PHI315: A Lecture on Virtual Minds, Worlds, and Probability

Today, given that we have had a shift in the movies we were going to watch and I could not make up my mind as to what movie we will watch for today, coupled with the fact that I feel the notion of virtual minds and intelligence and life has not been sufficiently addressed, I have chosen to give you a lecture on a series of arguments that a philosopher, Nick Bostrom, makes in his book *Superintelligence*. The arguments I am going to make require mathematical comprehension, so I will start with simple arguments to get you into the mathematical-probabilistic mindset. If you do not immediately grasp the mathematics, this is fine. We will take it one step at a time.

Traditional Probability

Traditional probability is relatively simple. Suppose you have a bag with ten marbles, of which five are red and five are blue. If you were to pull a marble at random, what would the probability be that the marble you pull is red?

The answer is 50%. Why? Becuase out of 10 marbles, there are 5 that are red. Hence, for the initial pull, there is a 5/10 chance you will pull a red, which reduces to 1/2, or .5, or 50%. This is simple probability given numbers we are used to. Probably something you learned in school. But probability can be far more complicated. There is Bayseian statistics, where probabilities are updated as new information comes in. Many philosophers think that the way in which our decision making process is Bayesian–we do not have perfect information, and when we act, we act in accordance to information as it comes in. Hence, for any given information set, our behavior has certain probabilistic dispositions.

Marbles with Large Numbers

Now, probability can get tricky when you look at meta-probability and very large numbers. We are now going to look at what happens when you have a standard easy probability problem turn into a complex large meta probability problem.

Scenario One:

Suppose you have two jars, A and B. **Jar A has 10** marbles and **Jar B has 10** marbles. Each marble is numerically labeled, corresponding, suppose, to the order they were put into the jar. Suppose further that I hand you a jar and you are unable to tell, from the outside, whether by shaking or by observation, which jar you have. If you click a button, the jar will randomly dispense a marble. You click the button and the jar dispenses a marble with a **#7** drawn on it. What are the odds you have **Jar A**? The odds are 1/2 or 50%, just as they are that its **Jar B**. However, things change very quickly if we change some conditions.

Scenario Two:

Suppose you have two jars, **Jar A** and **Jar B**. Much like in the prior scenario, **Jar A has 10 marbles**. Unlike the prior scenario, **Jar B has 1,000,000 marbles**. Now, suppose that you are unable to infer, from the mere appearance alone, which jar you are handed. Perhaps you are not handed a jar at all and they are behind a large glass pannel that you will never penetrate. Point is this, suppose you click a button and a marble with **#9** is dispensed. What are the odds you have **Jar A** versus **Jar B**?

You might think that the odds are 50%, since both **Jar A** and **Jar B** contain a marble with **#9** in it. However, counterintuitively, you are more likely to have **Jar A**. This is so because the odds of a **#9** number dropping from **Jar A** is 1/10 or 10%. However, the odds of a **#9** marble dropping from **Jar B** is 1/1,000,000 or 0.0001%. Based on that fact alone, you can infer that you most likely have **Jar A**. What are the odds?

To solve this problem, we can use Bayes' Theorem. Bayes' Theorem is a way of finding a probability when we know certain other probabilities, as stated earlier, as new information comes in. The formula is:

```
P(A|B) = (P(B|A) * P(A)) / P(B)
```

Where: - P(A|B) is the probability we are trying to find: the probability of having Jar A, given that we drew a #9 marble. - P(B|A) is the probability of drawing a #9 marble, given that we have Jar A. This is 1/10 or 0.1. - P(A) is the prior probability of having Jar A, which is 1/2 or 0.5, since we are equally likely to have Jar A or Jar B. - P(B) is the total probability of drawing a #9 marble.

The total probability of drawing a #9 marble, P(B), can be calculated using the Law of Total Probability as follows:

```
P(B) = P(B|A) * P(A) + P(B|B) * P(B)
```

Where: - P(B|B) is the probability of drawing a #9 marble, given that we have Jar B. This is 1/1,000,000 or 0.000001. - P(B) is the prior probability of having Jar B, which is 1/2 or 0.5, since we are equally likely to have Jar A or Jar B.

Plugging in the values:

```
P(B) = 0.1 * 0.5 + 0.000001 * 0.5
P(A|B) = (0.1 * 0.5) / P(B)
```

Calculating the probability yields:

```
P(A|B)  99.999%
```

This means that the probability that you have Jar A, given that a #9 marble has been drawn, is approximately 99.999%. This high probability reflects the much greater likelihood of drawing a #9 marble from Jar A, which has only 10 marbles, compared to Jar B, which has 1,000,000 marbles.

Kidnapper Hotel

The same mathematical truth holds if you are, suppose, kidnapped by a bunch of twisted philosopher-mathmatecians. Suppose that you are kidnapped and brought blindfolded into a motel room. You are told the following:

1. You will be released if you can guess which hotel you are in, A or B.
2. Your room number is 50.
3. Hotel A has 100 rooms, while Hotel B has 500 rooms.

Doing the same Bayesian analysis yields the following result: the odds of you being in hotel A are 98.04% as opposed to Hotel B, and thus you should guess that you are in hotel A as opposed to hotel B.

What does this mean?

We have looked at probability that is counterintuitive. It feels like the answer is still 50/50 but in fact it is not. You are overwhelmingly likely to have Jar A and be in Hotel A. The reasoning being sound and deductively valid, meaning that it is a fact. It is probabilistic, but it is a fact. When this gets interesting is when it comes to virtual minds and virtual worlds, where virtual is a neutral term to denote being instantiated by a computer rather than by nature.

Suppose that human beings become a technologically advanced civilization such that they are able to create virtual minds and virtual worlds. Why? For historical research, for assistance, for enterntainment, for whatever reason. If a society is capable of creating virtual minds and worlds, they will. Now, suppose that the minds that they create indeed have the experience of being minds–they are, for all concerns and purposes, real minds, merely instantiated via silicon as opposed to carbon, living in a world that exists in silicon, as opposed to in nature.

We can infer that as such a society progresses, it will create so many artificial minds that they will outnumber 'real' minds significantly. Moreover, there will be so many 'simulated' worlds while only one real world. I am sure you can see where I am going with this. But, suppose the following:

1. Virtual minds and virtual worlds are possible.
2. Virtual minds do not know they are virtual.
3. Virtual worlds are indistinguishable from the real world.
4. You have a mind and live in a world.

Which is more likely, that you have a real mind and live in the one real world or that you are a simulated mind living in a simulated world? Mathematically, if there are indeed virtual minds and virtual worlds, if they are at the very least possible in which case temporally they are probably actual, then you most likely are a virtual mind in a virtual world.

Extinction

This sort of argument is also made to bolster the case that our species will probably go extinct. Suppose that the human civilization does not go extinct.

If so, then it will probably spread throughout the solar system and beyond. There will probably exist billions upon billions of humans, and the species will last for millions of years.

Or, the species will die out in say 50 years or something. It has existed for 200,000 or so years. You do not know which timeline you are a part of, but you know you were born not when there are billions of billions of humans. Given that, much like in Hotel A in which there are only 100 rooms and you are in room 50, you are also probably in the extinction timeline, since you are born fairly early on, where there is a limited number of people compared to the intergalactic human species of billions of billions.