Preferences, Utility, and Rationality: Game Theory in the Lab

Pierfrancesco Guarino

This short note wants to address a very basic question: can we meaningfully talk about the rationality of a subject involved in an experiment and, if we can, how could we achieve this? To answer this question, first of all we have to be sure that we agree on how two words are used in economics, namely "utility" and "rationality".

Thus, what is the meaning in economics of attaching a certain *utility* to *something*? To answer this question we address it in the most basic setting conceivable, namely by assuming, without loss of generality, that there is no risk or uncertainty involved in the decision. The idea is that this *something*, e.g., an apple, belongs to given set of alternatives (finite, again just to simplify the exposition and without loss of generality), that we call $F = \{ apple, orange, strawberry \}$, on which a decision maker (henceforth, DM) has some preferences, where these preferences are captured by a relation $\succ$ on F, that we call *strict preference*. For example, we assume for a DM under our scrutiny that $apple \succ orange \succ strawberry$: this means that this DM prefers an apple to an orange or a strawberry, and an orange to a strawberry. In this very simplistic case, in order to be able to meaningfully talk about the utility that the DM attaches to an apple, the relation $\succ$ has to satisfy two properties called *asymmetry* (i.e., for elements x, y of F, if $x \succ y$, then it is not the case that $y \succ x$) and *negative transitivity* (i.e., for elements x, y, z of F, if it is not the case that $x \succ y$ and it is not the case that $y \succ z$, then it is not the case that $x \succ z$). If these properties are satisfied, economists say that $\succ$ is a *preference relation* over F and it is possible to represent $\succ$ with a function $u : F \to \mathbb{R}$, called the *utility function* of the DM. Thus, in our example, since it can be easily verified that these properties are satisfied, we could write that $u(apple) = 3$, or – in plain English – that for the DM an apple is equal to 3 *utils*, and also that $u(orange) = 2$ and $u(strawberry) = 1$: what is crucial, given the original preference relation of our DM, is that $u(apple) > u(orange) > u(strawberry)$ and not the actual numbers we decide to attach to the alternatives (this is the reason behind calling this notion *ordinal* utility: it only captures rankings). It is extremely important to notice that we can also start from our set F and create for a DM a function, called the *choice function* $c : 2^F \to F$, that for every subset of alternatives gives the element of this subset that the DM would choose (e.g., $c(F) = apple$, $c(\{ apple, orange \}) = apple$, $c(\{ orange, strawberry \}) = orange$, $c(\{ apple, strawberry \}) = apple$, and $c(\{x\}) = x$ for every x in F). Interestingly, if this function c satisfies one property, called the *α-condition* (that says that, for any two subsets A, B of F, if A is included in B and $c(B)$ belongs to A, then $c(A) = c(B)$), then we can safely state that the DM whose choices we have observed through her choice function c has a preference relation over F. Hence, her preferences can again be represented via a utility function!

Thus, from the previous paragraph we obtain two crucial points concering modern economic theory:

1. preferences can be *observed* and, in particular, can be *elicited* via the choice function (it is this point that gives empirical content to economics);

2. utility in economics is simply a function that happens to exist for a given DM if and only if the preferences that this DM shows to have satisfy a rational ordering.

Concerning "rationality", the issue is conceptually more problematic. Decision theorists have long fought (and are still fighting) with the idea of providing a formal definition of this notion. For game theorists the problem can be decomposed in two different cases:

a) in the equilibrium-based literature, whose goal is to refine the notion of Nash equilibrium to rule out all those equilibria that are unreasonable *from the point of view of game theorists*, rationality is not formalized, rather it is what arises as the result (outside the model itself) of an equilibrium notion that indeed rules out all the unreasonable equilibria;

b) in the epistemic game theory approach, rationality is formally defined from the outset as subjective expected *utility* maximization: a DM is rational if she does not choose any action s that is dominated by another action s' (i.e., such that no matter what the other DMs involved in the game are doing, it is better to choose s' instead of s).

For the sake of our argument, we take the notion of rationality expressed in (b), both because it can be formalized in the language of the theory and because it also intuitively captures the notion that economists seek to capture, but mind that this is (again) without loss of generality! One point needs to be emphasized, namely that a consequence of this definition is that we need a context where a DM expresses preferences enough well-behaved to derive a utility function (otherwise there would not be any utility to talk about in the first place) in order to make inferences concerning her rationality. Hence, rationality cannot be evaluated in a vacuum.

The natural question is now the following: how is all this related to our original problem? Consider an experimenter that sees a subject choosing an action that is dominated from the point of view of the material payoffs provided in the experiment. Can the experimenter infer that this subject is irrational (hence is not playing according to economic theory)? Before moving to an example that should clarify the question, one point. Occasionally in the literature it is mentioned this mythological beast called *homo economicus*, which prototipically is a selfish DM that cares only about her material payoffs. As a matter of fact this description does not capture the behavior of an ideal DM that behaves according to economic theory, since such DM would always maximizes her *subjective expected utility* (which can be different from her material payoff!). In other words, what we wrote about apples and other fruits applies to monetary allocations as well.

The following example will hopefully make more explicit what was written above: take a game between two DMs, call them Ann and Bob, where, given 10$, Ann has to divide them in whatever way she prefers and Bob has to simply accept her decision: this is called in the literature the *Dictator Game*. One question naturally arises: If Ann divides the amount of money in an equal way, i.e., 5$/5$ (where the first number is her material payoff and the second is Bob's), is she kind and rational or selfish and fool? As a matter of fact, we simply do not know, since by taking this single choice, we cannot say anything concerning her utility function and, in turn, we cannot talk about her rationality. To clarify, *if* Ann's utility is a mirror of her (and only her!) material payoff, then we could conclude that she is irrational. Indeed, since her preferences mirror her material payoff, we have that $10\$ \succ 9\$ \succ \cdots \succ 1\$ \succ 0\$$, and – by assuming that the number of utils coincides with the number of dollars – we have that $10 > 9 > \cdots > 1 > 0$, which in turn implies that, by taking just 5$, i.e., 5 utils, she is not maximizing her *utility*. However, if we do not know anything about her overall preferences over the material payoffs, there is no utility function in the picture. Indeed, observe that it is perfectly conceivable to have different

preferences than those just described above: for example, if Ann's preference relation is such that $5\$/5\$ \succ 6\$/4\$ \succ \cdots \succ 9\$/1\$ \succ 10\$/0\$ \succ 4\$/6\$ \succ \cdots \succ 0\$/10\$$, then we could infer that she cares for her well-being *and* also for Bob's well-being and her utility function should incorporate this. Hence, to conclude, this one observation does not give us enough information to meaningfully talk about Ann's utility function.

Considering that the argument in the previous paragraph can be extended (in an even more natural way) to games with a more complex structure than the Dictator Game, what could we say concerning our original question?

Bottom line: if we want to be able to sensibly state that a DM is rational, shouldn't we always divide an experiment from which we want to eventually draw conclusions on the rationality of the subjects in two stages? In stage 1 subjects' preferences should be elicited (in an incentive-compatible way) via their choices on all the possible subsets of the possible outcomes of the game that they have to play in stage 2, and only in stage 2 their actual choices in the game should be recorded and evaluated (along with their rationality... with the caveat that maybe it could not be possible to evaluate it, since stage 1 could show that some subjects' preferences simply cannot be represented via a utility function).