

The [Von Neumann-Morgenstern \(VNM\) theorem](#) states that any agent with rational preferences between probabilistic “lotteries” of outcomes must have a “utility” number for each possible outcome, and maximize the expected value of that number, or “expected utility.”

For example, suppose you’re choosing between a house, a car, and a television. Lotteries then look like “50% chance of a television, 10% chance of a car, 40% chance of a house,” or “10% chance of a television, 80% chance of a car, 10% chance of a house,” or anything else that adds to 100%.¹

The theorem assumes you have preferences between every possible pair of such lotteries, and that these preferences obey some axioms that you can think of as rationality requirements:

- “Completeness” says you always either prefer one option to the other, or are indifferent. You can’t consider the certainty of a car to be simply incomparable to a 50/50 coin flip between a television and a house. It’s either better, or worse, or exactly as good.
- “Transitivity” says if you prefer lottery A to lottery B, and you prefer B to C, then you prefer A to C: if a house is better than a car, and a car is better than a television, then a house must be better than a television.
- “Continuity” says if you prefer A to B and B to C, then there must be some lottery mixing the good option (A) and the bad option (C) that’s as good as the medium option (B). Even if some of your preferences are much stronger than others, like because you’d love a house and you don’t much prefer a car to a television, there’s still some mix — e.g., 0.01% chance of house and 99.99% chance of television — that’s no better or worse than a 100% chance of car. The axiom doesn’t allow a house to be “infinitely better” in the sense that adding *any* nonzero chance of a house (no matter how small) to the television deal would make it preferable to the 100% car deal. That would create a “discontinuity”: your preference at zero chance of a house would differ from your preference arbitrarily close to zero.
- Finally, “independence” says preferences don’t change as a result of mixing in some chance of a third outcome: if you prefer a car to a television, then you must also prefer a car/house coin flip to a television/house coin flip. Otherwise, if you saw the coin didn’t land on “house,” you’d be left preferring television to car, even though you prefer car to television outright. Your preference would depend on this event that could have happened but didn’t.

Given all these assumptions, the theorem says that there must be a function, called the “utility function,” that assigns a number, called “utility,” to each outcome, whose expected value always corresponds to which lottery you prefer. If the television, the car, and the house have utilities 10, 20, and 30, then the lottery “50% television, 10% car, 40% house” has expected utility $0.5 \cdot 10 + 0.1 \cdot 20 + 0.4 \cdot 30 = 19$, and the lottery “10% television, 80% car, 10% house” has expected utility $0.1 \cdot 10 + 0.8 \cdot 20 + 0.1 \cdot 30 = 20$; the second one is better because it has better odds of getting the

¹ To be clear, these are probabilistic mixes of possible futures; in each future, you get only one thing. If they were baskets of goods, it would be unreasonable to assume independence: maybe you’d rather have an apple than a right shoe, but you’d also rather have a left and a right shoe than a left shoe and an apple.

car. But if they have utilities 10, 20, and 40, then the expected utilities of these lotteries are 23 and 21; the first one is better because it has better odds of getting the house. A preference for the first lottery is consistent with the first utility function but not the second one. In this way, each preference helps narrow down the range of possible utility functions. The theorem says that there exist utilities consistent with *all* your preferences — that is, your preference is to maximize the expected value of this utility. If not, you'll have to bite one of the bullets in the bullet list above.

There's some limited freedom in choosing the utilities: you can add the same number to all utilities, or multiply all utilities by the same factor, and it will still fit the same preferences. But for every two pairs of outcomes, there will be a fixed ratio between their differences in utility. If the difference between the car and the house matters ten times as much as the difference between the television and the car, then it always matters ten times as much. Imagine your future is split into many different branches, each with the same probability. Then you'll consistently value "changing ten branches where you get the television into branches where you get the car" the same as "changing one branch where you get the car to the branch where you get the house." If an action upgrades nine television branches into nine car branches at the cost of downgrading one house branch into a car branch, you won't take that action. But if an action upgrades eleven television branches into eleven car branches at the cost of downgrading one house branch into a car branch, you will take that action.

In an agent that obeys the VNM theorem, expected value calculations could be an explicit part of its decision-making, or they could just match the outcome of some other complicated procedure that's doing the same math implicitly. And while there's no doubt that the theorem is true as a matter of math, there are disagreements about the extent to which the assumptions apply to ideal agents and AIs and humans, and disagreements about what the conclusions mean, some of which we'll discuss in other articles.

Related

Scratchpad