials as the gold standard, while observational studies are subject to confounding and other biases. The conventional wisdom, therefore, would be that an observational study has little to add to a randomized trial. Yet the school superintendent knows better! In this case, the combination of the randomized experiment and an observational study did better than either one in isolation. Sir Ronald Fisher would be rolling over in his grave!

Does the Food and Drug Administration know this? Can we use the combination of experiments and observations to improve the reliability of drug testing? Can we provide individualized guidance, based on counterfactual reasoning, to our patients? These are valid questions whose answers promise to revolutionize the art of experimental studies. They also hint at the practical value of counterfactual reasoning and the logic of counterfactuals, a subject we will take up in Chapter 8.