Beliefs and Off-the-Path Actions

Adlai Newson, adlai.newson@gmail.com

Let's suppose that you're going on a blind date with a guy, and all guys are either Nice Guys or Mean Guys. As you're getting ready for the date, your mom asks you what your belief is that the guy you are going to meet is a Nice Guy. How do you answer? Well, you don't know anything about him yet, so you would say something like "About 50% of guys are nice, and 50% are mean, so I put the probability at 50%." That is your prior belief (it's prior because you are not using any information from the guy, just general probabilities of nice vs mean guys in the population).

OK, now suppose that both types of guys have the same two possible actions (signals). When you walk into the restaurant they can smile at you or they can look bored. Based on what they do, you have a choice between staying for the date, or sneaking back out the door. Let's suppose that you prefer staying for the date if it's a nice guy, and sneaking back out the door if it's a mean guy.

Here's a reasonable-sounding seperating equilibrium:

- Nice guys smile
- Mean guys look bored
- You stay if you see a smile, leave if the guy looks bored.

This will be an equilibrium if nice guys want you to stay for the date, and mean guys like it when you leave. But what if the mean guys also want you to stay for the date? Probably they will start smiling too, so we'll have a pooling equilibrium:

- Nice guys and mean guys both smile
- When you see a smile, you flip a coin to decide whether to stay or leave.[1]

But notice the problem: what do you do if you see someone look bored? According to the equilibrium, both nice guys and mean guys will smile, so if you show up for the date and see your date looking bored, what should you believe about the type of the guy? First, why doesn't Bayes rule work:

[1]This will be rational if, for example, your utility from staying on a date with a nice guy is 1 and utility from staying on a date with a mean guy is -1, and your utility from leaving is 0; then you are indifferent between staying or not, so a coin flip is as good as anything for deciding whether to stay or leave.

$$
\begin{aligned}
Pr\left(\text{Nice}|\text{bored}\right) &= \frac{P(\text{bored}|\text{Nice})P(\text{Nice})}{P(\text{bored}|\text{Nice})P(\text{Nice}) + P(\text{bored}|\text{Mean})P(\text{Mean})} \\
&= \frac{0 \cdot \frac{1}{2}}{0 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2}} \\
&= \frac{0}{0}
\end{aligned}
$$

Which is undefined! This happens because in the equilibrium nobody is supposed to look bored. Here's the trick: as a game theorist, I can assume anything I want about your beliefs off-path in order to support the equilibrium. What should I assume in this case? Well, first I suppose that when you see your date looking bored, you assume he is a nice guy. Then if you see your date looking bored, you will stay. This means your date has two options: smile, and get a date with probability .5, or look bored, and get a date for sure! This is true for both a nice guy and a mean guy, so everybody will start looking bored. But this breaks the equilibrium, because now both types of guy have a profitable deviation. So what do I have to assume about your beliefs when you see somebody look bored? You have to believe that they are mean.[2] Then with these beliefs, the equilibrium above will work. This is the reason why in a Bayesian equilibrium you have to specify both the equilibrium *actions* of every player, and the *beliefs* of those players.

You might be thinking that there's too much freedom for me to invent whatever crazy beliefs I want to support all kinds of crazy equilibria. Well, as it turns out, there are other criteria for restricting this freedom somewhat. One such example is the Intuitive Criterion (IC), but that's a story for another time...

[2]Technically I only need to assume that when you see your date looking bored, you have to believe he is a mean guy with probability $q \geq .5$.