# Russian roulette: Why you can have a deterministic potential-outcome framework, or an asymmetric utility function, but not both*

Andrew Gelman† and Jonas Mikhaeil‡

16 Dec 2024

**Abstract**

It has been proposed in medical decision analysis to express the "first do no harm" principle as an asymmetric utility function in which the loss from killing a patient would count more than the gain from saving a life. Such a utility depends on unrealized potential outcomes, and we show how this yields a paradoxical decision recommendation in a simple hypothetical example involving games of Russian roulette. The problem is resolved if we allow the potential outcomes to be random variables. This leads us to conclude that, if you are interested in this sort of asymmetric utility function, you need to move to the stochastic potential outcome framework. We discuss the implications of the choice of parameterization in this setting.

## 1. Expressing the "first do no harm" principle as an asymmetric utility function on potential outcomes

Causal inference has traditionally focused on identification of average effects. In recent decades, there has been increasing interest in considering different forms of the average causal effect (Imbens and Angrist, 1994), flexible models for individual effects (Hill, 2011, Wager and Athey, 2018), and understanding how effects vary (Ding et al., 2019).

Individual causal effects can never be precisely estimated (without multiple measurements on the same units and strong assumptions about crossover effects), but they can be defined theoretically as latent variables using the potential-outcomes framework (Neyman, 1923, Rubin, 1974). In the simplest case of binary treatment $z$ and binary outcome $y$, the potential outcomes for unit $i$ are $(y_i^0, y_i^1)$, and these partition the population into four classes or principal strata (Frangakis and Rubin, 2002), which can be modeled as latent outcomes (Page et al., 2015).

Ben-Michael et al. (2024) and Christy and Kowalski (2024) have argued that, once effects are defined (even if only implicitly) at the individual level, it can make sense to consider asymmetric losses using some formalization of the "first do no harm" principle. Analyses using an asymmetric loss function have potential practical implications, because the average utility is no longer the same as the average causal effect.

Consider the case in which $y = 1$ is survival and $y = 0$ is death. Then the four principal strata correspond to $(y^0, y^1) = (1, 1)$ (patient survives under either treatment), $(0, 0)$ (patient dies under either treatment), $(0, 1)$ (treatment saves the patient's life), and $(1, 0)$ (treatment kills the person). In the first two principal strata, the treatment effect is zero. If the negative of the utility in the fourth stratum is taken to be greater than the positive utility of the third stratum, then the resulting decision rule will "prioritize safety over efficacy" (Christy and Kowalski, 2024).

There can be reasonable policy or ethical reasons to prefer low-risk treatments, but we have to be careful, as it is well known that any utility function that depends on unrealized outcomes can lead to incoherent decision recommendations; see, e.g., Bell (1982).

*We thank Amanda Kowalski for helpful discussions and the U.S. Office of Naval Research for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡Department of Statistics, Columbia University, New York.

For a very simple example, consider a decision between two options: A, a lottery which yields \$100 with probability 0.6 and \$0 with probability 0.4; and B, which with probability 0.5 yields \$100 and with probability 0.5 returns a lottery which yields \$100 with probability 0.2 and \$0 with probability 0.8. These two lotteries are identical in the distribution of outcomes and thus would receive identical utilities in the classical outcome-based utility framework of von Neumann and Morgenstern (1944). But now consider a utility function in which there is a negative value to uncertainty, which is evaluated at each branch of the tree. For example, suppose the utility of \$100 is set to 1, the utility of \$0 is set to 0, and the existence of uncertainty reduces the utility of any lottery by 10%. Then option A has utility $0.9\,(0.6 \cdot 1 + 0.4 \cdot 0) = 0.54$, while option B has utility $0.9\,(0.5 \cdot 1 + 0.5\,(0.9\,(0.2 \cdot 1 + 0.8 \cdot 0))) = 0.9\,(0.5 + 0.09) = 0.531$. Option B here has a lower utility because the utility function pays twice for the uncertainty. This is not to say that such a utility function is "wrong" (or, for that matter, that it is normatively or descriptively right); it just illustrates how it violates the axioms of classical (Neumann-Morgenstern) decision theory.

This example is *not* just a simple nonlinear utility function for gains and losses (which has its own coherence problems; see section 5 of Gelman, 1998); rather, it violates classical decision theory by depending on unrealized outcomes.

In general, when a decision rule is incoherent, it is possible to construct a scenario in which it yields inappropriate or nonsensical results. Again, that does not mean the rule is never appropriate. Working through simple examples can help us understand its domain of applicability.

## 2. Russian roulette

In the case of the "do no harm" principle, we can demonstrate incoherence using a simple example of two independent lotteries:

- $R_{1/6}$: You live with probability 5/6 and die with probability 1/6,
- $R_{1/7}$: You live with probability 6/7 and die with probability 1/7,

along with:

- $R_0$: You live.

If you have to choose between $R_{1/6}$ and $R_{1/7}$, the latter should clearly dominate. But, as we shall see, this does not always happen under a regret-based decision rule.

For a decision rule with potential outcomes $(y^0, y^1)$, we follow section 3.1 of Christy and Kowalski (2024) and assume the following relative utility function:

$$U(z = 1 \text{ relative to } z = 0 \,|\, y^0, y^1) = \begin{cases} 0 & \text{if } y^1 = y^0 \\ 0.5(y^1 - y^0) & \text{if } y^1 > y^0 \\ -(y^1 - y^0) & \text{if } y^0 > y^1, \end{cases} \tag{1}$$

so that the disutility of a loss is twice as large as the utility of a gain. For binary outcomes $y$, this rule corresponds to:

$$U(z = 1 \text{ relative to } z = 0 \,|\, y^0, y^1) = \begin{cases} 0 & \text{if you are sure to live} \\ 0 & \text{if you are sure to die} \\ +0.5 & \text{If the treatment would save your life} \\ -1 & \text{if the treatment were to kill you.} \end{cases} \tag{2}$$

First, we separately compare $R_{1/6}$ and $R_{1/7}$ to the status quo of $R_0$. Then we consider the scenario in which $R_{1/6}$ is the status quo and we are considering switching to $R_{1/7}$. That is where the counterintuitive result arises.

- $R_{1/6}$ *versus the status quo of* $R_0$. There is a 5/6 chance that $R_{1/6}$ has no effect and a 1/6 chance that $R_{1/6}$ kills. In the formulation of Christy and Kowalski (2024), the potential outcomes $(y^0, y^1)$ are $(1, 1)$ for 5/6 of a (large) population and $(1, 0)$ for 1/6 of the population. The expected relative utility is $-1/6$.

- $R_{1/7}$ *versus the status quo of* $R_0$. Similarly, here the expected relative utility is $-1/7$. Unsurprisingly, both $R_{1/6}$ and $R_{1/7}$ have negative utilities compared to $R_0$, with $R_{1/6}$ looking worse.

- $R_{1/7}$ *versus the status quo of* $R_{1/6}$. We have assumed the outcomes in $R_{1/7}$ and $R_{1/6}$ are independent, so we can easily work out the probabilities of the four potential strata: $(y^0, y^1)$ takes on the value $(1, 1)$ with probability 30/42, it takes on the value $(0, 0)$ with probability 1/42, it is $(1, 0)$ with probability 6/42, and it is $(0, 1)$ with probability 5/42. Under expression (2), the expected relative utility of $R_{1/7}$ compared to $R_{1/6}$ is thus $0.5(6/42) - 5/42 = -1/21$.

The result—that, in a head-to-head matchup, we should prefer to stick with $R_{1/6}$—makes no sense. Obviously we would prefer our chance of dying to decrease from 1/6 to 1/7. But, under the asymmetric utility function, you lose more from the 5/42 chance of the new treatment "killing you" than from the 6/42 chance of it "saving your life."

## 3. Moving from deterministic to stochastic potential outcomes

There are two natural ways to resolve this problem and avoid the nonsensical decision recommendation. The first approach would be to move to a classical utility function that depends only on outcomes, in which case one can just evaluate $U(R_{1/6})$, $U(R_{1/7})$, and $U(R_0)$ directly, and there is no incoherence. This solution is too easy, though, in that it eliminates the "do no harm" principle that motivated the idea in the first place.

A second approach is to preserve the asymmetric utility function but change the nature of the potential outcomes, replacing the *deterministic* potential outcomes of Neyman (1923) and Rubin (1974) with *stochastic* potential outcomes as in Greenland (1987) and Vanderweele and Robins (2012). In the Russian roulette example, this entails replacing the binary outcome $y^z$ with a binary random variable whose expectation is the probability of living under treatment $z$. We then generalize the utility function in Equation 1 to allow stochastic potential outcomes:

$$U(z = 1 \text{ relative to } z = 0 \,|\, y^0, y^1) = \begin{cases} 0 & \text{if } U(y^1) = U(y^0) \\ 0.5(U(y^1) - U(y^0)) & \text{if } U(y^1) > U(y^0) \\ -(U(y^1) - U(y^0)) & \text{if } U(y^0) > U(y^1). \end{cases} \tag{3}$$

In the problem we are considering, the $y^z$'s are binary random variables, and we set $U(y^z) = E(y^z)$. What is relevant in (3) is that the utilities of $y^0$ and $y^1$ are evaluated individually, and then the asymmetric utility for the decision is applied to these separate utilities. This follows the same principles as (1), with the only change being that the potential outcomes $y^0$ and $y^1$ are random variables.

Here are the three decision problems under probabilistic potential outcomes:

- $R_{1/6}$ *versus the status quo of* $R_0$. Here, $U(y^1) = 5/6$ and $U(y^0) = 1$ for everyone in the population, and so the decision will certainly reduce your probability of survival. The expected relative utility is $-1/6$.

- $R_{1/7}$ *versus the status quo of* $R_0$. Similarly, here the expected relative utility is $-1/7$.

- $R_{1/7}$ *versus the status quo of* $R_{1/6}$. Here, $U(y^1) = 6/7$ and $U(y^0) = 1/6$ for everyone in the population, so the decision will certainly *increase* your probability of survival. The expected relative utility is $0.5(6/7 - 5/6) = 1/84$, which is positive, as it should be.

Paradoxically, when we use stochastic potential outcomes in this example, the utility functions become deterministic. In general, this will not necessarily be the case. For example, consider a hypothetical medical condition where the survival rate is evenly distributed between 20% and 50% (depending current health status), and there is a treatment that only works on 20% of the population, but when it does work, it halves the death rate. In that case, one can set up a mixture model for potential outcomes in the population, where $y^0$ is uniformly distributed in the range $(0.2, 0.5)$ and $y^1 = y^0$ for 80% of the population and $y^1 = 1 - (1 - y^0)/2$ for the remaining 20%. Such a model could be elaborated by allowing the mixture probability to depend on $y^0$ and by letting the entire distribution depend on patient-level variables $x$.

## 4. Stochastic potential outcomes and the stable unit treatment value assumption

Stochastic models for potential outcomes require a generalization of SUTVA (the stable unit treatment value assumption; Rubin, 1980). The issue here arises not with the first assumption of SUTVA, the familiar rule of no interference within units, but rather with the lesser-known second assumption, which Rubin (2005) expresses as "there are no hidden versions of treatments; no matter how unit $i$ received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0." In the Russian roulette example there are no "hidden versions" of treatments in the usual sense of the phrase—you spin the cylinder and fire the shot; it's the same treatment every time—but the potential outcomes are not deterministic.

That is fine. SUTVA is not a requirement for causal inference; it is just a particular set of assumptions, appropriate in some settings but not for others. Just as we lift the no-interference-within-units assumption when modeling spillover effects, we can lift the second assumption of SUTVA when considering treatments that yield stochastic potential outcomes. There are many problems in biomedical and social research where outcomes are inherently noisy, so that such stochastic causal models make sense, even for treatments that are applied the same way to each unit.

Stochastic potential outcomes is not a new idea—Neyman et al. (1935) refer to them as "technical errors." In classical decision theory, they can be thought of as unobserved latent variables, so that any stochastic potential outcome model can be rewritten as a distribution over deterministic values, in the same way that logistic regression can be viewed as an integration over logistically-distributed latent variables. Vanderweele and Robins (2012) explain how a stochastic potential outcomes model can facilitate a convenient framework for modeling multiple causes.

This modeling choice is reminiscent of the distinction in classical statistical theory between parameters, which are conditioned on when evaluating inferential methods, and latent data or predictions, which are averaged over. For example, classical statistics distinguishes between unbiased *estimation* of parameters $\theta$ (the condition $\mathrm{E}(\hat{\theta}(y)|\theta) = \theta$) and unbiased *prediction* of latent data $u$ (the condition $\mathrm{E}(\hat{u}(y)|\theta) = \mathrm{E}(u|\theta)$). In Bayesian inference, there is no such distinction—all parameters are considered to be a form of latent data and are given probability distributions. This is consistent with the property of the Neumann-Morgenstern framework (which, confusingly, we call classical decision theory although it is fully Bayesian) that there is no fundamental distinction between deterministic and stochastic potential outcomes.

Even setting aside the problem of utilities that depend on unrealized potential outcomes, we have found stochastic potential outcomes to be useful as a way of partitioning into unit-specific

and unexplained variation (Gelman, 2021). More formally, one can imagine hypothetical multiple applications of the treatment to the same unit: in the Russian roulette example, different trials could yield different results on the same person, whereas if the potential outcome depends entirely on person-level characteristics, it would be the same every time. As discussed in the next section, the partition of variation depends on the extent to which the potential outcomes depend on immutable (if unobserved) unit-level characteristics, which in turn depends on the application.

## 5. Connecting the mathematical and substantive models

For the Russian roulette example, the problem with the deterministic potential outcome model is that it assigns the ultimate outcome under $R_p$ as a property of the individual rather than as a product of inherent randomness in the treatment. To make this clearer, consider the Russian roulette treatment as including the spinning of the cylinder as well as the firing of the shot: the assumption is that the outcome of the spin is random and is not predictable by any characteristics of the individual. To justify the probabilistic model, one could imagine the spinning to be performed by a machine under the control of an external random number generator.

The deterministic and stochastic potential outcome formulations imply the same joint population distribution for $(y^0, y^1)$ but correspond to different models of the individual $(y_i^0, y_i^1)$'s. When working with classical Neumann-Morgenstern utility theory, these differences have no decision implications because uncertainty about all potential outcomes is integrated out, but when considering asymmetric utilities that depend on unrealized potential outcomes, these differences can affect the decision, as we have seen for the example of comparing $R_{1/6}$ to $R_{1/7}$.

Another way to see this is to consider an analogous trolley problem, in which a runaway trolley is heading toward a crowd and, if nothing is done, it is expected to kill 1/6 of the people there, but there is an option to switch it to an alternative track leading to a less densely-populated zone where it would be expected to kill only 1/7 of the people in the crowd. Under classical utility theory, switching is an unambiguously dominant strategy. But under the asymmetric utility function (1), if we define deterministic potential outcomes—that is, if we condition on the set of people who would die under the different alternatives—it would be better not to switch.

Or, to morph this into a medical problem, suppose that for a certain otherwise-fatal snakebite the current treatment is an antidote which works perfectly except for 1/6 of the population who have a certain genetic condition. An alternative antidote is proposed which works perfectly except for 1/7 of the population who have a different genetic condition whose occurrence is statistically independent of the other condition in the population. Further suppose that only one antidote can be given to all patients—or, equivalently, that we are treating one patient whose genetic conditions are unknown, and only one antidote can be tried. The decision of whether to switch to the second treatment is equivalent in expected value to the earlier-considered decision to switch from $R_{1/6}$ to $R_{1/7}$, but it corresponds to a different distribution of principal strata and thus a different decision recommendation when using the asymmetric utility function that depends on unrealized outcomes.

When mapping this utility function to a decision recommendation, it matters where the conditioning is done. We see an analogy to the Heisenberg principle in physics, under which outcomes can depend on the order of measurement. In classical probability (in physics jargon, "Boltzmann statistics"), unmeasured data can be expressed as part of a joint distribution, and the order in which information is revealed does not affect observables. In quantum probability, indeterminacy manifests as superposition of wave functions, not averaging over probability, which is why we see the characteristic interference pattern of a two-slit experiment, for example.

The argument, then, is that if you want to use this sort of an asymmetric utility function, you cannot simply integrate out the uncertainty of stochastic potential outcomes ahead of time. The

difference between the Russian roulette problem and the snakebite problem is that, in the latter, the potential outcomes are latent variables that are deterministic attributes of the individuals.

How would this apply in real-world outcomes? We consider two areas: medicine and business.

- In a cleanly-specified medical example in which the outcome is binary (survival or death), when would it make sense to prefer to stick with a treatment that leaves more people dead? There is a vast literature on the trolley problem, and the resolution often turns on issues of agency. One rationale for preferring the status quo, even if the new treatment increases the expected number of lives saved, is that this sort of "social engineering" would reduce people's control over their own lives. In the Russian roulette example, this reasoning would not apply because the decision to switch from $R_{1/6}$ to $R_{1/7}$ does not harm any identifiable patient.

- When making a business decision—for which the negative outcome for an individual is not death but rather is to no longer be a customer—there are various reasons to prefer the status quo even if it is stochastically dominated by an alternative treatment. Making changes ("churn") has its own costs, and one can easily imagine a scenario in which a loss of 1/7 of a customer base, even if accompanied by a gain of 1/6, would represent a net loss. This would be the trolley or snakebite scenario. In the Russian roulette scenario, the decision to switch from $R_{1/6}$ to $R_{1/7}$ would simply result in a reducing the loss by 1/42, which is unambiguously better, as there are no particular customers who could be identified as being lost under one treatment or the other.

To push this further, consider a setting where the potential outcomes are mathematically deterministic but effectively random. For example, suppose that for some reason a company is about to lose all customers whose Social Security numbers are divisible by 6, and an intervention is proposed by which the company will instead lose all customers whose Social Security numbers are divisible by 7. Here the potential outcome is deterministic and we are in the trolley or snakebite scenario, but for all practical purposes this seems more like Russian roulette, in that there are no relevant distinguishing characteristics of people in the four principal strata defined by the deterministic outcomes $(y^0, y^1)$. In this case it would seem to make sense to integrate over this uncertainty, that is, to model the potential outcomes as stochastic. What this last twist on our example demonstrates is that the appropriate parameterization ultimately depends not just on the asymmetric utility function but also on its underlying motivation. In real-world settings, we would expect the units in the principal strata to differ in relevant ways, and we expect it would make sense to employ a stochastic potential-outcome model where the distribution for $(y^0, y^1)$ depends on observed and latent characteristics of the experimental units.

For another example, imagine there are two kinds of patients presenting with headaches. For one type of patient, the headache is caused by a migraine and is not responsive to treatment; for the other the headaches do respond to treatment with some probability of success. The potential outcome is deterministic for the migraine patients but, depending on how much is known about the mechanism of the treatment, could be stochastic for the non-migraine patients. This parameterization of potential outcomes would make sense even if individual patients' migraine status is unknown, because the possible action of the treatment depends on characteristics of the individuals. When comparing to a new treatment that works for some migraine patients but at the cost of reduced efficacy for others, one might want to use an asymmetric utility function that depends on unrealized potential outcomes, which in this case would map to migraine status.

## 6. Discussion

There is a large literature on decision rules and utility functions that go beyond classical decision theory. There could be a general conservatism or disutility of making changes, just because churn has its own costs—or, in other settings, novelty could be considered to having an inherent benefit as a counter to status quo bias. Preferences can also be based on agency: people can be more comfortable with risks that they feel they have the power to control and can resist an apparently clear improvement if it represents a perceived loss of power or status. And meta-decision issues come up, ideas such as satisficing (Simon, 1958), Type II rationality (Good, 1971), and fast and frugal heuristics (Gigerenzer and Goldstein, 1996) that recognize the cost of gathering and evaluating the information required to make theoretically optimal decisions. There is an expression in poker, "Evaluate the strategy, not the play," and one might choose relatively inefficient decision heuristics as part of a global decision-making strategy that is cheaper (less costly to evaluate, exerts a lower cognitive or moral load, etc.). One could imagine a "First, do no harm" principle being morally and socially valuable, to the extent that it could be worth making some objectively bad decisions so as to preserve a larger principle of consistently, in the same way that we like to say we live in "a government of laws, not of men," and so there will be unjust decisions that still need to be enforced in order to not lose that consistency.

The relevance of these ideas to the present paper is that, once we go beyond classical (Neumann-Morgenstern) decision theory and allow the utility of a decision to depend on something other than its outcomes (for example, life or death in the simple medical or Russian roulette example), stochastic potential outcomes can no longer be thought of as equivalent to population averages of discrete potential outcomes. This is a good thing, in that it allows us to apply the asymmetric utility function of Ben-Michael et al. (2024) and Christy and Kowalski (2024) to the trolley or snakebite scenario without having to swallow the uncomfortable conclusion that it is better not to switch from $R_{1/6}$ to $R_{/17}$ in the Russian roulette example. That is, we can attack those those different problems with the same utility theory, just expressing their potential outcomes in a different way.

When working with classical (Neumann-Morgenstern) decision theory in which the utility depends only on realized outcomes, there is no fundamental difference between stochastic and deterministic potential-outcome models: the stochastic model can be transformed into a deterministic model by simply interpreting the joint distribution of $(y^0, y^1)$ as representing a superpopulation distribution from which the deterministic outcomes $(y_i^0, y_i^1)$ are sampled. It is just a matter of convenience which model is easier to interpret in any given problem (Greenland, 1987).

Once we go beyond classical decision theory and allow utilities that depend on counterfactuals, though, the framing makes a difference. For decision analysis, going from stochastic to deterministic models for the potential outcomes has the effect of reordering of the steps of expressing uncertainty and evaluating the utility, and outside the Neumann-Morgenstern framework, these steps do not commute.

The Russian roulette problem is a simple example demonstrating how the relative utility of different decisions—even the ranking of which decision is preferred—can depend on how the potential outcomes are parameterized. Although this might seem awkward, it is a mathematical consequence of the use of a utility function that depends on unrealized outcomes.

To the extent that such decision rules are desirable, it behooves us to think of this dependence on parameterization as a feature rather than a bug. A principle such as "first do no harm" depends on what is considered "harm," which in turn depends on what aspects of a problem are considered to be potentially under control. By using potential outcomes that have stochastic components, we are making such decisions explicit.

## References


Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research* **30**, 961–981.

Ben-Michael, E., Imai, K., and Jiang, Z. (2024). Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*.

Christy, N., and Kowalski, A. E. (2024). Starting small: Prioritizing safety over efficacy in randomized experiments using the exact finite sample likelihood. `https://arxiv.org/abs/2407.18206`

Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association* **114**, 304–317.

Gelman, A. (1998). Some class-participation demonstrations for decision theory and Bayesian statistics. *American Statistician* **52**, 167–174.

Gelman, A. (2021). A counterexample to the potential-outcomes model for causal inference. Statistical Modeling, Causal Inference, and Social Science blog, 26 Jul. `https://statmodeling.stat.columbia.edu/2021/07/26/causal-counterexample/`

Frangakis, C., and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Gigerenzer, G., and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* **103**, 650–669.

Good, I. J. (1971). Twenty-seven principles of rationality. In *Foundations of Statistical Inference*, ed. V. P. Godambe and D. A. Sprott, 108–141. Holt, Rinehart, Winston.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* **125**, 761–768.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240.

Imbens, G. W., and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 463–480 (1990).

Neyman, J., Iwaszkiewicz, K., and Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society* **2**, 107–180.

Page, L., Feller, A., Grindal, T., Miratrix, L., and Somers, M. A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation* **36**, 514–531.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1980). Comment on D. Basu, "Randomization analysis of experimental data: The Fisher randomization test." *Journal of the American Statistical Association* **75**, 591–593.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100**, 322–331.

Simon, H. A. (1958). Rational choice and the structure of the environment. In *Models of Bounded Rationality* **2**, 259–268. MIT Press.

Vanderweele, T. J., and Robins, J. M. (2012). Stochastic counterfactuals and stochastic sufficient causes. *Statistica Sinica* **22**, 379–392.

von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Wager, S., and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242.