Linking Theory and Empirics: The importance of research communities.

Scott de Marchi, Duke University

> "The field covered in this book is very limited, and we approach it in this sense of modesty… Its first applications are necessarily to elementary problems where the result has never been in doubt and no theory is actually required. At this early stage the application serves to corroborate the theory. The next stage develops when the theory is applied to somewhat more complicated situations in which it may already lead to a certain extent beyond the obvious and the familiar. Here theory and application corroborate each other mutually. Beyond this lies the field of real success: genuine prediction by theory. It is well known that all mathematized sciences have gone through these successive phases of evolution" (p. 8, Von Neumann and Morgenstern).

Abstract

In this essay, I argue that cumulative knowledge is possible in the social sciences, but that our research culture is limiting our efforts. In part, we are hampered by our inability to coordinate on shared questions; this problem is worsened by claims of methodological exceptionalism which limit interaction between empiricists and theorists. We have also largely ignored the defining characteristic of much of the progress in the natural sciences: the ability of theory to make sharp *predictions* about the world.

Introduction

There are two hard facts that define our existence as social scientists. The first is that we have not seen the desired accumulation of knowledge about economic, political, and social phenomena. The second is that we must keep publishing anyway at an increasingly breakneck speed, if we wish to continue to enjoy the perquisites of employment at universities. Most researchers are, however, well-meaning and would like to make progress on the key substantive questions that motivate our research. What

are the causes of and remedies for poverty, racism, and violence? How are preferences aggregated by institutions into outcomes -- and, are these outcomes any good for the bulk of people affected by them?

A corollary of all this is that we tend to argue, as a profession, about epistemology – certainly more than the natural sciences to which we often (and unfavorably) compare ourselves. My impression is that most natural scientists spend their time going about the business of producing science and are convinced (probably rightly) that they are making "progress". For social scientists, our concern with epistemology is, however, more than a little like arguing about pizza toppings. Tastes vary even though the toppings themselves have been consistent for some time and no set of toppings can or should make everyone happy. Because of this, it is difficult to say anything new and it is easy to descend into a debate about whether pineapple belongs on pizza.[1]

That said, while the core arguments in epistemology have remained fairly static, there have been changes in the discipline that make it worthwhile to reconsider some of these arguments. In particular, the identification revolution has substantially affected the practice of political science. Empirical work, especially models that examine the causal effects of a single treatment variable in (natural) experiments, is ubiquitous. Theoretical work and the consideration of strategic behavior is on the wane.

I would like to make three points in this essay. The first is that there is the possibility of making cumulative progress – and perhaps more important, we should focus on prediction as a standard for evaluating research. The second, echoing Deaton (2010), is that cumulative progress is unlikely without more of a focus on generalizable theory than is currently the norm in political science. The third is that there are no perfect papers. Rather, we need to make progress as a community and accept that research about humans is a difficult endeavor.

The possibility of progress

[1] It does not.

Central to any definition of cumulative science is the notion of progress. Gerring (this volume) emphasizes that cumulative knowledge can be characterized as solving problems and that the best evidence of progress is that older, less useful solutions are no longer cited in the literature. Callis et al. (this volume) focus on the generalizability of treatment effects; progress is, for example, replicating the same empirical model and gathering evidence on a treatment effect in different contexts.

Obviously, some heterogeneity exists in how we define progress. This is likely due to the varied methodological commitments of the research communities that exist in the social sciences. That said, almost all definitions of progress refer to the natural science model where there is considerably more confidence that progress exists.

It is not surprising that many social scientists question whether or not progress is possible in the social sciences. Formal theorists, in particular, have raised doubts about progress and the utility of formal models in generating testable hypotheses:

> "I believe that as an economic theorist, I have very little to say about the real world and that there are very few models in economic theory that can be used to provide serious advice… We organize our thoughts using what we call models. The word 'model' sounds more scientific than 'fable' or 'fairy tale' although I do not see much difference between them… As in the case of a good fable, a good model can have an enormous influence on the real world, not by providing advice or by predicting the future, but rather by influencing culture. Yes, I do think we are simply the tellers of fables, but is that not wonderful?" (Rubinstein, p. 882, 2006).

Empiricists, while more hopeful that progress might be achieved, remain skeptical about the current level of progress in solving canonical substantive problems:

> "*Under ideal circumstances*, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us 'what works' in development, to design policy, or to advance

> scientific knowledge about development processes… Randomization is not a gold standard because 'there is no gold standard'" (Deaton, 2010).

It is worth noting that the above quotes are not from critics of quantitative social science or armchair philosophers – progress is an uncertain thing even for the most accomplished theorists and empiricists.

Underlying this debate on the possibility of progress is a definition that is considerably sharper and more burdensome than many might like. Progress is not an accumulation of unorganized "facts" or evaluations of individual programs from idiosyncratic contexts. Nor is it untested (and perhaps untestable) theories that have no mapping to empirical phenomena. If we rely on the natural sciences as our beacon, then progress depends crucially on prediction.

A sharper definition of progress in a cumulative science thus hinges on two related standards:

i. Ability to predict phenomena of interest
ii. Ability to perform (or evaluate the effects of) interventions in new contexts

It is worth nothing that (i) is a weaker standard than (ii) in the sense that prediction may be possible without understanding the mechanisms that would allow for meaningful policy interventions (de Marchi and Stewart, 2020). I also define "prediction" in the broadest sense here – it is not limited to predicting future events or large-N samples. Rather, prediction concerns testing a hypothesis derived through theory about empirical outcomes that (crucially) are not already known. For example, a theory that generations novel empirical implications can be tested with a handful of cases.[2]

There are social scientists who have embraced prediction as a standard. As noted in de Marchi (2005) and Ward et al. (2013), prediction serves a vital role in testing theory and

[2] And within a fixed sample, relying on k-fold cross validation to test a hypothesis, for example, can be considered prediction.

is an alternative approach to establishing causal relationships. If one commits to a theory *ex ante* and tests this theory out-of-sample, one is making a bet on the outcome. As one repeatedly tests a theory in novel samples, evidence accumulates that the relationship is causal. One can, of course, still be wrong in any number of ways, but the presence of a well-defined theory and the existence of repeated, out-of-samples tests is one approach to building cumulative knowledge.

Support for prediction as a standard is not, however, widespread in the social sciences. I think it is fair to say that while these standards describe the goals of the natural sciences, the difficulty in creating predictive models in the social sciences helps explain why many are suspicious about the merits of predictive work and others are gloomy about the possibility of progress. By and large, relatively little research has been able to predict substantive variables of interest, nor have we been able to conduct (policy) interventions in multiple contexts with certainty about the outcomes.[3] Many researchers are also skeptical of predictive work because of the existence of black box models that may predict with some accuracy but fail to illuminate the causal mechanisms involved.

One possibility is that social scientists largely avoid emphasizing prediction in our debates over epistemology because we are bad at it; given that we must keep publishing regardless, that is an uncomfortable situation to be in. But we may also avoid doing predictive work because alternative definitions of progress (as well as questioning the possibility of cumulative knowledge) protect the status of existing research and methodological approaches.[4] What to do?

My belief is that the hostility toward prediction is misplaced. In the natural sciences, prediction is both central *and* tightly connected to theory – there is no shortage of causal mechanisms that are tested with predictive models. The Eddington "experiment" serves as a shining example of how a theory (general relativity) generates sharp predictions that can be tested *observationally*. It is worth emphasizing that Eddington's test of

[3] An alternative explanation is that not many political scientists try to do predictive work. As noted elsewhere in this volume, we reward novelty and methodological innovation and are not always focused on achieving progress on substantive questions.

[4] For example, Johnson (2021) mirrors Rubinstein and question whether one should test formal models. He does, of course, believe in publishing formal theories even if they cannot be tested and does not adequately address the "many models" problems.

general relativity was purely observational; there was no experiment in the sense of treating a subset of units and holding the others constant. The idea, which is commonplace in the social sciences today, that the only way to establish causal relationships is by experiments is contradicted by numerous prominent examples in the natural sciences.

If we try and fail to build predictive models because such an endeavor is impossible in the social sciences – and that all we are capable of is producing fables and context-dependent empirical one-offs – we still lose nothing by committing to a natural science standard.[5] It is, of course, important to recognize when social science succeeds or falls short and clear standards based on prediction can alert us to these outcomes. To the extent we care about the substantive problems mentioned in the introduction to this paper and seek to give advice to real-world policy makers, it is important to continue to work toward replicable, prediction-centered research. Fables and empirical one-offs are not adequate to the task of changing policy.

In the remaining sections, I will explore the relationship between theory and empirical work and then turn to the reasons why we should focus on building scholarly communities focused on prediction.

Theory and cumulative science

Theory serves in two distinct ways in making cumulative science possible.

Measurement requires theory.

The first is underappreciated but vital to the scientific enterprise and concerns measurement. Simply put, if we want to examine / predict the effects of treatment variables in multiple contexts, we need to believe that a) the treatment variable is measuring the same empirical phenomenon across contexts and b) we understand what the measure is capturing. This can, however, be quite difficult.

[5] To quote the bards, "Don't stop believin'".

Example 1: Power. To take one salient example, a great deal of research in the social sciences concerns power. In the context of bargaining, a naïve understanding of power would be to collect data on the number of seats each party has in the legislature or the number of votes (or supporters) an actor has available in a given choice context. Experimental research has used power as a treatment variable but without considering theoretical work on the nature of power. One distinction with a long tradition in the bargaining literature is that pivotality is the best measure of power. E.g., if three actors have {10, 10, 1} seats / votes, then they are all equally "powerful" because each has the potential of forming two minimum winning coalitions. This is obviously an important distinction; if one were to simply use the "raw" integers counting seats, one would arrive at the (potentially) misleading belief that the first and second actors had ten times the power of the third actor. One would also miss the fact that many seemingly different settings are all in the same equivalence class; e.g., {5, 4, 2} or {1, 1, 1} are essentially the same strategic context.

This has large implications for the experimental literature. Diermeier and Morton (2005) notably use the following treatments which (for them) represent different levels of power:

Treatment 1: {34, 33, 32}

Treatment 2: {49, 33, 17}

Treatment 3: {46, 44, 9}

In these treatments, each actor has exactly the same pivotality and all three map to the same underlying minimum integer weight representation of {1, 1, 1}. Simply put, these may not be different treatments! This problem is compounded when one considers their seemingly innocuous choice in their experiments to base recognition probabilities in the bargaining game on the raw proportions from the above treatments.[6] This is a

[6] One could argue that the treatment concerns the recognition probabilities and is not focused on power. This is worrisome, though, in the sense that the setup used in this paper has a number of different choices that together interact in very unexpected ways. The version of the Baron-Ferejohn game implemented by Diermeier and Morton uses raw seats, recognition based on these (rather than the more common assumption in the literature of uniform

weighty choice in terms of producing their expectations for the experiment, but should recognition be based on pivotality? Should it be a uniform draw? Or are there institutions which determine which actors "go first"?

In experiments of this type, there are a number of choices that, while seemingly innocuous, nonetheless have a large effect on the results. We can't as a consequence easily compare the results in one study to any other and even relatively straight-forward findings (e.g., subjects rely on a fairness norm that ignores the raw vote share) might be the result of a subjects relying on a pivotality based understanding of power.

Example 2: Extremism. There are multiple studies in the close elections literature focusing on attributes of candidates in democratic systems. Hall (2015) examines the question of what the electoral consequences are for electing an "extreme" candidate in primaries in the United States. This raises the obvious question: given two candidates in a primary, how do we measure which one is the "moderate" and which is the "extremist"? Deriving a measure of these concepts is not straight-forward and requires a theoretical statement of the dimensionality of the underlying policy space. Hall's definition does not engage with this complexity and relies on the distance between the two candidates in a single-dimensional latent space.

Hall's approach to measuring relative extremism means that the following two cases are treated as equivalent in terms of the treatment:

1. two Republican candidates at $x_1 = .9$ and $x_2 = .6$; x_1 is an extremist
2. two Republican candidates at $x_1 = .15$ and x_2 at $-.15$; x_1 is an extremist

recognition), and a finite horizon game with five stages. The end result is spectacularly counter-intuitive – at each stage the continuation values of the players might change are (worse) players with more seats (and more power) might not have higher continuation values. The sample Diermeier and Morton used was undergraduates, but even if it were researchers working in bargaining theory, their expectations about this bargaining game would be violated by this experimental design.

Hall also includes a mix of very different primaries by including all cases that have a nearly even split of the vote between the "top two" candidates. This is reflected in the score variable which is defined as "the extremist candidate's share of the top two candidates' vote less 0.5" (Hall, p. 23). This means that California's 1982 district 43 Republican primary is included in the sample, which had 14 candidates with vote shares of [.166, .165, .158, .148, .106, .052, .048, .036, .028, .02, .015, .013, .01, .01] as well as more "ideal" elections where exactly two candidates split the primary vote 50/50. As noted above, our expectations about candidate behavior and what constitutes extremism would be very different in the 1982 California example than in the more straight-forward case.

It seems obvious that these problems require an examination of the data, which in this research is easy given the small sample size (N=83). Let us focus on the first issue: where are the "extremists" in Hall's model located in terms of ideology compared to the moderates? It is immediately obvious that a problem exists. One third of the 83 cases feature a primary race where one candidate is to the left of 0 and the other is to the right. Imagining that these are same party contests featuring a moderate and an extremist does considerable harm to the data; the races in fact feature two moderates. The 1994 Democratic primary in South Carolina is an example of these cases. A conservative Democrat had just retired and Tommy Moore and Jim Bryan – both even more conservative (i.e., moderate) Democrats – vied to replace him in an increasingly Republican district. The ideology scores for these two candidates in Hall's data place them at .07 and -.05 respectively, with Bryan (the winner) treated as the extremist in the race. Replicating Hall's regression with these cases excluded, the treatment variable for extremism is no longer significant.

Coding "power" or "extremism" are difficult problems. And, at root, they are *theoretical* problems. All of the care and attention that goes into setting up a laboratory experiment or regression discontinuity design is important, but it depends on a theoretical exercise that deserves more attention. In both of the above examples, the lack of theory makes it difficult to know what is being measured and how the choices made in performing the experiment affect the results or the external validity of those results. This makes it very difficult to perform experiments in different settings and combine multiple experiments to produce cumulative knowledge. Simply put, if we are interested in power, or agenda

control, or extremism and want to make predictions about new contexts, it would be difficult to do so given the lack of clarity on measurement.

I also do not mean to suggest that theory will uniquely identify the "right" measure in all cases. It may, for example, be perfectly reasonable to use the raw seats a party has in a legislature or a relative measure of policy distance in a study. But these are weighty choices. In both cases, substantial theoretical work exists to help researchers make judgements about the best way to pin down treatments and make choices on other parameters involved in conducting an experiment. As a discipline, we should get away from the idea that one can simply "scrape data" and move forward without additional effort. Instead, we should pause and ask "What is power?" or "What is policy extremism?" in contexts of interest and have at hand a theoretical justification for the particular measurement strategy we adopt. And, part of the evaluation of any experimental piece should not just focus on whether or not the model is identified; rather, it should also examine the choices made with respect to measurement and the mapping of the experimental results back to a general theory.

Questions require theory.

The second role for theory is to help us ask good questions and pin down causal mechanisms. This is especially acute in contexts in which actors make strategic decisions, but the need for strong theory is broader than this. As noted in Jenke (2022), it is often the case that experiments identify a treatment effect, but do not uniquely specify the causal mechanism involved.

Example 3: Information matters. The Metaketa project (as discussed by Samii, this volume) is a prominent example of an attempt to build cumulative knowledge and in many respects, there is much to laud about their approach. Here we will focus on the first collection of studies which concern the role of information in elections (Dunning et al., 2019). On the one hand, it is certainly the case that the role of information in elections is a "big" substantive question that has long occupied political scientists. On the other, the studies in the Metaketa 1 volume focusing on the role of information in elections are undertheorized and ignore much of what we know about causal mechanisms from prior political behavior research. I.e., the hypotheses in Metaketa 1

are clear but simplified to a degree that makes it difficult to understand the mechanisms involved:

> Hypothesis 1: Positive information increases voter support for politicians.
>
> Hypothesis 2: Negative information decreases support.

The core idea is sound: treat subjects with information across a range of contexts to determine what the general effect of information is in elections. But, given differences in these contexts (including institutional rules, social networks, etc.), the treatments are possibly quite dissimilar. E.g., in Benin, subjects are shown a video with their incumbent's rate of attendance in the legislature, rate of posing questions, committee attendance, and number of laws considered by the committee.[7] In Mexico, the treatment is a non-partisan leaflet detailing the amounts of infrastructure spending that benefits the poor or is unauthorized spending. Other treatments in the volume are equally diverse.

Whether or not these are suitably similar treatments is again a theoretical question and deserves close examination. And from prior work in the behavior literature, one would have to determine whether these treatments are reasonable ways to measure the causal relationship information has with voter learning and decision-making during a political campaign. The idea that information matters in democracies is suggestive, but a considerable amount of prior and empirically validated theoretical work exists on the topic; almost all of it calls into question the idea that a single treatment of information of the sort found in Metaketa 1 – much of which might not relate to voter concerns – would affect outcomes.

From prior research, we have good evidence that voters are exposed to a huge amount of information, especially during political campaigns. But despite this, much of their decision-making occurs prior to the campaign and results in relatively static decision-making close to an election (the "causal funnel" from Campbell et al., 1980). Worse,

[7] Additionally, an attempt to modify salience is included by instructing respondents on how important these features of legislator behavior are.

many voters lack the sophistication to rationally update based on new information; others, especially the most sophisticated, engage in motivated reasoning to discard disconfirming information (the RAS model from Zaller, 1992). We also know that affect plays a heavy role in processing information and attitudes are updated as a complex function of prior political beliefs that exist in long term memory (the ACT-R based JQP model from Kim et al., 2010).

The overarching finding from prior research is the idea that single pieces of information are unlikely to change voter behavior (at least, to a degree that would be detectable). We know that information matters – at least, under some conditions. Given strategic behavior and the heterogeneity of institutional contexts, though, it is no small matter to examine the conditions under which different types of information might matter in different contexts. There's a good reason that Kim et al. exposed their subjects to information across an entire political campaign and Gilliam and Iyengar (2000) created entire newscasts to study voter behavior given the presence of cognitive misers and motivated reasoners. And there are equally good reasons why political psychologists built complex theories to account for voter learning.

It is likely that the urge to simplify treatments by the researchers involved with the Metaketa project was due to the decision to use field experiments in multiple contexts, with an eye toward comparability across these contexts. But given prior research on the role of information in elections, it is not surprising that the results of Metaketa 1 are largely null:

> "Our findings are clear and robust. Despite the statistical power gained by pooing the results of the experiments, and despite what we believed to be the timely, relevant, and accessible nature of the information on candidate or party performance, the effects of the common interventions are largely null" (p. 10).

And it is not as if we can actually trust these null results. Prior studies have found that information matters, but they have established highly conditional effects based on better specified theoretical models of human learning. Metaketa gets a lot right – it is a community working together on a shared target variable. The lack of engagement with

theoretical work on voter behavior, however, sharply limits our ability to learn from these studies or make predictions about the effects of information in new contexts.

Et tu Brute?

While the above examples point to empiricists neglecting theoretical work that could inform their work, it is also the case that too little effort has been made by theorists to bridge the gap and produce causal mechanisms that could be tested empirically.[8] Often, theorists see a stylized empirical fact – people cooperate – and then produce a theory that "explains" this fact without providing any guidance to the real-world variability that is of interest to empiricists. More concretely, the IPD suggests that cooperation is possible in a repeated game with either an infinite horizon or an unknown horizon – given the folk theorem, this is not much guidance to empiricists. A theory with some empirical teeth, in contrast, would actually tell us something about the (measurable) conditions under which cooperation is likely to occur – i.e., it would explain the variance we see in the real world and point toward concrete policy interventions that would promote cooperation.[9] We know that cooperation is possible and we also know that it is not; we need sharper theories than currently exist to guide empirical tests and policy interventions.[10]

The other issue here is that many deductively consistent theories can explain the same outcome; picking parameters to achieve a particular result is always possible when that outcome is known *ex ante* (de Marchi, 2022). The Empirical Implications of Theoretical Models program (https://www.nsf.gov/sbe/ses/polisci/reports/pdf/eitmreport.pdf) was founded to help bridge the gap between theorists and empiricists, but no simple

[8] One key distinction between most of the theoretical work in the natural sciences and the theoretical work in the social sciences is that the natural science theories both accept inputs from and provide outputs to measurable phenomena in the real world.

[9] Explaining variance, rather than providing an existence proof, is hugely important. Consider Ostrom's (1990) groundbreaking work on common pool resource problems. She demonstrates that humans have efficiently solved these problems in a variety of contexts and derives, largely based on qualitative case studies, some lessons that might be useful when thinking about future problems. But, her hope is that better theory will someday be generated: "What is missing… is an adequately specified theory of collective action whereby a group of principals can organize themselves voluntarily to retain the residuals of their own efforts."

[10] The same issues apply to reputational / signaling games along the lines of the original work of Kreps and Wilson (1982). This class of formal model establishes a mechanism – based on actor types with varying payoff functions – that show the existence of reputational and signaling effects. But, they largely do not explain real-world variance in applied settings.

solution exists to the divide between theorists and empiricists! In fact, given the increased publication demands faced by younger scholars and the increasing sophistication and training requirements of empirical work, my dismal expectation is that theory will decline in importance over time. Simply put, there are already too many demands on junior scholars and intellectual breadth – while necessary to produce cumulative science – will be sacrificed unless the discipline changes our current research culture.

Research Communities

As the above discussion indicates, my main argument is that either community – empiricists or theorists – working alone is unlikely to produce cumulative science. On the empirical side, I am skeptical that simplistic, single-variable treatments will illuminate questions that are of substantive interest to political scientists. Callis et al. point to the large (and growing) literature on close elections, for example, where much of the focus is on the power of partisan and personal incumbency. At the risk of some impropriety, one has to ask if pinning down the exact level of incumbency advantage should be a central focus for political science. Or, is this growing literature a convenient application of regression discontinuity designs?

It is worth remembering that each natural experiment in this area of research is effectively an *n* of 1 (or a small sample) – it occurs in a very particular context. To provide cumulative knowledge, the results of these experiments would need to connect to theory and span multiple contexts to derive any sort of general result on the causal mechanisms responsible for the incumbency advantage. To do this, we must resort to strong assumptions to believe that this research generalizes from the samples used (which focus by design on close elections) to elections more generally (De la Cuesta and Imai, 2016). This assumption is problematic, especially in the United States, where the vast majority of elections in the population of interest are *not* close elections (and often, feature a candidate that runs unopposed, thereby changing the strategic incentives).

This sort of endeavor – connecting separate empirical studies to a larger theoretical argument – is clearly possible though and we should be optimistic as a result. The

important lesson to draw from these exemplars is that it is often the combined efforts of empiricists and theorists working together that produces progress.

One exemplar is Titiunik's (2009) finding that mayoral elections in Brazil had incumbents who were at a disadvantage. Given the rest of the empirical studies on the topic had produced strong support for a general incumbency advantage, this was a surprising result. Klašnja and Titiunik (2017), again using Brazil as a starting point, produced a more general theory about the (interactive) institutional effects of term limits and weak parties on incumbency. Then, to test the main implications of their theoretical argument (i.e., that term limits in the context of a weak party system change the horizons and performance of incumbents), they examined new empirical cases that varied over the main parameters established by the theoretical model.

Crucially, however, this sort of work takes time and diverse skillsets, both to perform the empirical studies, generate a general theory, then perform even more empirical studies to span the theoretically interesting parameters of the theoretical model. Progress occurred because the researchers involved took the time to iterate between theory and empirics and the main threat to this sort of progress is the temptation to simply mass-produce purely empirical or purely theoretical studies.

We need empiricists and theorists talking more than they do now. Part of this shared endeavor, though, hinges on the realization that research about human behavior is hard. While efforts to ensure that each individual article is "correct" are worthwhile (e.g., pre-registration, editorial replications of the results in papers, etc.), my view is that papers will make mistakes and that strategic behavior by goal-seeking academics will undermine most of these efforts. Instead, we should focus on progress made by a community of scholars focused on the same substantive questions.

Even relatively simple studies are quite complex as the debate over Gerber and Green's (2000) get out the vote study illustrates (Sekhon, 2009). We need to understand that even our best scholars will make mistakes and that (on top of methodological issues) the world is a noisy place to conduct research. In my mind, we do not gain much from obsessing over single studies; rather, we should try, whenever possible, to make

theoretical commitments to study important substantive questions, agree on target variables to measure the progress of our research, and have repeated tilts at the windmill with teams that include theorists and empiricists working together. We will, all of us, individually make mistakes, but the possibility exists that the research community over time will generate progress.

Conclusion

Von Neumann and Morgenstern were acutely aware that progress in science depends on linking theory to empirics and that prediction was central to the exercise. Much of the framework of game theory was created to solve simple games so that we could build intuition that would help in creating more complex theories that might apply to the much more complex real-world games we see in economics and politics. Unsurprisingly, given the incentives to publish, we have had very little patience for this sort of endeavor and have doubled down as formal theorists on solving simpler games – in most cases, even simpler games than Von Neumann and Morgenstern proposed initially.[11] For Von Neumann and Morgenstern, games like poker were good targets. But, social scientists discovered that poker, at least as played by actual humans, was much more complicated than it initially appeared. Accordingly, game theorists simplified their models of poker until the games under consideration bore little resemblance to the games played by human poker players.

In contrast, computer science as a discipline committed to studying poker *as played by humans in the real world*. By mixing computational methods with game theory, they have succeeded in generating theories – with spectacular empirical results – that solve Texas Hold 'em and illuminate canonical human behavior like bluffing (Billings et al., 2003; Brown and Sandholm, 2019). My argument is that if we committed ourselves to a similar stance on epistemology – that theories and empirical work should be closely linked and prediction should be the ultimate goal – we might achieve similar successes in the social sciences.

[11] It's thus not surprising that even more complex human behaviors – like forming a government in parliamentary systems – has resisted our best theoretical efforts.

What can we do to solve the problems that exist with our research culture? First, as noted above, we should abandon the notion that an article can be "correct". Rather, our progress will be fitful and we should expect that results are frequently challenged and replaced by newer research. We will never have a perfect theory that covers all cases, nor enough data to illuminate all possible mechanisms, nor the perfect experiment that maintains external validity. Preregistration and other costly mechanisms are well-intentioned, but easy to defeat with strategic behavior.[12] Instead, we should embrace the idea that we exist in a research community and that the community will make progress over time. If we care about progress, then the main corrective will be later researchers critiquing and building on the efforts of those that preceded them.

Second, we should agree on a set of shared dependent variables whenever possible. Cumulative knowledge without shared benchmarks seems to me to be a very uncertain enterprise and not descriptive of the periods in which the natural sciences have made progress.[13]

Third, we should slow down, collaborate more than we do, avoid methodological purity tests, and create teams of theorists and empiricists working together to solve substantive problems. To the extent we aspire to a natural science ideal of cumulative knowledge, we need question-focused, heterogenous teams (Page, 2014). Human behavior is complex and we do not have the luxury of looking for low-hanging fruit. Because of this complexity, we must be ecumenical with respect to methods and accept and incorporate evidence from a variety of approaches. Too often, scholars focus on one approach and then make exceptionalist claims – e.g., "the only way to determine causality is via experiments" or "theories are deductively true, they shouldn't be tested".

[12] The trend toward "robustness tests" and ever-expanding online appendices in journal articles should, however, stop immediately. Once a result is known and given the huge incentives to publish, almost all authors will magically demonstrate that the robustness tests support their original result.

[13] Consider the Copernican revolution. The existence of widely accepted measures (i.e., using direct observation or later, telescopes) and an agreed upon target (i.e., the movement of celestial bodies) was crucial in making progress. This progress was not, of course monolithic (see Kuhn, 1957), but it is difficult to argue that it was not genuine.

Last, we need, as a research community, to privilege scholarship that represents a genuine advance in our understanding of political phenomena and give (especially junior) scholars the time and the space to tackle bigger questions. In many respects, the institutions of knowledge production and in particular our journal system are not providing the best incentives for cumulative knowledge (see, de Marchi et al., 2020). Unlike most of the natural sciences, our journals are slow and wildly selective; combined with the increasing publishing demands for hiring and promotion, the obvious consequence has been to impede scientific progress.

Whether progress is possible or not is an open question given that humans are difficult to study. But given the rewards for making progress on understanding the substantive questions we all care about, it seems worthwhile to try. For me, the best way to do this is to adopt the standards of the natural sciences – no other approach to cumulative knowledge has had similar success.

Citations

Billings, D., Burch, N., Davidson, A., Holte, R., Schaeffer, J., Schauenberg, T., & Szafron, D. (2003, August). Approximating game-theoretic optimal strategies for full-scale poker. In *IJCAI* (Vol. 3, p. 661).

Brown, Noam, and Tuomas Sandholm. "Superhuman AI for multiplayer poker." *Science* 365.6456 (2019): 885-890.

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1980). *The American voter*. University of Chicago Press.

Deaton, Angus. "Instruments, randomization, and learning about development." *Journal of economic literature* 48.2 (2010): 424-55.

De la Cuesta, Brandon, and Kosuke Imai. "Misunderstandings about the regression discontinuity design in the study of close elections." *Annual Review of Political Science* 19.1 (2016): 375-396.

de Marchi, Scott. *Computational and mathematical modeling in the social sciences*. Cambridge University Press, 2005.

de Marchi, Scott. "The complexity of polarization." Proceedings of the National Academy of Sciences 119.17 (2022): e2115019119.

de Marchi, S., Gerring, J., Hartman, A., Nalepa, M., Pemstein, D., & Seim, B. (2020). A Political Science Peer Review and Publication Consortium. *PS: Political Science and Politics*.

de Marchi, Scott and Brandon Stewart. *Wrestling with complexity in computational social science: Theory, estimation and representation*. SAGE Publications, 2020.

Diermeier, Daniel, and Rebecca Morton. "Experiments in majoritarian bargaining." *Social choice and strategic decisions*. Springer, Berlin, Heidelberg, 2005. 201-226.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., & Nellis, G. (Eds.). (2019). *Information, accountability, and cumulative learning: Lessons from Metaketa I*. Cambridge University Press.

Gerber, Alan S., and Donald P. Green. "The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment." *American political science review* 94.3 (2000): 653-663.

Gilliam Jr, Franklin D., and Shanto Iyengar. "Prime suspects: The influence of local television news on the viewing public." *American Journal of Political Science* (2000): 560-573.

Hall, Andrew B. "What happens when extremists win primaries?." *American Political Science Review* 109.1 (2015): 18-42.

Jenke, Libby. "Introduction to the special issue: Innovations and current challenges in experimental methods." *Political Analysis* 30.S1 (2022): S3-S7.

Johnson, James. "Models-as-fables: An alternative to the standard rationale for using formal models in political science." *Perspectives on Politics* 19.3 (2021): 874-889.

Kim, Sung-youn, Charles S. Taber, and Milton Lodge. "A computational model of the citizen as motivated reasoner: Modeling the dynamics of the 2000 presidential election." *Political Behavior* 32.1 (2010): 1-28.

Klašnja, Marko, and Rocío Titiunik. "The incumbency curse: Weak parties, term limits, and unfulfilled accountability." *American Political Science Review* 111.1 (2017): 129-148.

Kreps, David M., and Robert Wilson. "Reputation and imperfect information." *Journal of economic theory* 27.2 (1982): 253-279.

Kuhn, Thomas. *The Copernican revolution: Planetary astronomy in the development of Western thought*. Vol. 16. Harvard University Press, 1957.

Ostrom, Elinor. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press, 1990.

Page, Scott E. "Where diversity comes from and why it matters?." European Journal of Social Psychology 44.4 (2014): 267-279.

Rubinstein, Ariel. "Dilemmas of an economic theorist." *V Econometrica* 74 (2006): 865-883.

Sekhon, Jasjeet S. "Opiates for the matches: Matching methods for causal inference." *Annual Review of Political Science* 12.1 (2009): 487-508.

Von Neumann, John, and Oskar Morgenstern. "Theory of games and economic behavior." *Theory of games and economic behavior*. Princeton university press, 2007.

Ward, Michael D., et al. "Learning from the past and stepping into the future: Toward a new generation of conflict prediction." *International Studies Review* 15.4 (2013): 473-490.

Zaller, John R. *The nature and origins of mass opinion*. Cambridge university press, 1992.