Branislav L. Slantchev[1]

On the Proper Use of Game-Theoretic Models in Conflict Studies

[1] Department of Political Science, University of California, San Diego, CA, USA, E-mail: slantchev@ucsd.edu

Keywords: Game-Theoretic Models, Jan Tinbergen European Peace Science Conference, economics
DOI: 10.1515/peps-2017-0041

1 Introduction

As I stand here today, I am suddenly keenly aware that I am a theorist in the clutches of a bunch of empiricists. Perhaps that should not matter. After all, most of us here would agree that the ideal dissertation consists of a theory (perhaps formalized, perhaps not), hypotheses derived from that theory, and an empirical assessment, preferably multidimensional (large-N statistical analyses, case studies, process-tracing, experiments), of these hypotheses. Me might also agree that the theory should be internally consistent: its assumptions do not contradict each other and its conclusions follow from its premises. Some might even agree that formal modeling could be helpful in ensuring that last part. So one big happy family, right?

Well, obviously not quite right or else I would not be here talking about that. As a practicing theorist who uses applied formal game-theoretic models – and these, by the way, are the ones I will be exclusively talking about today – I cannot help but notice a serious divergence between this supposed ideal and the reality of the discipline. Now, there are good reasons not to have that ideal: it's indefensible as an approach to science. But the underlying problem is more fundamental: there is serious disagreement about the role of theory in general, and formal models in particular, in the advancement of knowledge.

Today I wish to suggest that a lot of this disagreement stems from misunderstanding about what models can and cannot do. And that it is both proponents of the use of models and their critics that are perpetuating this misunderstanding. Proponents often play fast and loose with wonderfully protean terminology, and critics often latch onto some specific meaning of that terminology to attack the entire enterprise. Moreover, many proponents are, I believe, thoroughly confused about what it is that they are doing, and as a result their bombastic claims are turning others into critics of the approach. So today, I would like to take a first step toward lifting some of the fog that is clouding our communication.

There are two main ideas I would like to convey today (along with the subsidiary points that constitute arguments to support them):

1. The negative one: The EITM approach to models as hypothesis-generating devices is wrong, for two reasons. One is that models should not be judged by their empirical performance. I will not talk about that, but you should read Nancy Cartwright and Jim Johnson.[1] The other, which I will talk about, is that even on their own terms, models are not as "scientific" as many wish to believe.

2. The positive one: We should think of models as creative tools for conceptual exploration and more precise communication. We should use them to tell stories.

These arguments apply to any social science research that uses formal models, but since this is NEPS and my examples tend to be drawn from the field I study, I slapped "in conflict studies" in the title. Pretend-humility no doubt.

2 The fake rigor of game-theoretic models

> *More data is better than less, and the attention to research design that accompanies the renewed interest in causal inference is integral to the maturation of our science, as is* ***the clarity and rigor inherent in formal theory****.* (Clark & Golder, 2015)

Branislav L. Slantchev is the corresponding author.

One of the most common defenses you hear about modeling is that they are rigorous. Presumably, this special rigor derives from the fact that mathematical operations are performed on some variables defined by the model. This is somehow supposed to make modeling better than nonmathematical theorizing, among other things. The only problem is: when it comes to using models in social science, there is no rigor. Or, more precisely, the actual rigor that involves mathematics turns out to be merely a form of accounting: useful to communicate with other like-minded researchers and ensure that the derivation of the argument follows some commonly agreed upon rules, but nothing particularly insightful beyond that. Specifically, we can think of the modeling exercise as having three phases – building a model, analyzing it, and interpreting it – and none of them is rigorous: the first and last not at all, and the middle only in a limited sense.

2.1 Building models: the Chinese-speaking flying pig problem

Consider the statement "all flying pigs speak Chinese". It is true. Why? Because nobody has ever seen a flying pig that does not speak Chinese. The problem, of course, is that there is no such thing as a flying pig, and as a result the statement describes properties of an empty set. When we build models, we must take special care that we do not end up with the mathematical equivalent of a flying pig. If we do, we could analyze its properties to our hearts' content without learning anything useful in the process.

Building good models is tough. This is the creative part of modeling. Unlike analysis, it cannot really be taught well. Models have to be complex enough to capture the interaction one wishes to analyze: make them too simple and they yield trivial results. But they have to be sufficiently simple to enable us to understand how the interaction works once it is analyzed: make them too complex and they become insoluble or unintelligible. It takes quite a bit of insight and practice to come up with an interesting but tractable model. It takes a lot more of both to come up with one that is not a flying pig.

Let me illustrate this with an IR example. We are all aware of the venerable claim that wars are caused when nations disagree about their relative strength and neither is willing to concede enough to make the other unwilling to fight. This is the "mutual optimism" explanation for war and it's best informal statement is by Blainey (1988). Subsequent research built on this insight by searching for causes of that optimism and, in the "rationalist" school, by coming up with a rationalization of the mechanism (Blainey himself favors the non-rational factors, at least when it comes to the rise of that optimism). One can easily expand Fearon's (1995) "private information with incentives to misrepresent" into a full-fledged version of the mechanism.

This happy state of affairs was rudely disrupted by a very provocative paper published by Fey and Ramsay (2007). Their central claim was that mutual optimism cannot lead to war in a rationalist framework, even if we were to allow for some forms of bounded rationality. They assumed, as is common in the bargaining approach to war, that war is costlier than peace. Unlike existing models, however, where war could occur whenever any player chose to initiate it, they required that both players agree to fight for war to begin. Reasonably, this is what one needs to do if one were to analyze **mutual** optimism where both players choose to fight. They then used standard tools developed to model knowledge, combined them with standard tools used to analyze strategic situations independent of the game tree, and showed that there exist no (Bayesian) Nash equilibria in which war occurs with positive probability.

To say that this paper caused a commotion (at least among people interested in the causes of war) would be an understatement. The result threw into doubt a long tradition of research, and challenged some deeply held convictions. I was among the sceptics. I never quite understood why such an intuitive idea as "mutual optimism causes war" would fall apart. I also did not quite understand how the analysis worked: I could follow the math, everything seemed correct, the conclusion followed… and I just did not believe it. For my money, it was in Chinese and I started to suspect that the model might be a flying pig: how else could one get such a seemingly ridiculous conclusion?

Things came to a head a few months after the article was published as I was mulling a possible response. My friend Ahmer was visiting UCSD and told me that he had just had a paper rejected because it used the standard incomplete information model of crisis bargaining and one of the referees said that this whole approach has been disproven by the Fey and Ramsay paper. From Ahmer's perspective, their article was not merely an intellectual curiosity, it was now doing active harm to his research agenda. So we resolved to write a response together (Slantchev & Tarar, 2011).

Our immediate problem was that every extensive game form that we were familiar with produced equilibria with war. They all seemed reasonable and yet they obviously did not fit in the general category of models considered by Fey and Ramsay. We wrote down other variants that we could think of, but no, they did not work either: war was always there under the right conditions. Obviously, we did not understand the general class described by Fey and Ramsay.

After spending more time on that, we finally realized that the innocuous definition of mutual optimism they use is actually extremely restrictive. In particular, it requires that either player can avoid war by reaching

for a negotiated settlement that is *unrelated* to the expected payoffs of war. That is, a player could avoid war by inducing a peaceful outcome that left the other player with a payoff worse than war. This would be equivalent to assuming that Iraq could end the war with ISIS by stopping the fight and imposing a peace that gives ISIS nothing. This is, of course, absurd, and we were able to show that in the class of models Fey and Ramsay analyze, this is precisely the structure they use (this means that they do not get war under complete information even when they should). Digging deeper, we discovered that two assumptions are jointly needed to generate the result under incomplete information: each player can unilaterally avoid war by choosing to negotiate, and the terms of the peace settlement cannot depend on the behavior of the players. Needless to say, if one makes these assumptions, the notion of crisis bargaining is eviscerated.

How did that happen? The problem was that Fey and Ramsay came at the problem from an analogy with economics, where there is a result known as a "no trade theorem" that says that trade/speculation cannot be the result of private information. The reasoning is simple: if a trader has private information about the value of an asset and wishes to trade it for a more valuable one, the other trader will infer the existence of such information from the willingness to trade, and will revise his valuation accordingly. If he, in turn, is still willing to trade, the first trader would infer that there is also information he isn't privy to, and will revise his valuation accordingly. The process continues until their valuations converge, and no trade can occur.[2]

The problem with using this analogy is that crisis bargaining is not voluntary in the sense trade is. If I conclude that my asset has a higher value than your asset, I can just walk away from the proposed trade, and that would be it. If you still wish to trade, too bad for you. It's not like you could clobber me on the head with a bat and take that asset from me. But that's precisely what crisis bargaining amounts to. If war is not to occur, any voluntary peace deal must be such that its terms satisfy the minimal war expectations of both actors. It cannot be the case that the terms of peace are independent of the expected war payoffs because anyone can start a fight when they think peace does not give them a good enough deal. Fey and Ramsay ignored this, and as a result ended up modeling an empty set of crisis interactions. They might find all sorts of fancy properties of this class, but their usefulness to the study of crisis behavior is nil.

Why do I bring this up? Because the argument I just made is not rigorous at all, at least not in the mathematical sense. It is a logical (and hopefully convincing) one, but you can choose to remain unpersuaded. Fey and Ramsay (2016) have done so, for example. It's not like I showed you that they made a mathematical mistake in their analysis (this would have been a rigorous way to disprove a result). I made a verbal argument. I happen to think that Ahmer and I are right, but at the end of the day, I cannot prove that, I can only try to persuade you.

Building models – choosing what to put in and what to leave out, how the interaction unfolds, what the players know, what their preferences are – is messy and not at all rigorous. There is simply no list of commonly accepted criteria that a model must satisfy for us to agree that it is a good model.

So much for the input: if you're not careful, garbage will go in.

2.2 Interpreting models: the flaming trousers conjecture

Let me ask you: do you think it is possible to write a reasonable game-theoretic model of the banking system, where in equilibrium bank presidents publicly set their pants on fire? Richard Rumelt certainly thought so (and didn't like it), or as he put it:

> *The trouble with game theory is that it can explain anything. If a bank president was standing in the street and lighting his pants on fire, some game theorist would explain it as rational.* [3]

Well, can one do it? Sure. Consider a banking system where the solvency of each individual bank is private information to its president. Assume that potential customers prefer to deposit their money in safer banks. Finally, assume that setting the pants on fire is less costly for the president of a solvent bank (perhaps because he would get a higher bonus to buy new pants and medical treatment). That's it: you can easily construct a separating equilibrium, in which only presidents of solvent banks publicly set their pants on fire. Having observed that costly signal, customers deposit their money into the solvent bank. A relaxation on the difference in the costliness of setting the pants on fire would further produce a semi-separating equilibrium, where some presidents of insolvent banks sometimes set their pants on fire too, and the customers randomize among the banks. So there you have it: a rational explanation for people setting their pants on fire in the street.

Or is it? People tend to use this illustration to assert that game theory is ridiculously adaptable. Any behavior could be rationalized with some model. In fact, I suspect many scholars believe that this is what we do – find some interesting (or not so interesting) pattern of behavior and then throw together a model that yields it in equilibrium. Well, OK, many models do, in fact, look precisely like that. But they are usually easy to spot because you can always see where the rabbit goes into the hat: there's almost invariably some quirky assumption that

buys the intended result almost directly, and the rest of the math machinery is just baroque ornamentation designed to distract both the author and the reader from the fact that the model is adding nothing of value. (I think this is the most common reason I reject papers with models in them.) But this is merely shoddy modeling. It is not a problem with the method itself but with its bad application.

The flaming trousers example shows a deeper problem: the model is only as good as its interpretation. The signaling game is standard, and the result turns on the parameter labeled "cost" of the action "setting pants on fire". While the model is analyzed rigorously (and even that, as we shall see shortly, has its own issues), attaching labels to parameters – the interpretation of what the model is saying – is most decidedly not rigorous at all. We could have just as easily, and perhaps more convincingly, called the action "opening up the bank to a public audit" or "depositing the president's entire wealth into the bank". Since the audit is more likely to reveal healthy finances when the bank is solvent, the argument would go through. Also, since the president would not risk his own money on a failing enterprise, the second interpretation could also go through.

Did we just make the model more reasonable? No, we did not. The model is exactly the same. But its usefulness to us went from ludicrously irrelevant to potentially quite important merely because we attached different labels to a variable. Is there a rigorous way to label variables? Of course not, and that is the point: even the best-constructed and correctly-analyzed model would only be useful if we interpret it appropriately, and that final step is creative, interesting, and entirely non-rigorous (in the mathematical sense). Reasonable people can disagree on what a variable represents. They can also agree that it could represent a great many things. But how persuasive these agreements and disagreements are depends on the arguments being made, not on a standard criterion that can settle them independently.

Models do not come pre-interpreted, and neither do they define their own interpretations. The application is quite external to the models themselves. As with the building stage, the interpretation stage is an exercise in persuasion, not assertion of truth or a standard definition.

Think now what this implies for the EITM approach that exhorts us to test hypotheses "derived" from the model. I put "derived" in quotation marks to indicate that the hypotheses are not, in fact, derived from the model but from our interpretation of its variables. This pseudo-scientific exercise often leads to bogus scientism in coding of parameters that are then plugged into regressions.

Consider, for example, Fearon's (1994) article were he tests a crisis bargaining model. He looks at crises in which a potential challenger decides whether to threaten a defender, who then responds by deciding whether to resist by mobilizing, giving in either case the challenger an opportunity to back off, and if she does not, the defender can decide whether to fight. In this model, general deterrence success is when the challenger does not issue a threat, and immediate deterrence success is when she backs off after the defender mobilizes. The question is: does the success of either type of deterrence depend on measures of power?

This question had generated a lively debate with mixed results. Fearon's main point is that the debate had ignored selection effects. Suppose the military balance of capabilities is observable before the crisis but the challenger is uncertain about the resolve of the defender on the issue. When this balance favors the defender, the challenger would only threaten when she thinks the defender does not care all that much about the issue or when she herself is more resolved. Thus, the more powerful the defender, the more likely is general deterrence to succeed. But since the threat is now issued only by more resolute challengers when they believe the defender might give in, all else equal, immediate deterrence should be more likely to fail. The "all else equal" part is important because if the defender could somehow reveal his actual resolve during the crisis, any challenger that issued a threat but was not prepared to fight would back down. Thus, measures of relative military strength (or defender interest) revealed during the crisis should make immediate deterrence more likely to succeed.

Fearon then uses the Huth and Russett (1984) deterrence dataset to reassess previous results in the light of the important distinction between *ex ante* and *ex post* indicators of interest and relative strength. He argues that two variables – long- and short-term balances of power – are properly understood as measures of strength available prior to the crisis. (The first measures overall military and industrial capabilities – the familiar CINC scores from COW, and the second measures capacity to mobilize troops.) In contrast, the immediate balance of power, which is measured as the ratio of forces *present at the point of conflict immediately prior to the onset of hostilities or retreat*, is properly understood as a measure of strength that was revealed during the crisis. It is, in fact, the only ex post indicator among the variables, and as such crucial for the discussion that follows.

Fearon himself never actually took a closer look at the ex post variable (probably because it showed up with the same sign as the ex ante variables, which contradicted the expectations of the theory). Instead, he argued that the short-term balance of forces should be a proxy for the challenger's initial beliefs about the defender's willingness to use force, and should thus be positively correlated with immediate deterrence success (because weak challengers would only threaten an observably stronger defender if they were really unsure that he would respond, making the response more effective). Notice here how an *interpretation* of a seemingly straightforward variable was used to explain its rather unexpected effect if one were to treat it as an ex ante measure of relative power. The problem of labeling is not limited to theoretical models.

Signorino and Tarar (2006) took the analysis further. They examined very closely the effects of all three variables and found that both short-term and immediate balance of forces improved the chances of immediate deterrence success, but that it was the immediate balance that had a much stronger impact. Thus, they replicated the (somewhat puzzling absent the ad hoc reinterpretation) finding about short-term balance of forces from Fearon's analysis, and in addition seem to have provided a much more direct evidence for the selection mechanism since the only ex post indicator was behaving precisely as it should according to the theory.

Great!

Not so fast. First, let me ask you: do you think that the selection logic makes sense? Of course it does! Does the mechanism that explains how prior and updated beliefs might influence crisis behavior and therefore affect crisis outcomes make sense? Of course it does! So what is the role of the statistical analysis? Is it supposed to enhance our confidence in the argument? Or to demonstrate how widely applicable it is?

What if I, ignoring the ad hoc interpretation of the short-term balance of forces, told you that the one and only crucial ex post measure, the intermediate balance of forces, is bogus? This is through no fault of either Fearon or Signorino and Tarar. You see, while the description of the variable asserts that it measures forces *present at the point of conflict immediately prior to the onset of hostilities or retreat*, it does no such thing. Instead, it is a measure of the overall military capabilities "adjusted to reflect the impact of distance. If both defender and attacker share a border, or share one with the protégé, then no adjustment is made. Otherwise the indices of military capabilities are adjusted for the distances between the attacker's and defender's loci of power and the protégé, and for the principal military transport capabilities of the day (expressed in travel days)".[4] Basically, they used Bueno de Mesquita's (1983, p. 105) formula for the "loss of strength" gradient. As Huth and Russett (1984, pp. 509–510) say, "Ideally it might have been desirable to measure actual locally available military forces (divisions, gunboats, etc.) but that would have required a level of information not easily accessible... The procedure we did use seems less arbitrary."

Now, I can attest that it is quite difficult to measure the balance of forces available at the point of impact, so to speak. It is, in fact, my effort to do so that led me to the discovery of the curious mismatch between label and actual content of the variable in question. The point here is not whether this coding is more or less arbitrary than others. The point is that *the variable is composed entirely of ex ante observable indicators*. It is no ex post measure of relative capabilities. It contains no new information revealed during the course of the crisis. It cannot possibly be measuring the thing that Fearon had in mind, and what Signorino and Tarar thought it did. More importantly, it cannot have the effect on immediate deterrence a true ex post variable should have. Even worse, since it is, in fact, an ex ante indicator, the effect of the variable uncovered by the statistical analysis actually contradicts the theory.[5] At the very best, we have a finding in search of an explanation.

Does this shake your confidence in the argument? Well, it shouldn't. Any reasonable person can see that the argument makes sense. What you should really be worried about is whether the interpretation of the variables in the model *and* their empirical representation make sense.

Now, what if I told you to consider a model where the immediate deterrent threat isn't just "resist or not" but, say, mobilization of troops? (This actually brings the model closer in line with the informal discussion and with what the variable had attempted to measure as the local/immediate balance of forces.) What if I told you that, because of that same selection logic, this mobilization tends to be more aggressive, and so immediate deterrence becomes *more* likely to succeed? And since ex ante stronger defenders can afford to mobilize more aggressively, the ex ante measures should be correlated with success in both general and immediate deterrence (Slantchev, 2011)? The only difference is that they would have to mobilize relatively more aggressively than others.

Does this shake your confidence in the original argument? It should. The new argument is quite convincing, if I may say so. But there is nothing empirical about it. It is a conceptual exercise. And not really rigorous in the mathematical sense.

So, if you're not careful how you interpret the model, then garbage comes out.

2.3 Analyzing models: the goldilocks property

Perhaps it is the middle part of modeling – the analysis – that is rigorous? If we limit ourselves to the question of whether some behaviors meet a particular definition of rationality or not, then this is so. That is, we can rigorously answer whether a set of strategies is an equilibrium or not. The definition of equilibrium itself is, of course, rigorous as well. What isn't rigorous, however, is our choice of a particular equilibrium as the solution concept; that is, as the set of criteria strategies must satisfy. To put it bluntly, the choice of a definition of rationality is not rigorous. To see this, we shall make a brief detour to dispel some very common misconceptions about rationality in game-theoretic models.

Game-theoretic analysis is based on an assumption that people are rational (Myerson, 2006).

Let me start with a grating misconception that stems from poor nomenclature: despite its name, game theory is *not* a theory of behavior. It is a *method* for analyzing strategic interactions (Kreps, 1990). It has no substantive content beyond the notion of what we should consider important (actors, beliefs, options, information, payoffs) and what constitutes optimal behavior within the model (equilibrium solution concept).

Let me unpack this a bit. There is one thing that game-theoretic models assume, and that is that behavior is goal-oriented. Some people, e.g. Harsanyi (1977, 84) call this "rationality" but this is a rather empty concept. When it comes to explaining behavior – with or without game-theoretic models – this sort of rationality should have a "presumptive priority" in the absence of some evidence to the contrary (Johnson, 2017). Without further elaboration, this definition of rationality is not useful.

What game theory does is *provide a specific definition, or rather, several definitions, of rationality*. We call them "solution concepts". They are lists of requirements that behaviors must satisfy to be considered solutions to the model. For example, the solution concept of Nash equilibrium requires that each player's strategy is a best response to whatever all other players are doing. That is, it *defines* rational behavior as the action that maximizes the player's payoff (best response) when he expects all other players to be taking actions that maximize their payoffs under analogous expectations.

But Nash equilibrium is not the only definition of rationality in game theory.

One might wish to relax the requirement that the player expects everyone else to be playing best response strategies, and instead define rational behavior as the action that maximizes his payoff given some subjective belief about what the other players might be doing. Instead of restricting this belief to the expectation that they are choosing best responses, one might only wish to restrict it to the expectation that they are not choosing strategies that are never best responses for them (i.e. strictly dominated strategies). This definition – it is called "rationalizability" – will label many more behaviors as "rational" compared to Nash equilibrium.

Alternatively, one might wish to strengthen the requirement that players expect everyone else to be playing a best response given what the other players are doing, and instead require that they play best responses even in contingencies that would not arise given the strategies of the players. That is, we now require optimal behavior in hypothetical, off-the-path of play, situations. This definition – it is called "subgame perfection" – will usually eliminate some Nash equilibria, and so restrict the set of behaviors we would label "rational".

I will talk more about that in a bit, so for now the only point is that game theory does not *assume* that players are rational; it *defines* what "rational" behavior looks like. That is, defining what "rational" means more specifically is itself part of game theory. Not surprisingly, game theory does not even have a single such definition despite the dominance of Nash equilibrium, and there is quite bit of discussion about which definition is "more appropriate" than others.[6] (The answer, it seems, is rather context-dependent.)

So where does this leave us? What is it that the models are doing if all they seem to be telling you is whether particular strategies satisfy a seemingly arbitrary list of criteria that are collectively, and somewhat contentiously, labeled "rationality"? Rather than telling you that some observable behavior is rational because it can occur in equilibrium, they are telling you that you could understand that behavior (rationalize it) as arising from incentives and constraints that the definition of equilibrium represents. It is telling you that if you consider these incentives and considerations important, then you might be interested in knowing they can interact in particular ways, resulting in some specific behaviors.

Even the best practitioners often get this wrong (or at least they are careless with their talk about what it is that they are doing), as the quote from Myerson shows. In fact, theorists often play fast and loose with the term "rationality," applying it to several distinct and unrelated concepts. The broad, imprecise, and implicit definition of what we mean by "rational" creates a lot of confusion. I am not going to burden you with an exhaustive (and exhausting) list of possible definitions. Instead, I will focus on several that are pertinent to our discussion.[7]

2.3.1 Purposeful action in pursuit of some objective

Rationality here simply means that people act in a way consistent with the pursuit of some objective. That is, their behavior is purposeful, and its purpose is to achieve some outcome that they prefer to others. This is a very thin, almost vacuous, notion of rationality and I would say it is absolutely essential if we are to hope for any explanation as social scientists. Why? Because at the end of the day, any social phenomenon is the aggregate of individual behaviors. To understand it, we need to explain those behaviors. We need to understand why an individual would do what we observe them doing. But how do we "understand" behavior in everyday life? We *rationalize* it: that is, we try to relate it to preferences the individual might have and constraints that might affect their behavior. We look for reasons for them to behave in particular ways. When we do not understand the behavior, we call it *irrational*.

The desire to rationalize is so extreme that we tend to attribute other people's behavior to their internal motives and beliefs even while we see our own behavior more conditioned by external constraints. (Psychologists call these the "dispositional" and "situational" attributions.) The key here is that we intuitively understand behavior as arising from internal motivations but modified by external forces. That is, we relate it to preferences/beliefs and the context. When we get cut off on the freeway, we get mad at the other driver: What a jerk! That's because we implicitly infer from their behavior that they do not care about our safety or the rules of the road: that is we rationalize them cutting us off by making an assumption about their preference ordering. Some of you might be more charitable and assume that perhaps they did not see you: again, we are rationalizing the behavior by assuming they have the "right" preference ordering but not sufficient information to act "properly". Some might even go further and assume that perhaps they have a good reason for the driving; some sort of emergency that could reasonably override the usual preference for safety (a woman in labor?). And again, we are rationalizing the behavior by trying to reconcile it with some sort of preferences and information. What we almost never do is immediately assume that the other driver is crazy or that their driving is unrelated to anything they want.

This is how we understand behavior in everyday life, and this is how social science should explain it to us. *Our models must rationalize behavior. They must show how it can be related to preferences and beliefs individuals have, and constraints under which they must operate.*

The next definition of rationality is a very unfortunate case of bad labeling.

2.3.2 Minimally logically coherent preferences

In their defense of modeling, theorists sometimes assert that the only assumption they make is that preferences are rational, by which they mean they are complete and transitive. This has absolutely nothing to do with any intuitive understanding of rationality, and I have always disliked its use here.

What is being assumed here is that (a) individuals have preferences over the relevant outcomes that might arise from their behavior, and (b) these preferences are logically coherent.

The first requirement is important in two ways. First, if individuals fail to consider some relevant outcome, their choices can easily end up as mistakes in the sense that they do not relate to their preferences in the way they desire. (This is not the same as estimating a low probability of an outcome, taking a considered action, and that outcome getting realized: what looks like a mistake ex post was, in fact, a rational decision *ex ante*.) Second, if we omit these outcomes from the model, we would misdiagnose the behavior. (For instance, leaving out the possibility of a woman in labor in the other car.) In reality, of course, people sometimes fail to consider all potential outcomes, but then they regret that and they realize they have made a mistake. That is because *we want to be rational in our choices*. They will usually correct that mistake for the next iteration, if any. Notice, however, that the mistake only becomes intelligible as such in the context of having complete preferences.

The second requirement is what allows us to make consistent inferences. It states that preferences are ordered in a coherent manner, like numbers. If I prefer A to B and B to C, then I must prefer A to C. That is so because we rationalize behavior by relating it to a desire to achieve an outcome that is higher on that list of preferences. The notion of "higher" is well-defined and unambiguous only when preferences are coherent. Otherwise, any outcome can be "higher" than the others depending on the order of comparisons. But if any outcome could be ranked as the most preferred one, then any behavior can be rationalized, which means that none of it can be explained.

This is why *our models need to assume preferences are defined over all relevant outcomes and that they are minimally logically consistent.*[8] I would not call these preferences "rational" though.

And, in fact, the assumption is not as innocuous as it sounds when you move to choices that involve uncertain outcomes. When uncertainty is involved, we have to deal with preferences over "lotteries"; that is, choices that generate probability distributions over outcomes. Ranking these lotteries in a consistent manner is not straightforward because we have to make assumptions about how people deal with risk, and people are not very good at comparing probabilistic outcomes. What von Neumann and Morgenstern did was show what you need to assume about preferences over lotteries in order to be able to rank them consistently.[9]

Why is this necessary? Because when you can rank them consistently, you can represent the resulting ranking with numbers. Although not arbitrary in the sense that their ordering also encodes the intensity of preferences, the representation is not unique (because the rank ordering is preserved under affine transformations). This is why the so-called "inter-personal comparisons of utilities" – where you say that person A's utility for an outcome is higher than person B's utility – are nonsense. Given these numbers, it is then possible to show that you could generate them by simple multiplication of numbers assigned to certain outcomes (the misleadingly called "utilities" that represent the rank ordering of certain outcomes) and the probabilities with which these

outcomes occur, and then summing over all possibilities. In other words, you can generate the numbers that represent the preference ordering over uncertain outcomes by expected utility calculations.

Why is this important? For starters, because it allows us to use math techniques to analyze choices over preferences: the choice with the highest expected utility is the choice the decision-maker prefers most, and so we would expect her to make that choice. This brings me to the reason I wanted to discuss this: *it is not the case that the player makes this choice because it yields the highest expected utility; it's exactly the opposite: it yields the highest expected utility because it is her most-preferred choice, which is why she picks it.*

In other words, and contrary to much misplaced criticism of the "people are not expected utility calculators," nobody claims that people actually compute expected utilities or carry around some numbers for payoffs from outcomes. What is being claimed that if the preferences over risky choices satisfy three assumptions, it will be possible to represent their rank ordering with expected utilities, and so we (the analysts) can use these calculations to analyze changes in the ordering under various conditions, which in turn allows us to infer which choices would be preferable under different circumstances.

For instance, we might want to know how this preference ordering changes when the probability distributions over the outcomes change because of what the player expects the other players to do. With numbers, it is a simple matter of plugging these probability distributions into the expected utility formula and recalculating. What we do with the result, however, depends on how we restrict what beliefs about these probability distributions the player can have.

2.3.3 Set of characteristics behavior must satisfy

Relating behavior to preferences is straightforward in a world where the outcomes only depend on your choice and perhaps external factors. (This is what decision theory is all about.) It is much less so when the outcomes also depend on what other people do. That is so because in order to figure out what your best course of action is, you need to make predictions about what the others might do under different circumstances.

2.3.3.1 Iterated elimination of strictly dominated strategies

Sometimes what they do does not matter: you might have an option that gives you the best outcome irrespective of what the others do. Consider the single-shot Prisoner's Dilemma, where defection is the best choice whether or not the other cooperates or defects. In our lingo, defection is a strictly dominant strategy. No reasonable person should be expected to cooperate under circumstances where this invariably leaves them worse off. One *definition of rationality*, then is that it describes behavior that does not involve strictly dominated strategies.

Is this definition of rationality useful? It is certainly intuitive. It is also strong enough to eliminate some types of behaviors as failing to satisfy it. For example, in the PD game, it would tell us that we should not expect either player to choose to cooperate. Eliminating this strategy for each player as failing to satisfy our definition of rationality, we obtain defection as the only surviving strategy.

Our definition of rationality has enabled us to rationalize a unique outcome (mutual defection). In other words, this definition of rationality can provide us with an understanding of how it is that this outcome could arise given the individual preference orderings, where it is next-to-last on the list of preferred outcomes and where it is strictly dominated for both players by the mutual cooperation outcome.

This definition of rationality can be useful in more complicated settings too if we are willing to apply it to make interim inferences. For instance, suppose that my opponent has a strictly dominated strategy but I do not. Since our definition of rationality says she would never use it, I can eliminate it from my expectations about her behavior. Suppose now that focusing only on her remaining undominated strategies I find that I have a strategy that is strictly dominated. Under the definition of rationality, I would never choose it. Note now that my opponent's rationality (in the sense of not playing dominated strategies) combined with her knowledge of my rationality (I do not play such strategies either) allows her to infer that I must ignore her dominated strategy, which means I would then ignore my (newly) dominated strategy, which in turn enables her to remove that strategy from her expectations about my play. Continuing iteratively in this way, we can eliminate strictly dominated strategies until we are left with undominated ones.[10]

Sometimes this process will leave only one strategy for each player, giving us a unique solution. Most often, however, it will not. The problem with the definition of rationality as players not choosing strictly dominated strategies is that it is too demanding: in most social situations there will be very few strategies that can be eliminated from consideration this way, leaving a whole lot of possibilities that we would not know what to do with. This definition is not terribly useful.

2.3.3.2 Rationalizability

One "easy" solution to the problem of not knowing what to do once all strictly dominated strategies are eliminated is to say that all remaining strategies are rational. That is, we can define behavior as rational if it yields the best expected payoff given the player's subjective beliefs about what other players might do. In other words, a strategy is rationalizable if a player can justify using it by explaining that it is the best response to what she thinks the other players could do. If a strategy is never a best response, then our definition of rationality requires that it be dropped from consideration. As before, we can iterate on the elimination of strategies that are never best responses.

In 2-player games, never best response strategies are strictly dominated and vice versa. This means that finding rationalizable strategies is equivalent to finding all profiles that survive iterated elimination of strictly dominant strategies. (This is what I meant when I said that we are essentially defining all surviving strategies as rational.) This will also be the case in n-player games if we assume that the strategies can be correlated. If strategies are independent, there could be cases where some strategies are never best responses even though they are not strictly dominated (Osborne & Rubinstein, 1994, 4.1 and 4.2).

This definition of rationality also makes sense, and solves the problem of indeterminacy by declaring that it is reasonable to have it. This might be so. But, it is a very serious drawback because in many situations too many strategies will satisfy that definition of rationality. In a sense, "rationalizability" has the opposite problem of using strictly dominant strategies: is too permissive as a definition of rationality. In many games, in fact, *all* strategies would be rationalizable (e.g. Matching Pennies), leaving us with the very unsatisfying conclusion that the choice is unpredictable. If everything is rational, then nothing is. For this reason, rationalizability is also not terribly useful.

2.3.3.3 Mutual best responses (Nash equilibrium)

The definition of rationality has a Goldilocks flavor to it: it should be strong enough to eliminate many possible behaviors, and yet not so strong that no behavior satisfies it. One reason Nash equilibrium has become so popular is precisely because it defines rationality in such a way for a great many situations. (The definitions are actually related: every Nash equilibrium can only involve strategies that are rationalizable, and if a unique strategy profile survives iterated elimination of strictly dominated strategies, then it is the unique Nash equilibrium.) It strengthens rationalizability by requiring not merely that the strategies are best responses to some conjecture, but that these conjectures are also consistent with the choice of best responses. That is, it places restrictions on what players are allowed to conjecture about the choices of the other players: Nash requires that they limit these conjectures to the other players also choosing best responses. In other words, a Nash equilibrium is a set of strategies that are mutually best responses.

Now, Nash equilibrium is not the ultimate definition of rationality. Despite the Goldilocks quality to it, it might still sometimes be too strong (so it eliminates reasonable profiles) and it might still sometimes be too weak (allows unreasonable ones).

The first possibility is not widely considered to be a problem. The second, on the other hand, has animated a long tradition of equilibrium refinements; that is, attempt to provide a stronger definition of rationality that would eliminate behaviors that appear unreasonable. Thus we ended up with trembling-hand, subgame-perfect, proper, strong, perfect Bayesian, sequential, and so on equilibria, along with additional impositions on what beliefs could be reasonable. In all of these refinements, Nash equilibrium remains the core: all other definitions eliminate some Nash equilibria under their stricter criteria but they never admit profiles that are not Nash equilibria. Although it is possible to commit oneself to a definition of rationality that yields a unique Nash equilibrium – Harsanyi and Selten's (1988) tracing procedure – most of us would consider this to be unreasonably demanding. Unreasonable because we think that many social situations can, in fact, have multiple reasonable behaviors associated with them, and we should not artificially reduce that number for analytical purposes.

> *My general point here is that the reason we use Nash equilibrium is because it has proven to be a useful definition of rationality on account of its* ***Goldilocks Property****: it is neither so weak that too many behaviors satisfy it nor so strong that none do. There is nothing particularly rigorous or self-evident about choosing it as the fundamental solution concept in game theory.* [11]

So, the rigorous analysis in the middle is only as good as the solution concept it seeks to satisfy.

2.4 Conclusion about rigor

There are several benefits of using models, and I shall enumerate some of them in a moment. But rigor – in the sense of having a common standard that is universally and unequivocally applied to reach the same conclusions

– is not one of them. Because of this, the precision that models seem to yield is rather bogus. I am always amused when a model deploys a concept like "power" and I am then confronted with statistical analysis with substantive effects that involve precision with four digits past the decimal point. Really?

To invoke a metaphor used by Cartwright, modes are not vending machines (Cartwright, 1999, p. 184). We don't insert currency (assumptions, parameters), wait for the internal mechanism to activate the appropriate machinery (analysis), receive the desired confection (equilibrium), and then consume it (empirical tests of its predictions). This approach eliminates creativity in building the model, analyzing it, and interpreting it. It also injects false scientism into the empirical component.

Quite apart from the philosophical reasons that models should not be treated as mere hypotheses-generators, we have reasons to be skeptical of any "tests" of models. Not because models are neither true nor false, and so it makes no sense to test them in the first place. Not because finding evidence consistent with the model tells you nothing about the validity of the model (that's the common fallacy of affirming the consequent).[12] Not because many models make no empirical predictions whatsoever (the so-called "chaos" theorems) or are pitched at such a high level of abstraction that they yield only the most trivial predictions (the bargaining model of war). All of the above are valid reasons. My point is that trying to assess the usefulness of a formal model by treating its variables, structure, and magnitude of comparative statics as if they are came from some rigorously defined mathematical object with precise properties is, well, unhelpful. That is not what models really are, and that's not what they really do.

3 So, what are models good for?

Let me sketch a few answers to that. For a clearer, less telegraphic, and better reasoned version of this argument, see Johnson (2017). Now, I am not going to suggest that models are some abstractions that float above empirical reality and that should avoid any contamination through contact with data. Ultimately, we have to bring them to bear on some actual phenomenon that we wish to understand. So it is not the case that I am against putting them to empirical use, so to speak. I am simply against doing this in a way that sells models short.

3.1 Making the abstract concrete

Models must somehow represent reality. This is where people usually start talking about models-as-maps, and the simplification of reality. Briefly, the idea is that the model is not a full description of reality. Like a map of a country with a 1:1 scale, that would be impossible. More to the point, it would also be useless. The point of a map is to make reality understandable by eliminating all irrelevant detail, which in turn allows for a great reduction in scale. A good map only has the elements that are necessary for its purpose. And so does a good model, which throws out all variables and interactions irrelevant for its purpose.

I am obviously being vague here. For instance, what is that purpose? How do we decide what is relevant and what isn't? If I am in New York City and wish to visit several landmarks efficiently without walking or biking great distances, and without splurging on a cab, I good topographical map of Manhattan is not going to be a good choice. It will have a lot of detail but most of it would be merely clutter given what I wish to use the map for. A tourist map listing all the attractions (even one of those hand-drawn ones that does not respect distances) might be a better bet: it will at least give me a sense of where the sites are with respect to each other, along with some minimal information on how to get to them. A map with the attractions and public transportation routes will be even better. If the sites are easily reachable by subway, then a map of the subway routes will be sufficient and probably best. Note that the highly stylized subway map does not show distances aside from number of stops, and it rarely reflects the true routes of the trains. In that sense, it is a drastic and very distorting simplification of reality compared to a topographical map. But it is by far the best map for my purposes.

And so it is with models: purpose determines scale and degree of simplification. Sometimes a great model will offer such a dizzying simplification that one would be tempted to dismiss it as useless. Don't do it too fast. Just think of Fearon's (1995) bargaining model of war. Only two actors? Disputing over a single dimension? One makes a demand, the other only agrees or not? Whatever one agrees to gets implemented, and the implementation is costless? Disagreement automatically leads to war? War is a costly lottery with only win and lose outcomes? The outcomes are winner-take-all? Uncertainty is over the costs of fighting? Each player knows their own costs? No further interaction once war begins? This is just a sampling of what the simplification has left out; I could go on.

But so what? Despite these simplifications, or, rather, *because of them*, the model serves its purpose admirably. What is its purpose? Contrary to the widespread belief in the discipline, it is not to enumerate three reasons bargaining could be inefficient. Two of them – incomplete information and commitment problems – had been

well-known already, and the third is mechanical.[13] It is not because it provides an exhaustive list of causes – it does not (there are very many others, as subsequent research has shown).

The purpose of Fearon's model is conceptual: it frames war as a method of dividing a benefit that is less efficient than peace (because fighting is both costly and risky), and shows that this implies that it is always possible to divide the benefit peacefully in a way that satisfies the war expectations of each player and leaves something extra on top of those. In other words, it is always possible to design a peaceful distribution of the benefit such that neither player would find it preferable to fight to overturn it. This creates the famous puzzle: if it is always possible to satisfy the war expectations of the players peacefully, then why would they ever fight?

Fearon, of course, goes on to list three reasons players might do so, but these are tangential to the conceptual exercise. The fundamental insight – the framing of the problem of war – remains the crucial contribution irrespective of whether we find that a particular cause he identified is not as robust as might have appeared initially (Leventoglu & Tarar, 2008). The insight has generated a lively literature on the causes of war, and has led to generalizations that show how the same basic commitment problem due to large rapid power shifts underpins explanations heretofore thought to be distinct (Powell, 2006).

Even disagreements with the conceptualization enhance the value of the insight because it allows one to pose the alternative very clearly. For example, one might disagree that war is always costlier than peace. In that case, Fearon's puzzle would not longer apply since players would be choosing war because it is the most efficient way to divide the benefit. This might seem uninteresting until one asks what might make war, which we all agree is costly and risky, a more efficient dispute resolution mechanism than peace (Coe, 2011; Slantchev, 2012). Whatever interesting answers this line of inquiry provides, its indebtedness to the original puzzle is beyond doubt.

Just as important as choosing *what* to include in the model is choosing *how* to represent it. This is sometimes obvious with physical objects but far less so with concepts. For instance, we can model money, factory outputs, legislative votes, and so on in a way that is probably clear and uncontroversial. But what about "power" or "uncertainty"? How does the model represent power? And even if we agree that uncertainty is properly represented through some probability distribution, we still want to know, distribution over what? Clearly, we wish to be more precise with the meaning of these concepts, and this is what models do: they make abstract ideas more concrete by specifying some particular form they can take. Models can render "power" in many different ways, and we can argue which rendition is more useful than others. But we will agree on the specific way "power" is being represented (when analyzing its effects).

One model is supposed to represent real arrangements in the world (representative) but the other (interpretive) is designed to render abstract concepts (power, institutions, justice) more concrete. So our *theory* might make a use of a concept, say *power or justice*, and then we can have a variety of models that articulate what this concept might mean. Instead of abstracting from reality, as often claimed, models move us closer to reality but taking abstract concepts and making them more concrete/specific so we can argue about them and use them (Johnson, 2017, p. 45).

Take, for instance, "uncertainty". Usually, it is introduced and then a bunch of things that potentially relevant to the payoffs and that one might be uncertain about are enumerated. These end up being lumped together in non-formal theorizing. Since formal models force one to be specific, one usually picks among the possible things: distribution of power, costs of fighting, interest at stake, and so on. We used to do this without thinking too much about what is being picked, and as a result it was often done for convenience: pick the variable to be uncertain about such that it is easier to solve the model. That is, until Fey and Ramsay (2010) showed that it does matter what one is uncertain about. Purely conceptual development, and an important one at that.

The only thing we can get from models is the urge to think about what they try to make specific. Like fables, they translate the abstract into concrete, the general into the particular. They are not merely illustrations; rather they reveal something "not sufficiently recognized" about the general by showing it in a particular instance. There's a whole models-as-fables approach for this. The point here is that you don't "test" what the model tells you. You think about it. You could derive some fairly broad lessons that you might want to apply to real world analysis. They help you appreciate, for instance, how uncertainty might operate on incentives, so when looking at cases you would be cognizant of that possibility. From this view, the perennial "problem" of multiple equilibria – indeterminacy, which leads to ambiguity – the thing that generations of game theorists have tried to solve, might not be a problem at all. It could be that the moral is ambiguous. It could be that the real world admits a variety of solutions.

The bottom line here is that one should not always think of models as abstract simplifications of reality. One should also think of them as concrete representations of abstract concepts.

3.2 Telling stories while keeping them simple

Game theorists use models as simplified versions of life that are meant to clarify some of the logic of life's dilemmas, just as people everywhere use stories to develop new perspectives on important social problems. People regularly tell stories to help themselves in understanding society and its problems. To understand an international crisis, we might seek useful analogies by retelling, for example, the story of the 1938 Munich appeasement (to justify resolve) or the 1962 Cuban missile crisis (to justify restraint). … Models in game theory are just stories of another kind (Myerson, 2006).

It seems terribly non-rigorous, even flippant, to suggest that models are just stories. Or, rather, that they enable us to tell stories. But this is precisely their most valuable role. What is the point of abstraction and/or reification that models offer if not to provide the structure of an argument we wish to make and aid us in its clear exposition? A good model captures the elements of the story and lays bare the workings of the mechanism. The best model does this unfussily, with the minimally necessary assumptions to get things going. For me, the best models are as simple as they have to be but not simpler. One can go wild with complexity, sometimes to the point of having to resort to computer simulations or numerical techniques to solve the model, but I prefer to go in the opposite direction and keep stripping the model of features until any further simplification would not let me tell the story I wish to tell. Sometimes (very often, actually) the process of rebuilding the model using different assumptions teaches me something about the story itself, and sometimes I find out that the story I wanted to tell is not the one that I end up telling. Sometimes I discover that my intuition was incorrect, and the model can show me precisely where I had gone wrong. Sometimes I discover that while my intuition was right about the behavior, I had not quite grasped the reasons for it. Models help with all of that, so they are the great disciplinarians of my thinking.

This is an appropriate point for a confession: for me, modeling is a kind of a crutch, an aid in reasoning logically through complex interactions. I use models because I cannot see my way through without them. Some people do not seem to need models for that – Tom Schelling and R. Harrison Wagner come to mind – but I am not among them. I like building models, and I like interpreting them. In that sense, I am most interested in the conceptual problems of how to represent the situation I wish to analyze, and how to bring the insights of the model to bear on the problem. I am not very fond of actually solving models. Scratch that. I really do not like solving them. But I don't outsource the analysis to smart graduate students. That's because it is a necessary evil: when you do not understand how the analysis works, you do not really understand what the model is telling you. Oh, I can usually summarize the logic for an audience, but that's only after I have properly understood it myself. For me, the proof of the model is in the… proof. And it is often invaluable to attempt solutions and fail because this teaches you how the model works. Models have the capacity to surprise me, and it feels good when they are smarter than me. Learning that my intuition was not quite right makes the pain worth it.[14]

This is why I insist on simple models. Many scholars deride simple models, presumably because they fail to capture reality adequately. (Or maybe because they don't demonstrate one's modeling chops.) This is a strange notion since all models, yes, all of them, without exception, make silly simplifying assumptions that nobody can possibly relate to the real world. Some silly assumptions have become normalized through repetition and through their usefulness in permitting us to build even more models that we can solve. In IR, this would be the two unitary actors assumption that is ubiquitous and rarely challenged by modelers. But one must not forget that this assumption is really quite silly if one's goal is to approximate the real world where the sphere of international relations is populated with a bewildering diversity of actors from almost any level of aggregation. It is not, however, silly if one wishes to illuminate some strategic aspect of an interesting situation. It also happens to be sufficient to demonstrate a variety of important mechanisms (e.g. the various rationalist explanations for war) that would also occur in models that dispense with the assumption, but where the complexity would tend to obscure the fundamental incentives so starkly revealed by the simple models.

Minimalist models are also very useful in communicating the insight to others. This is where formal models can really outperform a story told with just "plain" words. Models make communication more precise because they demand specific definitions of the concepts they utilize. We might disagree whether the particular way a model represents some concept is the most appropriate way to do it but we will not disagree on what the representation is. There are tremendous benefits to this precision that cannot be had with lengthy definitions of terms that invariably accompany many works using "plain" words precisely because many of these words are anything but plain. I have already given an example with the formal definition of rationality, but one can point to others. For instance, it was formalism that eventually led us to the realization that many seemingly disparate problems – instability produced by first-strike advantages, dynamic commitment problems, bargaining over things that influence future bargaining power, and even changing power relationships between domestic political factions – actually reduce to a simple mechanism that generates inefficiencies due to large rapid power shifts between the actors (Powell, 2006). Without the precise definition of the concepts in the existing models it is very doubtful that Powell would have realized that they all relied on a common mechanism.

As an amusing aside, I also prefer simple models because I find them more realistic. This might sound like an oxymoron until one considers the fact that most of our IR models are attempts to represent incentives for

behavior of individuals. Even when we talk about actors at high levels of aggregation, the stories we tell usually involve highly placed policy-makers. But if we are trying to understand these incentives, we are essentially trying to reason like these policy-makers. I very much doubt that people, even highly placed policy-makers, have particularly complicated internal models of the world, are able to consider numerous variables, or reason through convoluted strategic situations. More likely, they use simplifying shortcuts like the rest of us. If that is so, then adding bells and whistles to the model is actually moving it further away from the reality it is supposed to represent. Complexity might, in fact, make them worthless.

4 Conclusion

Learn to love simple models. Remember, simplicity, sometimes even of the silly variety, can be a virtue when it comes to clarity, and clarity means reliable communication, which in turn produces replicable arguments that (hopefully) lead to persuasion.

And lay off me with demands for empirical tests of models. Really. Lay off.

Prepared for the NEPS Lecture at the 17th Jan Tinbergen European Peace Science Conference, University of Antwerp, Belgium, 26-28 June, 2017.

Whatever it is, I'm against it.

Groucho Marx, in *Horsefeathers*

Notes

1 Cartwright 1999; 2010; Johnson 1996; 2017.
2 Milgrom & Stokey, 1982. This is related to Aumann's (1976) famous result that it is not possible for Bayesian players to agree to disagree. That is, players who start with private information but update their beliefs based on new information using Bayes rule cannot have different posterior beliefs when these beliefs are common knowledge.
3 Cited in Postrel (1991), who constructs a signaling game to make the points I am making here.
4 Huth and Russett (1984). The description of the coding is on pp. 509–510. I suspect the reason for the misinterpretation of this variable is that this is the only place where its coding is actually described. All subsequent papers that use the dataset refer to this article for description of the variables, and so subsequent researchers seem to have trusted the label much more than they should have.
5 Unless, of course, you argue that for some reason this measure is a good proxy to the true local balance of capabilities. But then another problem arises: if the local balance is strongly correlated with whatever this variable measures, then there cannot be any new information conveyed by it. It should not have the anticipated effect either.
6 Another limitation of the Nash definition of rationality and its progeny is that it is indivdualistic: one only considers deviations by a single player while holding everyone else's strategies constant. In some settings it might be appropriate to consider a group of players deviating jointly, and further demand that such group deviations are credible in the sense that no player would individually choose to break the agreement to deviate. This would restrict the set of Nash equilibria even further. In some cases it might leave it completely empty.
7 Another widespread criticism of game theory is that it assumes that people are driven by materialistic (or, even more crassly, monetary) rewards. This one is complete nonsense. Anything can go into preferences. Anything.
8 These are for choices over certain outcomes. When there is risk involved, the requirements become more stringent because they essentially involve imposing completeness and transitivity on preferences over lotteries: actions with uncertain outcomes.
9 The assumptions are (1) completeness and transitivity, (2) continuity – given some lottery and two others, one that is better and the other worse than the first, we can construct a compound lottery of the two such that the decision-maker is indifferent among it and the original one; (3) independence – adding the same outcome with the same probability to two lotteries should not alter the ranking of the lotteries. The assumption of continuity rules out lexicographic preferences. Although these orderings might be useful for organizing dictionaries, it is not clear to me that people actually have them in real life. The innocuously-looking assumption of independence, on the other hand, might be violated empirically because people are truly awful at dealing with compound lotteries and might not notice equivalent parts they would ignore (Allais Paradox).
10 I am going to ignore the additional requirement of iterated knowledge that allows to process to unfold, and assume common knowledge of our definition of rationality among the players.
11 I am not aware of a definition of rationality in game-theory that would select anything other than the mutual defection outcome in the single-shot Prisoner's Dilemma. This seems to suggest that there is some agreement about what minimally rational behavior should look like – no strictly dominated strategies.
12 Of course, if one found no evidence to support the hypothesis generated by the model, then one could reject the model. Ironically, this is precisely the type of research that does not get published all that often.
13 Private information with incentives to misrepresent causing war shows up in Morrow (1989). A very general result that shows why stronger actors must run higher risks of war to obtain better peace deals because of incomplete information appears in Banks (1990). The literature on bargaining inefficiencies due to asymmetric information in economics was already quite large, as was the literature on

commitment problems wreaking all sorts of havoc (often under the moniker "dynamic inconsistency"). The third cause, indivisibility, is mechanical rather than strategic – it just asserts that there is no way to partition the benefit to satisfy the players' minimal demands because some partitions are simply impossible – and so not very illuminating as a cause (it is illuminating if one wishes to consider how actors make things indivisible).

14 See, for instance, the section on making mistakes in Varian (1997).

References

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics*, *4*, 1236–1239.

Banks, J. S. (1990). Equilibrium behavior in crisis bargaining games. *American Journal of Political Science*, *34*(3), 599–614.

Blainey, G. (1988). *The causes of war* (3rd ed.). New York: Free Press.

Bueno de Mesquita, B. (1983). *The war trap.* New Haven: Yale University Press.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science.* Cambridge: Cambridge University Press.

Cartwright, N. (2010). Models: parables v. fables. In R. Frigg & M. C. Hunter, (Eds.), *Beyond mimesis and convention: Representation in art and science* (pp. 19–32). Dordrecht: Springer.

Clark, W. R. & Golder, M. (2015). Big data, causal inference, and formal theory: Contradictory trends in political science. *PS: Political Science & Politics*, *48*, 65–70.

Coe, A. (2011). *Costly peace: A new rationalist explanation for war*. Los Angeles: Department of Political Science, University of Southern California.

Fearon, J. D. (1994). Signaling versus the balance of power and Interests: An empirical test of a crisis bargaining model. *Journal of Conflict Resolution*, *38*, 236–269.

Fearon, J. D. (1995). Rationalist explanations for war. *International Organization*, *49*, 379–414.

Fey, M. & Ramsay, K. W. (2007). Mutual optimism and war. *American Journal of Political Science*, *51*, 738–754.

Fey, M. & Ramsay, K. W. (2010). Uncertainty and incentives in crisis bargaining: Game-free analysis of international conflict. *American Journal of Political Science*, *55*(1), 149–169.

Fey, M. & Ramsay, K. W. (2016). *Mutual optimism in the bargaining model of war*. Princeton, NJ: Department of Politics, Princeton University.

Harsanyi, J. C. (1977). *Rational behavior and bargaining equilibrium in games and social situations.* Cambridge: Cambridge University Press.

Harsanyi, J. C. & Selten, R. (1988). *A general theory of equilibrium selection in games.* Cambridge, MA: MIT Press.

Huth, P. & Russett, B. (1984). What makes deterrence work? cases from 1900 to 1980. *World Politics*, *36*, 496–526.

Johnson, J. (1996). How not to criticize rational choice theory: Pathologies of 'Common Sense'. *Philosophy of the Social Sciences*, *26*, 77–91.

Johnson, J. (2017). *Models-as-fables: An alternative to the standard rationale for using formal models in political science*. Rochester, NY: Department of Political Science, University of Rochester.

Kreps, D. (1990). *Game theory and economic modeling*. Oxford: Oxford University Press.

Leventoglu, B. & Tarar, A. (2008). Does private information lead to delay or war in crisis bargaining? *International Studies Quarterly*, *52*, 533–553.

Milgrom, P. & Stokey, N. (1982). Information, trade, and common knowledge. *Journal of Economic Theory*, *26*(1), 17–27.

Morrow, J. D. (1989). Capabilities, uncertainty, and resolve: A limited information model of crisis bargaining. *American Journal of Political Science*, *33*(4), 941–972.

Myerson, R. (2006). *Force and restraint in strategic deterrence: A game-theorist's perspective*. Chicago: Humanities Festival on Peace and War.

Osborne, M. J. & Rubinstein, A. (1994). *A course in game theory.* Cambridge, MA: MIT Press.

Postrel, S. (1991). Burning your britches behind you: Can policy scholars bank on game theory? *Strategic Management Journal*, *12*, 153–155.

Powell, R. (2006). War as a commitment problem. *International Organization*, *60*, 169–203.

Signorino, C. S. & Tarar, A. (2006). A unified theory and test of extended immediate deterrence. *American Journal of Political Science*, *50*, 586–605.

Slantchev, B. L. (2011). *Military threats: The costs of coercion and the price of peace.* Cambridge: Cambridge University Press.

Slantchev, B. L. (2012). Borrowed power: Debt finance and the resort to arms. *American Political Science Review*, *106*(4), 787–809.

Slantchev, B. L. & Tarar, A. (2011). Mutual optimism as a rationalist explanation of war. *American Journal of Political Science*, *55*, 135–148.

Varian, H. R. (1997). How to build an economic model in your spare time. *American Economist*, *41*(2), 3–10.