Original Article

EJIR

European Journal of International Relations
1–25

DOI: 10.1177/13540661221137037
journals.sagepub.com/home/ejt
SAGE

Better than a bet: good reasons for behavioral and rational choice assumptions in IR theory

James W. Davis
University of St. Gallen, Switzerland

Abstract

Behavioral IR is enjoying newfound popularity. Nonetheless, attempts to integrate behavioral research into the larger project of IR theory have proven controversial. Many scholars treat behavioral findings as a trove of plausible ad hoc modifications to rational choice models, thereby lending credence to arguments that behavioral IR is merely residual, empirical, and hence not theoretical. Others limit their research to cataloging outcomes consistent with the basic tenets of behavioral models. Although this expands the empirical base, it is insufficient for theoretical progress. In this article, I explore various answers to the question of when rational choice or behavioral assumptions should guide efforts to build IR theory. I argue that no single answer trumps all others. Examining the various conditions under which actors reason highlights the importance of macrofoundations. Macrofoundations condition the effects of microprocesses and help identify relevant scope conditions for both rational choice and behavioral models of decision-making. Examining the various purposes of IR theory also provides answers to the question of when rational or behavioral assumptions are likely to be most useful. Although many behavioral scholars premise the relevance of their findings on claims of empirical realism, I argue that under certain conditions, deductive theorizing on the basis of as-if behavioral assumptions can lead to powerful theories that improve our understanding of IR and may help decision-makers promote desired ends.

Keywords

International Relations, rationality, rational choice, behavioral IR, theoretical assumptions, political psychology

Corresponding author:
James W. Davis, Institute of Political Science, University of St. Gallen, Müller-Friedberg-Strasse 8, 9000 St. Gallen, Switzerland.
Email: James.Davis@unisg.ch

Introduction

Behavioral International Relations (IR)—understood here as the empirical study of political decision-making by individuals, groups, and institutions with influence on international outcomes—is enjoying newfound popularity. The number of published behavioral studies is increasing as is the number of symposia and special issues devoted to the subject in leading journals (Bleiker and Hutchison, 2014; Davis and McDermott, 2021; Hafner-Burton et al., 2017; Mintz, 2007; van Aaken and Broude, 2019). In large part explained by the growing influence of behavioral economics, this newfound popularity also is driven by a younger generation of scholars applying a broader range of methods and data to long-standing questions of interest and leads to questions of how to integrate their findings into models and theories of IR.[1]

Two dangers face the further integration of behavioral IR into the larger project of IR theory. First, behavioral research presents a tempting trove of plausible ad hoc modifications to rational choice models confronting disconfirming evidence. It is an approach to incorporating behavioral research that both reflects and justifies arguments that the findings are residual, merely empirical, and hence not theoretical (Hafner-Burton et al., 2017: S1–S2; Posner, 1997). Nevertheless, even prominent proponents of behavioral models have fallen into the trap.

Thus, Colin Camerer writes,

> The recipe for behavioral game theory I will describe has three steps: start with a game or naturally occurring situation in which standard game theory makes a bold prediction based on one or two crucial principles; if behavior differs from the prediction, think of plausible explanations for what is observed; and extend formal game theory to incorporate these explanations. (Camerer, 1997: 167–168)

By definition, however, rational actor models cannot accommodate systematic deviations from rational choice and retain the label.[2] Incorporating behavioral findings ad hoc, while retaining a rational choice core, eventually will produce an unfit hybrid. As the internal contradictions mount, the likelihood of surviving repeated encounters with the empirical world is likely to decline, with the blame attributed to the baroque behavioral ornamentation affixed to the clean lines of the original rational choice facade.

The second danger is to succumb to the law of the instrument. Armed with the findings of behavioral research, IR scholars may be tempted simply to catalog the universe of outcomes that appear consistent with basic tenets. If conducted systematically, the approach might produce data sufficient for the establishment of strong correlations that could prove useful for prediction. But merely expanding the empirical base of behavioral regularities is insufficient for the sort of theoretical advancement that would take us beyond what we already know (James, 2007: 164).

How should we think about integrating behavioral IR into IR theory? Before trying to answer the question, a brief discussion of what is meant by the term "theory" is in order. Theories, in the context of this discussion, are understood not as collections of hypotheses derived from direct observation, but rather abstract models based on a set of interrelated assumptions (some of which may be based on observations) that allow the theorist

to deduce hypotheses on how actors are likely to behave under certain conditions. The understanding differs from ideal type analysis in that theory is not being used as a standard against which empirical observations are measured in an effort to uncover deviations that require further explanation, but rather is used to generate expectations about what is most likely to be observed.

The focus of this article is the ongoing debate over the theoretical assumptions we make about how individuals reason and make decisions (Hafner-Burton et al., 2017: S3–S4; Powell, 2017; Stein, 2017). For some, the choice of theoretical assumptions is simply a matter of taste (Bueno de Mesquita, 2014: 45). For others, it is akin to placing a bet: Is investing in models and theories informed by psychology likely to yield a sufficient payoff (Lake and Powell, 1999; Powell, 2017:: S274)?

The goal is to recast the debate over decision-making assumptions at the core of IR theories. Rejecting claims of epistemological certitude as well as superficial dismissals of considered bases for such choices, I discuss various reasons for assuming outcomes will approximate the result of rational choice, as well as those that would favor behavioral assumptions. If the question is whether to adopt rational choice or behavioral assumptions in pursuit of powerful theory, no single answer trumps all others.

One set of answers derives from examining the conditions under which actors make choices. Decision-making is more likely to approximate at least substantive rationality (Simon, 1976) when actors deliberate in competitive environments where survival is precarious; where actors share causal beliefs and relevant information is accessible and cheap; and where the states domestic institutions and processes correct for individual bias. Directing our focus toward relevant scope conditions for decision-making models, the analysis aims to move beyond theoretical trench warfare. The analysis differs from efforts to establish the psychological microfoundations of decision-making given observed variation in the degree to which individuals engage in rational deliberation (Rathbun, 2007; Rathbun et al., 2017). To establish when individual variation should be integrated into general IR theory, I turn to an analysis of *macrofoundations*. These structural conditions create constraints that mediate the effects of microprocesses. Analyzing the structural context of deliberation provides one set of answers to the question of when individual variation matters.

A second set of answers derives from an understanding of the various purposes for which we employ theory. Theories are used to explain, anticipate, and indeed promote certain behavioral outcomes. Sometimes the choice of assumptions is best determined by the function the theory serves.

The argument proceeds in three steps. First, through lenses offered by various levels of analysis, I examine justifications for the rationality assumption based on market analogies. Particular attention is given to arguments that international anarchy promotes rational decision-making. Because the effects of anarchy are variable, we cannot assume that states or decision-makers always act as if they are synoptically rational. Second, I discuss the various purposes for which we develop and employ theories. This leads to the recognition that both rational choice and behavioral theories serve important purposes and sometimes complement each other. A dogmatic preference for one type of theory at the expense of the other can hinder our ability to reach desired goals. Having established conditions under which it makes sense to employ behavioral theory, in a third step I

discuss how best to do so. Recognizing the value of pursuing empirical realism and mechanismic explanations of political behavior, I nonetheless argue that deductive theorizing on the basis of as-if behavioral assumptions can lead to powerful theories that can improve our understanding of IR and serve as a guide to more effective action.

Market analogies and the levels-of-analysis question

Skeptics of the utility of grounding theoretical assumptions in the empirical findings of behavioral science often point to the problem of aggregation and what they regard as "the mismatch between behavioral findings about individuals and the fact that the actors in most IR models and theories are aggregate actors" (Powell, 2017: S265). But the levels of analysis problem is neither new nor specific to behavioral IR (Rosenau, 1961; Singer, 1961; Waltz, 1959).[3] Nonetheless, it does provide a way of framing the issues that takes us beyond a critique that is as applicable to rational choice as it is to the alternatives. After all, "[s]tates cannot think, process information, estimate probabilities, or calculate, only their leaders can" (Stein, 2017: S256). Characteristically, Arnold Wolfers framed the issue well:

> The psychology of the actors in the international arena, instead of operating in limitless space, is confined in its impact on policy by the limitation that external conditions . . . place on the choices open to governments in the conduct of foreign relations. (Wolfers, 1962: 44)

Thus, the relevant question is not whether we should attribute to aggregates a form of rationality that has not been demonstrated empirically at the level of the individual; rather, to what degree the institutional context or external environment compels aggregates to behave in certain ways, regardless of the psychology of relevant individuals.

Perfect markets

A related question was once central to economics but largely has been forgotten by a younger generation of "empirically oriented" researchers who rather blindly go about testing all manner of hypotheses based on the assumption of fully rational actors. Although many justify the practice with reference to arguments popularized by Milton Friedman, he did not assert that we should always assume that economic actors are rational expected utility maximizers. Rather, Friedman argued that *if* the market environment is highly competitive, we then can assume that firms will act *as if* they were rationally seeking to maximize profits, for otherwise they would go bankrupt given the competition (Friedman, 1953: 21–22). Under such *theoretical* conditions, markets would clear.

Adapting the arguments of Friedman to IR, Kenneth Waltz argued that the lack of an international sovereign creates an inherently competitive realm where states can be assumed to act as if they are rational security maximizers, for otherwise they would "fall by the wayside" (Waltz, 1979: 76–77). But this justification for the rationality assumption is rather different from the suggestion that assumptions are little more than wagers

(Powell, 2017). In Waltz's neorealist theory, the assumption of rationality is neither metaphysical nor foundational in the sense that it is analytically independent of other, more fundamental claims. Rather, it is subsidiary to the premise that states interact in a dangerous and unforgiving anarchic environment for which a particular psychological feature—the perception of fear—conveys a survival advantage.[4]

The implications of anarchy, however, are not constant (Milner, 1991). Geography and technology condition anarchy's effects (Jervis, 1978; Taliaferro, 2000). Empirical studies document significant variation in the death rate of states over time, with most state "deaths" occurring in conspicuous clusters (Fazal, 2008). Hence, unless we are willing to conclude that large numbers of states periodically fall victim to a form of collective amnesia—simultaneously forgetting the strategies effective for survival in an anarchic system—it makes more sense to assume that the effects of anarchy vary over time and place. During periods of heated competition and rapid technological or political change, even small mistakes can prove fatal (Cohen and Gooch, 1990; Horowitz, 2010).

Such was the nature of the European state system from roughly 1450 to 1550, when a rapid increase in the effectiveness of artillery and the sudden maturation of the cannon altered the balance between offensive and defensive weapons and strategies. Whereas extensive networks of castles had provided effective defenses and allowed feudal lords of decentralized states to weather most attacks—in large part because the cost of starving castles one by one outweighed the benefits of sustained siege—the ability to efficiently breach their walls now provided an advantage to the larger armies of more aggressive centralized states. With the increased minimum effective size of the state came a sharp reduction in the number of smaller independent states. The new equilibrium persisted more or less until the French Revolution (Bean, 1973).

Under such conditions, assuming that the system will "punish" those who fail to adapt while "rewarding" behaviors that maximize the state's ability to survive makes sense. The evolutionary dynamic ensures that only those actors who behave as if they are strategically rational are around to establish a new equilibrium. Rationality in this formulation is simply the ability to translate fear into actions that promote survival.

Yet if strategic rationality depends on perceptions, the causal logic of neorealism is not free from reductionism. Nonetheless, the independence of the systemic level of analysis is maintained via an evolutionary logic. Actors whose behavior appears puzzling in light of the environmental challenges facing all—either because they appear fearless or are unable to translate fear into effective defensive measures—are unlikely to make it to subsequent rounds of interaction.[5]

If the effects of anarchy on states' prospects for survival vary, so do the systemic incentives for rational deliberation. This is the essence of Rathbun's defense of neoclassical realism against charges that it is an ad hoc or theoretically degenerative modification of neorealism because it incorporates first and second image variables. When systemic constraints are weak, "significant departures from the stylized neorealist depiction of unitary actors and objective perception" are theoretically justifiable (Rathbun, 2008: 296).

A close reading of the IR literature on social preferences also suggests the benefits of assessing the level of threat present in the international system prior to making assumptions about the psychology of foreign policy choice. Both behavioral economists (Fehr

and Fischbacher, 2002; Fehr and Schmidt, 1999) and IR scholars (Kertzer et al., 2014; Lumsdaine, 1993) have documented the fact that actors often are not purely self-interested but consider the impact of their choices on others' welfare. Nonetheless, experimental and case study evidence provided by Kertzer and Rathbun supports the notion that in situations where egoistic "proselfs" can exploit a position of strength for unilateral advantage, the behavior of "prosocial" actors converges over time with that of proselfs. The authors, both leading behavioral scholars, concluded that in IR, "particular contexts make prosocial action extremely improbable" (Kertzer and Rathbun, 2015: 636).

When, however, the international system is more forgiving or less than national survival is at stake, it seems reasonable to assume that there is more room for the impact of idiosyncratic psychological factors on behavior (Wolfers, 1962). When structural constraints are weak, we should not expect uniform behavior or "like actors," as decision-makers not only enjoy a wider range of choices for which individual-psychology is relevant, but also face the challenge of establishing others' motives and intentions when the goals they seek are the product of interdependent choice. For the theorist interested in explaining international outcomes under such conditions, as-if assumptions of egoistic rational actors are theoretically questionable and require additional justification (Milner, 1991).

Efficient markets, information, and causal beliefs

Freidman's as-if argument derived from the theoretical notion of competitive markets. A closely related argument from financial economics that also has gained traction in rational choice models of IR is the efficient market-hypothesis (EMH). The EMH maintains that market prices reflect all available information on asset values. Relevant past, present, and future events are "priced in," such that subsequent price fluctuations are stochastic (random walk) and determined by information in the price series. Should price anomalies appear in the market, they are expected to quickly disappear owing to arbitrage. When efficient, markets will clear, and investors will be unable either to purchase undervalued assets or to sell their assets for inflated prices. Hence, searching for "deals" is irrational. The market will always outperform the individual investor (Fama, 1970; Samuelson, 1965).

By definition, all perfect markets are efficient, but a market need not be perfect to be efficient. Yet few markets exhibit the necessary features: numerous buyers and sellers, none of which can influence price; perfect and costless information available to everyone; an absence of entry barriers; a homogeneous "product"; and an absence of taxes and transaction costs. "The issue therefore reduces to whether (and to what extent) efficiency can be observed in actual (imperfect) markets, and the ultimate test of applicability must be empirical" (Thompson et al., 2003: 108). Stiglitz has demonstrated that when markets are characterized by costly information and/or information asymmetries, they do not clear and a market populated by profit maximizing firms does not produce the social welfare benefits (e.g. efficient distribution of scare resources; full employment, etc.) predicted by classical theory (Rothschild and Stiglitz, 1976; Shapiro and Stiglitz, 1984; Stiglitz, 1979).

Efficient market analogies are implicit in many rational choice models of IR and provide the theoretical baseline against which supposed anomalies are identified. If markets clear, by analogy we expect: power in the international system to be balanced; the welfare diminishing costs of war will be avoided by negotiated settlements; foreign policies will reflect the aggregate interests present within the state; and that trade barriers will be rare. "Puzzles" emerge when outcomes deviate from the theoretical baseline.[6]

Thus, the puzzle of interstate wars emerges from the observation "that war is costly and risky, so rational states should have incentives to locate negotiated settlements that all would prefer to the gamble of war" (Fearon, 1996: 380). Or, as Powell put it: What prevents leaders from reaching more efficient negotiated settlements to those disputes that otherwise result in war (Powell, 2006: 169)? The answer is held to be incomplete, costly, or asymmetric information. If, however, it is the condition of complete and costless information that is rare in international politics, it is far from obvious that the assumption of market efficiency should serve as the default theoretical baseline in IR. Privileging explanations that highlight the implications of incomplete and private information—a general condition in IR—can lead analysts to overlook more proximate causes of war.

A simple thought experiment brings some of the relevant issues into focus and demonstrates how the efficient market analogy can be profoundly misleading when used indiscriminately as the baseline for IR. Was there a "price" for which war with Hitler could have been avoided? Given widespread knowledge in Britain and France of Hitler's industrial and military investment, the incentive to find a negotiated settlement was high. But was the "market failure" of the late 1930s a function of incomplete information and the Western allies having failed to make sufficient concessions? Or was the efficient bargaining outcome unknowable in September 1938, because Hitler's level of ambition both before and during the war was not so much private as it was unstable, an effect of the strategic interaction itself (Hilderbrand, 1976; Hillgruber, 1955; Rich, 1973–1974)? Or was there no price at which war could have been avoided—either through deterrence or concessions—because war itself was Hitler's objective (Kennedy, 1987: 338; May, 1984: 540; Richardson, 1988: 67–68)?

When information is plentiful and cheap, the EMH predicts rational actors will converge on efficient outcomes. But unless we ascribe converging expectations to the realm of chance, an efficient solution to problems such as war, inflation, or employment also requires that agents share the same (valid) causal beliefs (Kirshner, 2015: 169–170; Muth, 1961: 316). Lessons of the past and their implications for the future must be clear and unambiguous and any actors with idiosyncratic beliefs randomly distributed (Muth, 1961: 317).[7] This seems to be what one prominent proponent of rational expectation models in IR had in mind when he wrote:

> [I]f two rational agents have the same information about an uncertain event, then they should have the same beliefs about its likely outcome. The claim is that given identical information, truly rational agents should reason to the same conclusions about the probability of one uncertain outcome or another. Conflicting estimates should occur only if the agents have different (and so necessarily private) information. (Fearon, 1996: 392; Harsanyi, 1968)

But when is it reasonable to assume that foreign policy decision makers share the same information set? Even when they do, the data seldom speak for themselves. Diverging interpretations of the same data are the stuff of both policy and academic debates. Lessons of the past are rarely unambiguous and are as much a product as the source of our causal beliefs. We are interested in the decisions of Hitler, Churchill, and Trump precisely because most observers believe that rational assessments of the situations they confronted should have led them to make quite different choices (Rathbun, 2019). When individuals reach different judgments about the nature of the situation they are confronting or hold different beliefs about how to reach their goals even when they agree on the nature of the situation they confront, their behavioral choices are likely to diverge.

A half century of behavioral research documents how human beings employ numerous cognitive shortcuts to reduce complexity, resolve ambiguity, and avoid the psychological stress produced by the need to confront value trade-offs. These decision-making heuristics bias perceptions and subsequent inferences in systematic and predictable ways. Because people exhibit a strong predisposition to perceive what they already know or expect, they tend to ignore information that contradicts prior beliefs; they assimilate ambiguous information to pre-existing beliefs; they tend to neglect or discount relevant base rate information when a person or situation appears similar to others in a class; and they are quick to reach conclusions, as theory-driven perceptions seem to confirm these preexisting beliefs. Moreover, choices are shaped by the framing of options or the order in which they are presented, in ways that violate fundamental tenets of rational choice.[8]

Many of these findings can be subsumed under the concept of bounded rationality, which variously is regarded as a hybrid of rational and psychological models (Gigerenzer et al., 1999) or a rational choice model of decision-making under constraints (Stigler, 1961). If rational choice depends on ascertaining all possible behavioral alternatives and associated consequences (logical omniscience) and a subsequent comparison of the alternatives in terms of their consequences for desired ends, then our capacity for fully rational decision-making is diminished by constraints on any of these operations.

In the original articulation, bounded rationality referred to the fact that

> [t]he capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world — or even for a reasonable approximation to such objective rationality. (Simon, 1957: 198)

The formulation again draws our attention to the interplay between actors' environments and the psychology of decision-making. Although the international system is always highly complex (Jervis, 1997; Rathbun, 2007), the number of features actors regard to be relevant for a particular decision may vary across contexts. This seems to be what Waltz has in mind when he outlines the virtues of bipolarity (Waltz, 1979: 168–169). Bipolar structures, he argues, allow for more efficient balancing because the number of relevant equations the major actors need to solve is smaller than in multipolar systems. At the same time, deviations among the major actors from rational choice are likely to have greater effects under bipolarity than multipolarity. Although

necessarily brief and superficial, the analysis suggests that absent a serious discussion of relevant macrofoundations, IR theories built on a microfoundation of bounded rationality—conceived either as rational choice under constraint or a rational-psychological hybrid—are conceptually unmoored.

It should be clear from the forgoing discussion that I am not challenging all efforts to model political behavior on the basis of the rational actor assumption. Rather, I am arguing that theorists need to provide compelling reasons for any theoretical assumption. Analogies to perfect or efficient models are legitimate if the relevant scope conditions obtain. When they do not, the analogies may be misplaced. When market forces are weak, equilibria tend to be less stable and more sensitive to the particular characteristics of individual producers and consumers. Under these structural conditions, biases cannot be assumed to be idiosyncratic and randomly distributed with their effects canceled out as the number of transactions increases. As transparency in the market is reduced and the ability of market actors to misrepresent their preferences and assets increases, equilibria grow more sensitive to the specific characteristics of individual actors. Consequently, the weight of explanation should shift from the features of the situation (e.g. the market or the international system) to characteristics particular to the actor (e.g. buyers and sellers, or political decision-makers). The assumption of interchangeable rational actors loses its justification. Under such conditions, behavioral assumptions gain appeal.

Domestic politics and institutions

When market analogies break down, rational choice theorists often argue that societies have developed institutions to compensate for individual deviations from rationality. Such arguments draw our attention to levels of analysis between the international system and individual decision-makers. But though the state's domestic arrangements certainly condition the effects of an individual decision-maker's personality traits, it is by no means obvious a priori whether we should assume that domestic politics ameliorate or enhance the patterns stressed by behavioral science.

Take, for example, the well-known phenomenon of "loss aversion." Central to prospect theory, it refers not only to the fact that humans do not like losses, but more importantly that a loss of a certain magnitude causes more psychological pain than the pleasure caused by an equivalent gain. Because losses loom larger in our deliberations than gains, preferences and corresponding choices are sensitive to whether the framing of a situation highlights the potential for loss or gain relative to a reference point. Furthermore, our aversion to losses leads us to take greater risks than we would to secure equivalent gains and to continue to engage in risky behavior in an effort to recover past losses (Kahneman and Tversky, 1979; Tversky and Kahneman, 1981, 1984, 1986).

Coupled with some rather basic insights from realist theory, loss aversion provides an interesting lens into early 21st-century American foreign policy and the willingness of the United States to engage in preventive war. Realism tells us that the state's definition of its interests tends to grow with its relative power. Because US relative power was at its greatest in the early post-Cold War period (Brooks and Wohlforth, 2008; Krauthammer, 1990/91/91), negative developments in almost any part of the world, if large enough, would threaten its relative position. If the pressure to act is greatest when leaders come

to believe that the price of inaction is a certain loss, loss aversion offers one explanation for George W. Bush's willingness to blunder into risky adventures, such as the war in Iraq, as well as his subsequent resistance to cutting losses and withdrawing when the gamble failed to pay-off (Jervis, 2004).

The foregoing analysis suggests that such risk taking will be less common when the international system is highly competitive, and states enjoy a narrow margin of security. By contrast, the relative security enjoyed by the United States during a period of structural unipolarity allowed for a wider range of action and a greater role for the psychology of the president (Jervis, 2009). But was Bush's reluctance to withdraw from a losing proposition evidence of the failure of domestic institutions to correct for individual psychological bias?

If democratic leaders seek reelection and fear that citizens will punish them at the ballot box for costly boondoggles, then even efforts of a risk averse decision-maker to recover losses appear quite rational (Downs and Rocke, 1995). To understand the effects of domestic institutions on foreign policy decision-making requires first an appraisal of the motives behind leader's choices: are these driven by personal psychological needs, an assessment of the effects of policy on their reelection, or both? Second, even when the politician fears retribution, the origins of voter preferences require explanation. Do voter preferences reflect citizens' judgments of the likely influence of this or that policy or politician on their net assets or does the general validity of prospect theory mean that when voting to stay or change course, voters themselves may be engaging in a risky wager in an effort to recoup their state's lost position?[9] The second possibility suggests that even if common behavioral biases relevant for foreign policy decision-making aggregate to groups as suggested by recent experimental results (Kertzer et al., 2022) one could model these as domestic constraints on the choices of rational elites. The implication is that neither rational nor behavioral assumptions should be privileged a priori in theories cast at the level of the state's domestic politics. Rather, a theorist's choice of assumptions requires justification. Under what conditions should we assume that the state's domestic politics correct for pervasive biases and decisional heuristics and move aggregate behavior into ranges corresponding to the predictions of rational choice?

Moving from the level of the electoral system to the specialized bureaucracies of the state provides no clear-cut answer. Both rational choice and organizational psychology models suggest that bureaucracies sometimes are the source of, rather than the corrective for, biased decisions (Allison, 1971). Moreover, as behavioral researchers have demonstrated, the state's institutions can both enhance and mitigate the influence of individual decision-makers' biases on outcomes (Kahler, 1998; Saunders, 2017).

Recent scholarship links variations in decision-makers' propensity to miscalculate to differences in the organization of the state's bureaucracies and leaders' capacity for oversight. Leaders of states with insular bureaucracies and weak oversight capacity appear to be more prone to foreign policy miscalculation than states with integrated bureaucratic structures and strong oversight capacity (Jost, 2021, 2022) Another way of putting it is that domestic structures mediate the impact of the leader's psychology on the state's foreign policy. Again, the question is one of relevant scope conditions: When do the institutions of the state correct for individual biases and thus produce behaviors more

consistent with the assumption of rational choice (Bendor and Hammond, 1992; Goldgeier and Tetlock, 2001; Jervis, 1976: 28)?

Identifying the most appropriate level of analysis for a given phenomenon and thinking about how the different levels interact with one another can help identify the scope conditions for our models and thus overcome the sterile debate over whether rational or psychological assumptions should be the default option when trying to explain international behavior.[10] When the environment or institutional framework within which actors behave is likely to push heterogenous actors to behave in similar fashion, then the assumption of rational decision-making can be justified on the grounds that it provides the most parsimonious explanation. Even dictators can face strong incentives to create institutions that mediate the deleterious effects of their personality on foreign policy (Weeks, 2014: 14–36) if the neighborhood is dangerous enough (Jost, 2021). But when the environment is less compelling or institutions serve to enhance the effects of individual decisional heuristics and biases, the assumption of rationality loses its appeal. Thinking about the levels of analysis question in this way reveals the weaknesses of the "all or nothing" approach to the assumption of rationality (Powell, 2017: 265–267) and opens space for a more fruitful discussion between scholars representing both rational choice and behavioral traditions.

Purposes of theory

Another way of thinking about the relationship of theories based on behavioral assumptions to those based on the assumption of rational choice directs our attention to the various uses of decision-making theories. Is the purpose of a given theory to predict choices, explain decision-making processes, inform decision-makers of the likely outcomes of various courses of action, or promote desired behavior?

Predicting, explaining, or informing choices

Game theory provides a good example of the diverse purposes for which theories are employed, as it is variously used to predict individual choices and thus outcomes, to account for the reasoning processes that lead to strategic choices, and to inform individual choices in strategic situations. But asking every theory to predict behavior, provide optimal solutions, and describe how people actually deliberate is to ask too much.

Game theory originally was held to explain how rational actors would behave when playing others whom they believed to be rational. Building on the pioneering work of Zermelo (2010 [1913]), game theorists developed the concept of backward induction to provide subgame perfect solutions for sequential games with perfect information (von Neumann and Morgenstern, 1944). Although defenders of the concept's theoretical validity did not argue that real people actually engage in backward induction (Luce and Raiffa, 1957: 97–102; Selten, 1978: 138), systematic violations of rational play, even under conditions of perfect information, are commonplace (Camerer, 1997). Even when playing for high stakes, people generally are nicer—that is, other regarding—than standard models predict, with up to 50 percent cooperating in single shot prisoners' dilemma games (van den Assem, van Dolder, and Thaler, 2012).[11] Similarly, most people tend to

contribute to the provision of public goods even though collective action theory tells us it is irrational to do so (Fehr and Schmidt, 1999; Rabin, 1993).[12] Moreover, studies of negotiation behavior in well-defined games suggest that individuals systematically overlook available information and simplify complex problems in ways that lead players to construct mental models that no longer accurately reflect the essential features of the game they are playing (Camerer, 1997; Neale and Bazerman, 1991). Thus, even if they are engaged in backward induction, they likely are solving for a private game of their own making.

In IR, actors routinely confront situations that appear ambiguous and would support multiple interpretations. Thus, before they can decide which information might be relevant for strategy or begin to invest mental energy to reduce complexity, they first need to establish the parameters of the game they (think they) are playing. If both the payoffs as well as the range and meaning of available moves are uncertain, it is unclear that starting from a stylized depiction of the situation makes much sense, regardless of whether the goal is to explain or inform agent's choices.

The issue was at the heart of Lebow and Stein's critique of rational deterrence theory: "The most important determinant of strategic decisions is not the process of choosing among options but the prior definition and construction of the problem to be decided" (Lebow and Stein, 1989: 214). Whereas abstract models of deterrence encounters assume that the status quo is unambiguous and thus identifying challengers and defenders is clear-cut, detailed analyses of real-world inter-state crises revealed that these defining features of the game were obvious neither to scholars nor to the actors themselves. In many cases, they found both parties to the encounter considered themselves to be defenders of the status quo. If "[l]eaders' conceptions of themselves as initiators or defenders have important consequences for their behavior," then we need to understand the processes that lead to their subjectively constructed mental models (Lebow and Stein, 1989: 221). A related issue concerns the nature of the strategic choices—the moves—available to the agents. Standard game-theoretic treatments of deterrence comprise two: "stand firm" or "back down." Yet carefully conducted historical studies find a wider range of options available to decision-makers. Moreover, individual moves are often less clear-cut, with admixtures of firm and accommodating strategies possible (Davis, 2000; George and Smoke, 1974, 1989; Jervis, 1979; Kecskemeti, 1964). Efforts to inform or explain behavior that begin from the premise that the roles, stakes, and available moves in a strategic encounter are obvious are bound to come up short.[13]

Does this mean that traditional game theory is useless as a guide to strategic choice? *It depends*. Knowing, for example, that most people find backward induction difficult and concentrate on the current round of play at the expense of future rounds of iterated games can provide a strategic advantage to the informed player, a fact that may provide an incentive to invest the mental energy to solve across the many branches of the decision tree (Camerer et al., 1993). In this example, traditional game theory recovers its strategic utility for the player who can make descriptively accurate assumptions about others. To prevail in the real world, one needs to recognize that games often have both logical and psychological equilibria (Goldgeier and Tetlock, 2001; Green and Shapiro, 1994). Paradoxically, if both sides are aware of such tendencies and react in similar fashion, outcomes will shift toward those predicted by the rational model. The role of theory

in such situations is complex, as theory seems both to cause and explain the outcomes of interest (Gartzke, 1999; Jervis, 2008). Rather than regarding rational choice (or just straightforward cost-benefit analysis) and behavioral theories as necessarily competing, such examples suggest that each may play a role in different phases of decision-making processes. For many situations, rational choice analysis will remain the gold standard for identifying options that are likely to maximize individual or collective returns. Nonetheless, knowing that deviations from the normative strictures of rational choice are frequent and systematic is helpful whether the goal is to take advantage of or correct for them.

Nudging

A more straightforward effort to use behavioral scholarship to reach desired ends is the growing literature on "choice architecture." Armed with predictive insights from behavioral research, choice architects now routinely are included in everything from the design of school cafeterias to hospital operating rooms (Thaler et al., 2012). Some of the basic ideas have long been familiar to students of electoral behavior—and not too few state electoral boards—who know, for example, that ballot design produces systematic effects on electoral outcomes (Walker, 1966). Although less common in the field of IR, recent governmental applications have been notable (Chetty, 2015; Johnson and Goldstein, 2003; Thaler and Sunstein, 2009). Many extensions to the optimal design of international regimes, treaties, and the standard operating procedures of international organizations are fairly straightforward.

Thus, Galbraith suggests negotiators can adapt choice architecture principles in the drafting of treaties in an effort to "nudge" those involved in national ratification processes to accept their preferred provisions. Based on an analysis of the negotiations and ratification processes of over 300 international treaties, she argues that states are more likely to enter into optional commitments—such as supplementary protocols, or precommitting to International Court of Justice (ICJ) jurisdiction for disputes arising under the treaty—if treaties are written in a way that requires states to "opt-out" of otherwise required commitments rather than "opt-in" to additional obligations (Galbraith, 2013).

Others have suggested harnessing loss aversion to achieve behavioral change in pursuit of international goals and standards. Setting a goal provides a reference point and divides prospective outcomes into two domains: those above and those below the reference point. Prospect theory would lead us to expect leaders to engage in strenuous efforts to avoid outcomes that fall short of the goal. Teichman and Zamir suggest that the relative success of states in achieving the United Nations Millennium Development Goals (MDG) reflects such processes. The MDG established specific targets in areas such as poverty reduction, infant mortality, HIV/aids, and gender equality. Following their adoption in 2000, the United Nations established a comprehensive system to monitor states' progress toward meeting the targets. Although the agreement made no provisions for sanctioning underperformance, many of the goals were achieved. For example, by 2015, the number of people living in extreme policy was cut in half (Teichman and Zamir, 2020: 1271).

Such effects may be stronger when achievement is coupled with material benefits. For example, increases in worker productivity have been found to be greater and more sustained when bonus payments are provided up front yet subject to loss if performance targets are not met. The immediate payment of a benefit establishes a new status quo, induces loss aversion, and thus spurs employee effort (Hart and Moore, 2008). The finding may have implications for the design of international institutions. For example, norm compliance should be higher in institutions where the benefits of membership are secured up front and only lost in the event of non-compliance, than in those institutions where benefits are structured as a reward only granted after a period of compliance (Davis and McDermott, 2021: 165).

Building behavioral IR theory

Rejecting dogmatic either-or arguments in favor of a pragmatic, tool-box approach to IR theorizing, I have argued that it is important to assess the conditions under which actors reason and the diverse purposes for which theories are employed. Under certain structural conditions, models based on the assumption of synoptic rationality are reasonable. Under others, beginning the analysis from the empirical findings of psychology and behavioral science makes more sense. But if behavioral IR is going to move beyond a mere catalog of empirically documented heuristics and biases that deviate from the equilibria outcomes predicted by rational choice models, then behaviorally oriented scholars need to devote more energy to task of theory building. This raises the question of how to do so.

One answer is suggested by Kertzer and Tingley, who argue that "the era of grand systemic theories" in IR is over, in large part due to a growing interest in micro-level phenomena. Because micro-foundational accounts of decision-making processes focus on the operation of psychological mechanisms, they maintain that behavioral IR implies a mechanismic understanding of explanation (Kertzer and Tingley, 2018: 321).

Insofar as the operation of causal mechanisms varies across different contexts, adherents to a mechanismic conception of causality (George and Bennett, 2005; Gerring, 2007) should be predisposed to the argument that both behavioral and rational choice scholars should pay more attention to the scope conditions of their theories. But does increased interest in the micro-foundations of human behavior necessarily imply we abandon the enterprise of developing theories of IR that anticipate general patters and the bounds within which most outcomes will fall? To do so would impoverish the discipline and leave it incapable of providing compelling answers to the important questions that gave rise to the academic study of IR in the first place. Moreover, I am not persuaded that the best path to theoretical progress in behavioral IR runs through an effort to nail down real psychological processes, if for no other reason than that psychology and neuroscience have yet to provide a coherent understanding of empirically realistic decision-making mechanisms.

The most prominent approach to replacing standard rational choice assumptions with an empirically based alternative is dual process theory (DPT). DPT conceives and categorizes mental processes according to two systems or types. System 1 (or Type 1) processes are depicted as fast, reactive, automatic, intuitive, heuristic, associative and

unconscious or preconscious mental activities and are contrasted with System 2 (Type 2) processes, which are associated with slow, controlled, effortful, reflective, serial, rule-based and conscious processes including deductive, hypothetical, and counterfactual reasoning.[14] This has led many behaviorally oriented economists and political scientists to erroneously assume that System 2 is more or less synonymous with rational decision-making, whereas System 1 is held to be responsible for decisions that appear irrational. Although conceptually distinct, the two systems likely do not operate in isolation of one another, but rather interact. Whereas Kahneman and Frederick (2002) have suggested that each competes for control over overt responses to decisional needs, others maintain they operate sequentially, with System 2 monitoring and intervening when the requisite mental resources are available (Lurquin and Miyake, 2017; Pennycook et al., 2015; Sinayev, 2016). Moreover, deviations from the expectations of rational choice can occur from errors specific to System 2. Individuals can fall victim to a form of "over-thinking" where relatively simple information is distorted as a result of deliberation and they often engage in pseudo-rational thought in post hoc efforts to rationalize behavior engaged in for other reasons (Bonnefon, 2018; Mercier and Sperber, 2011; Nisbett and Ross, 1980). DPT simply is not yet able to answer the question of when it is reasonable to assume that System 2 can override judgments that result from the unconscious mental processes of System 1 and produce decisions corresponding to rational choice. Most likely a reflection of the early stages of research in this vein, the fact nonetheless has led some critics to wonder why we should make the distinction at all (Grayot, 2020: 113).

Although less well known, Damasio's "somatic marker" hypothesis (Damasio, 1996) provides a somewhat different explanation for when psychological processes are most likely to dominate decision-making. Damasio postulates that feelings generated by experience are physically embodied as emotional associations. These somatic markers are activated when individuals subsequently experience similar situations and provide rapid information on which people or outcomes should be approached and which should be avoided. Rather than hindering, somatic markers facilitate rational calculation by constraining the individual's decisional space (by simply ruling some things out). Absent the affective guidance provided by somatic markers, decision-making becomes nearly impossible as individuals get lost in a process of infinite regress.

Fortunately, we do not need to wait for a resolution to these issues. The question for IR is not whether we have grasped how the brain really works, but whether we can develop theories that help us to explain and anticipate behavior. While welcoming—or even contributing to—continued research directed at uncovering real mental processes, IR scholars can focus on developing theories that incorporate behavioral assumptions and provide powerful insights into political choices under relevant scope conditions. In fact, this is precisely the approach followed by luminaries in the field of behavioral economics.

One of the motivations for early psychological research was a rejection of the notion that people really could conduct the sort of mental calculations that would approximate the basic axioms of rational choice. Yet ironically, the assumptions supporting the most influential psychological models of choice actually *increase* the number of relevant decisional parameters. Thus, whereas both prospect theory and expected utility theory postulate that risky choice follows a process of weighting and averaging all relevant

information, prospect theory includes additional parameters to account for the systematic under- or over-weighting of probabilities (Kahneman and Tversky, 1979). To account for observed other-regarding preferences, the social preference function postulated by Fehr and his colleagues (Fehr and Schmidt, 1999) amends the classical utility function by adding new parameters weighting the decision-maker's concern for receiving more, or less, than the group average. Similarly, Laibson's (1999) model accounts for hyperbolic discounting and time-inconsistent preferences by postulating that choosers conduct an exhaustive search of all feasible consumption sequences rather than the narrower set of currently available consumables.

Reflecting on these developments, Berg and Gigerenzer note,

> Aside from this paradox of increasing complexity found in many boundedly rational models, there is the separate question of whether any empirical evidence actually supports the modified versions of the models in question. If we do not believe that people are solving complex optimization problems—and there is no evidence documenting that the psychological processes of interest are well described by such models—then we are left only with as-if arguments to support them. (Berg and Gigerenzer, 2010: 141)

But the utility of theories is not reducible to the descriptive accuracy of their assumptions. We can build theories with realistic assumptions and explain little (Leamer, 1983; McCloskey and Ziliak, 1996). Accurate predictions are possible with theories based on absurd assumptions (Bunge, 1996: 55; Quackenbush, 2004: 91). Forecasting tournament have demonstrated that predictive accuracy can be enhanced by aggregation the predictions of many. In the process of aggregation, however, the importance of the individual theories and the assumptions behind them washes out (Kertzer, 2017: 85). In fact, any move from description to explanation requires abstraction, which is a move away from descriptive accuracy. Theories are as much about the organization as the descriptive accuracy of our assumptions. As-if assumptions are legitimate to the extent that we can give reasons for adopting them.

I have argued that when the environment is highly competitive or approximates efficient market conditions, assuming that decision-makers behave as if they were synoptically rational utility maximizers is not unreasonable. When, however, the environment is less compelling, when interactions are not repeated, and when there are good grounds to believe that the idiosyncrasies of individuals matter, then assuming the choices of decision-makers are governed by the operation of postulated psychological mechanisms is justifiable.

The approach to theorizing I am proposing bears some resemblance to arguments regarding the notion of "ecological rationality" (Gigerenzer et al., 1999: 3–34). Decision heuristics can be considered ecologically rational to the degree that they are functional adaptations to the structure of the decision-maker's environment. Whereas evolutionary arguments provide an account for the emergence of heuristics (Buss, 2009; McDermott et al., 2008), the sort of theorizing I am proposing requires a prior assessment of the conditions under which actor's reason and maintains that our assumptions should match these environments.[15] Once we have done so, we can move to the development of deductive theory.

Even when structural conditions clearly favor a rational or a behavioral assumption about the psychology of decision-making, substantive theories contain additional assumptions about actors' goals and preferences. For example, a rational voter theory makes different substantive assumptions about actors' goals than does a rational deterrence theory. Such assumptions may, but need not be, real in an ontological sense. After all, in what sense do "utility" or "self-esteem" refer to something real? Both are concepts that help us to theorize neurobiological processes that are the products of our evolved brains (Glimcher et al., 2006: 2014). The understanding of theory adopted here regards assumptions to be analytical devices for developing theories that help us to anticipate behavior and provide a coherent account for why it tends to fall within expected bounds (Goddard and Nexon, 2005: 17).

Building models on the assumption that actors reason as if they were guided by the findings of behavioral researchers and then deducing the implications of the models avoids the temptation of simply matching observed behavior with some heuristic, bias, or emotion with which it appears to be consistent. The latter is an approach to theoretical explanation that is resistant to refutation and rarely takes us beyond what we already know. By contrast, integrating behavioral assumptions into deductive theorizing can lead to novel hypotheses that can be confirmed or rejected through subsequent empirical analysis.[16]

Take, for example, the study of deterrence. The original articulations of deterrence theory assume that the decisions of rational actors with fixed risk profiles can be influenced by threats and promises that change their estimations of the net assets associated with various prospective outcomes (Kaufmann, 1954; Schelling, 1966). For such models, the effects on a target of enhancing the value of the status quo with promised rewards and of diminishing the value of challenging the status quo through threats of punishment are functionally equivalent. By contrast, a behavioral deterrence model that assumes actors frame prospective outcomes around a reference point, have utility functions that are convex below and concave above the reference point, and thus are risk averse to outcomes above, but risk acceptance to outcomes below the reference point, leads to different expectations regarding the effect of threats and promises on targets' calculations (Davis, 2000). If decision-makers are willing to take greater risks to avoid losses than to make gains of a similar magnitude, then threats of punishment will prove less effective in deterring aggression by a state that is motivated by a desire to avoid a deterioration of the status quo than they will be at preventing aggression aimed at enhancing it. Given the shape and slope of the utility function, a promised reward of the same size will represent a larger prospective increase in utility in most regions below the reference point than in those above. Incorporating behavioral assumptions into an IR theory in this way takes us beyond well-documented empirical regularities and hypothesized psychological mechanisms to substantive questions of international politics.

Conclusion

Can it be that the human capacity for reason makes rational analysis with corresponding choice possible, and that the brain that evolution has provided us operates in ways that violate basic tenets of rational choice analysis? Of course. Is it possible to develop a

theory of decision-making that assumes human reasoning follows the normative strictures of rational choice except when it doesn't? Not if theories are understood to be falsifiable when observations do not corroborate the expectations they generate. So, does that mean that we are reduced to gamblers placing outside bets at the roulette table on the basis of some unconscious heuristic or bias? I think not. As trained theorists, we can invest the mental energy to allow for more reasoned choices.

The reasons for behavioral theories in IR are many. Nevertheless, it should be clear by now that I am not suggesting that models of foreign policy decision-making based on behavioral assumptions need to displace or subsume existing rational choice modes (in the sense that they not only explain the cases claimed by existing models but then some) in order to be useful. My argument is rather that we should devote our energies to constructing models and theories based on axioms derived from behavioral research and to identifying the scope conditions that condition their utility. For some sets of questions, the issue may indeed be whether behavioral theories provide a superior explanation than those based on the assumption of synoptic rationality. I have suggested that this is the case for isolated interactions among agents in situations where competitive pressures do not threaten survival and when information is scarce, costly, and/or ambiguous. No doubt there are other, equally reasoned justifications. For other questions, it may be a question of sequencing, where either rational or behavioral analysis provides a (normative or positive) baseline for which the other approach suggests an optimal response.

Acknowledgements

My thanks to the editors and two anonymous reviewers at the *European Journal of International Relations* for their constructive critiques and helpful suggestions on earlier versions of this article. I also owe a debt of gratitude to many colleagues who provided feedback at various stages of this project, in particular David Baldwin, Tim Büthe, Simon Evenett, Robert Jervis, Jonathan Kirshner, Rose McDermott, Kate McNamara, Brian Rathbun, Jack Snyder, and Janice Gross Stein.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

James W. Davis https://orcid.org/0000-0003-1851-8193

Notes

1. The term behavioral IR is a frequent source of confusion. In psychology, behaviorism regards behavior as an objective and observable conditioned response to an individual's environment and eschews reference to factors such as cognition or emotion that it regards to be idiosyncratic and subjective. The origins of behavioral IR lie in the application of cognitive psychology to the study of political decision-making. Now part of the larger field of behavioral science, behavioral IR increasingly makes use of data and methods from a broader range of disciplines, including neuroscience, genetics, and evolutionary biology.
2. For an accessible introduction to rational choice, see Kirchgässner (2008).
3. For an argument that the challenge of aggregating behavioral findings is both exaggerated and less important than often claimed, see Gildea (2020).

4. For a similar formulation, see Rathbun (2007: 539–540).
5. Goddard and Nexon (2005) argue that Waltz's theory, properly understood, relies on a structural-functionalist logic. Explaining how the system maintains order requires an understanding of functional mechanisms that emerge from the interaction of the structure and units of the system. Balances of power among like units constitute an emergent property of the system as a whole. Strategic rationality is thus a functional requirement of the system, which itself is an analytical device. I thank an anonymous reviewer for pointing this out.
6. For IR scholarship that takes efficient markets as a baseline, see Fearon (1996), Milner (1997), Potoski and Prakash (2009), and Powell (1999). In the subfield of comparative political economy, examples include Alesina (1988), Alesina et al. (1989), Koremenos (2005), Lohmann (1992), and Pollack (1997).
7. An additional condition—that all equations of the system are linear—is also problematic though not addressed here. See Jervis (1997).
8. For reviews of these findings, see Davis and McDermott (2021), Levy (2003), and Stein (2017).
9. For arguments to this effect, see Berejikian (2004), Masters and Alexander (2008), and Nincic (1997).
10. For a similar argument that is more sanguine about the prospects for rational decision-making, see Druckman (2004).
11. Despite common claims to the contrary, evidence does not support the notion that actors always behave more rationally if the stakes are raised. See Grether and Plott (1979).
12. For the conditions under which real communities have solved problems of collective action, see Ostrom (2003).
13. Early deterrence theorists were not so much unaware of these issues as they were committed to the possibility of solving them through logic. See Trachtenberg (1989).
14. Popularized by Kahneman (2011), the foundational studies include Evans (1989), Lieberman et al. (2002), Shiffrin and Schneider (1977), Sloman (1996), and Stanovich and West (2000).
15. McDermott et al. (2008) develop a similar methodological argument, but suggest that under some competitive conditions, actors with prospect theoretic utility functions will take greater risks than, and subsequently outperform, actors whose decision-making approximates an economic model of procedural rationality.
16. In pointing out advantages of deductive theory I am claiming neither that this is the only way to conduct social science, nor that a straightforward appeal to objective facts as arbiters of theoretical claims is possible (since what we see is also a function of preexisting concepts, themselves often given by theory). Because theories serve a variety of purposes, the criteria for judging their utility should also vary by context. See Hellmann (2002) and Kratochwil (2018: 141–191; 325–326; 394–399).

References

Alesina A (1988) Macroeconomics and politics. In: Fischer S (ed.) *NBER Macroeconomics Annual 1988*. Cambridge, MA: MIT Press, pp. 13–69.

Alesina A, Mirrlees J and Neumann MJN (1989) Politics and business cycles in industrial democracies. *Economic Policy* 4(8): 57–98.

Allison GT (1971) *Essence of Decision: Explaining the Cuban Missile Crisis*. New York: Harper Collins.

Bean R (1973) War and the birth of the nation state. *Journal of Economic History* 33(1): 203–221.

Bendor J and Hammond TH (1992) Rethinking Allison's models. *American Political Science Review* 86(2): 301–322.

Berejikian JD (2004) *International Relations Under Risk*. Albany, NY: State University of New York Press.
Berg N and Gigerenzer G (2010) As-if behavioral economics. Neoclassical economics in disguise? *History of Economic Ideas* 18(1): 133–165.
Bleiker R and Hutchison E (2014) Forum: emotions and world politics. *International Theory* 63(3): 490–594.
Bonnefon JF (2018) The pros and cons of identifying critical thinking with system 2 processing. *Topoi* 37(1): 113–119.
Brooks SH and Wohlforth WC (2008) *World Out of Balance: International Relations and the Challenge of American Primacy*. Princeton, NJ: Princeton University Press.
Bueno de Mesquita B (2014) *Principles of International Politics*. London: SAGE.
Bunge M (1996) *Finding Philosophy in Social Science*. New Haven, CT: Yale University Press.
Buss DM (2009) The great struggles of life: Darwin and the emergence of evolutionary psychology. *American Psychologist* 64(2): 140–148.
Camerer CF (1997) Progress in behavioral game theory. *Journal of Economic Perspectives* 11(4): 167–188.
Camerer CF, Johnson EJ, Rymon T, et al. (1993) Cognition and framing in sequential bargaining for gains and losses. In: Binmore K and Tani P (eds) *Contributions to Game Theory*. Cambridge, MA: MIT Press, pp. 27–47.
Chetty R (2015) Behavioral economics and public policy: a pragmatic perspective. *American Economic Review* 105(5): 1–33.
Cohen EA and Gooch J (1990) *Military Misfortunes: The Anatomy of Failure in War*. New York: Free Press.
Damasio AR (1996) The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions: Biological Sciences* 351(1346): 1413–1420.
Davis JW (2000) *Threats and Promises. the Pursuit of International Influence*. Baltimore, MD: Johns Hopkins University Press.
Davis JW and McDermott R (2021) The past, present, and future of behavioral IR. *International Organization* 75(1): 147–177.
Downs GW and Rocke DM (1995) *Optimal Imperfection? Domestic Uncertainty and Institutions in International Relations*. Princeton, NJ: Princeton University Press.
Druckman JN (2004) Political preference formation: competition, deliberation and the (ir)relevance of framing effects. *American Political Science Review* 98(4): 671–686.
Evans J (1989) *Bias in Human Reasoning: Causes and Consequences*. Mahwah, NJ: Erlbaum.
Fama EF (1970) Efficient capital markets: a review of theory and empirical work. *The Journal of Finance* 25(2): 383–417.
Fazal TM (2008) *State Death: The Politics and Geography of Conquest, Occupation and Annexation*. Princeton, NJ: Princeton University Press.
Fearon J (1996) Rationalist explanations for war. *International Organization* 49(3): 379–414.
Fehr E and Fischbacher U (2002) Why social preferences matter: the impact of non-selfish motives on competition, cooperation, and incentives. *Economic Journal* 112(478): C1–C33.
Fehr E and Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114(3): 817–868.
Friedman M (1953) The methodology of positive economics. In: Friedman M (ed.) *Essays in Positive Economics*. Chicago, IL: University of Chicago Press, pp. 3–43.
Galbraith J (2013) Treaty options: towards a behavioral understanding of treaty design. *Virginia Journal of International Law* 53(2): 309–363.
Gartzke E (1999) War is in the error term. *International Organization* 53(3): 567–587.

George A and Bennett A (2005) *Case Studies and Theory Development*. Cambridge, MA: MIT Press.
George A and Smoke R (1974) *Deterrence in American Foreign Policy. Theory and Practice*. New York: Columbia University Press.
George A and Smoke R (1989) Deterrence and foreign policy. *World Politics* 41(2): 170–182.
Gerring J (2007) The mechanismic worldview: thinking inside the box. *British Journal of Political Science* 38: 161–179.
Gigerenzer G and Todd PR and ABC Research Group (1999) *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.
Gildea RJ (2020) Psychology and aggregation in international relations. *European Journal of International Relations* 26(1): S166–S183.
Glimcher PW, Dorris MC and Bayer HM (2006) Physiological utility theory and the neuroeconomics of choice. *Games and Economic Behavior* 52(2): 213–256.
Goddard SE and Nexon DH (2005) Paradigm lost? Reassessing theory of international politics. *European Journal of International Relations* 11(1): 9–61.
Goldgeier JM and Tetlock PM (2001) Psychology and international relations theory. *Annual Review of Political Science*: 467–492.
Grayot JD (2020) Dual process theories in behavioral economics and neuroeconomics: a critical review. *Review of Philosophy and Psychology* 11(1): 105–136.
Green DP and Shapiro I (1994) *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven, CT: Yale University Press.
Grether DM and Plott CR (1979) Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69(4): 623–238.
Hafner-Burton E, Haggard S, Lake DA, et al. (2017) The behavioral revolution and international relations. *International Organization* 71(S1): S1–S31.
Harsanyi J (1968) Games with incomplete information played by "Bayesian" players, part III. The basic probability distribution of the game. *Management Science* 14(7): 486–502.
Hart O and Moore J (2008) Contracts as reference points. *Quarterly Journal of Economics* 123(1): 1–48.
Hellmann G (2022) Practising theorizing in theorizing praxis: Friedrich Kratochwil and social inquiry. In Hellmann G and Steffek J (eds) *Praxis as a Perspective on International Politics*. Bristol: Bristol University Press, pp. 72–93.
Hilderbrand K (1976) Hitler's war aims. *Journal of Modern History* 48(3): 522–530.
Hillgruber A (1955) Der Faktor Amerika in Hitlers Strategie, 1938-1941. *Aus Politik und Zeitgeschichte. Beilage zur Wochenzeitung Das Parlament* 39: 3–21.
Horowitz MC (2010) *The Diffusion of Military Power: Causes and Consequences for International Politics*. Princeton, NJ: Princeton University Press.
James P (2007) Behavioral IR. Practical suggestions. *International Studies Review* 9(1): 162–165.
Jervis R (1976) *Perception and Misperception in International Politics*. Princeton, NJ: Princeton University Press.
Jervis R (1978) Cooperation under the security dilemma. *World Politics* 30(2): 167–214.
Jervis R (1979) Deterrence theory revisited. *World Politics* 31(2): 289–324.
Jervis R (1997) *Systems Effects. Complexity in Political and Social Life*. Princeton, NJ: Princeton University Press.
Jervis R (2004) The implications of prospect theory for human nature and values. *Political Psychology* 25(2): 163–176.
Jervis R (2008) Bridges, barriers, and gaps. Research and policy. *Political Psychology* 29(4): 571–592.
Jervis R (2009) Unipolarity: a structural perspective. *World Politics* 61(1): 188–213.

Johnson EJ and Goldstein DG (2003) Do defaults save lives? *Science* 302(5649): 1338–1339.
Jost T (2021) Authoritarian advisers: institutional origins of miscalculation in Chinese foreign policy. Working paper, Brown University, Providence, RI.
Jost T (2022) Leaders, bureaucracy, and miscalculation in international crisis. Working paper, Brown University, Providence, RI.
Kahler M (1998) Rationality in international relations. *International Organization* 52(4): 919–941.
Kahneman D (2011) *Thinking Fast and Slow*. New York: Macmillan.
Kahneman D and Frederick S (2002) Representativeness revisited: attribute substitution in intuitive judgement. In: Gilovich T, Griffin DW and Kahneman D (eds) *Heuristics and Biases: The Psychology of Intuitive Judgement*. New York: Cambridge University Press, pp. 49–81.
Kahneman D and Tversky A (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47(2): 263–291.
Kaufmann WW (1954) *The Requirements of Deterrence*. Memorandum 7. Princeton, NJ: Center of International Studies.
Kecskemeti P (1964) *Strategic Surrender*. New York: Athenaeum.
Kennedy P (1987) *The Rise and Fall of the Great Powers: Economic Change and Military Conflict from 1500-2000*. New York: Random House.
Kertzer JD (2017) Microfoundations in international relations. *Conflict Management and Peace Science* 34(1): 81–97.
Kertzer JD, Powers KE, Rathbun BC, et al. (2014) Moral support: how moral values shape foreign policy preferences. *Journal of Politics* 76(3): 825–840.
Kertzer JD, Holmes M, LeVeck BL, et al. (2022) Hawkish biases and group decision making. *International Organization* 76: 513–548.
Kertzer JD and Rathbun BC (2015) Fair is fair: social preferences and reciprocity in international politics. *World Politics* 67(4): 613–655.
Kertzer JD and Tingley D (2018) Political psychology in international relations. Beyond the paradigms. *Annual Review of Political Science* 21: 1–23.
Kirchgässner G (2008) *Homo Oeconomicus: The Economic Model of Individual Behavior and Its Application in the Economic and Social Sciences*. New York: Springer.
Kirshner J (2015) The economic sins of modern IR theory and the classical realist alternative. *World Politics* 67(1): 155–183.
Koremenos B (2005) Contracting around international uncertainty. *American Political Science Review* 99(4): 549–565.
Kratochwil FV (2018) *Praxis: On Acting and Knowing*. Cambridge: Cambridge University Press.
Krauthammer C (1990/91) The unipolar moment. *Foreign Affairs* 70(1): 23–33.
Laibson D (1999) Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112(2): 443–477.
Lake DA and Powell R (1999) International relations: a strategic choice approach. In: Lake D and Powell R (eds) *Strategic Choice and International Relations*. Princeton, NJ: Princeton University Press, pp. 3–38.
Leamer EE (1983) Let's take the con out of econometrics. *American Economic Review* 73(1): 31–43.
Lebow RN and Stein JG (1989) Rational deterrence theory. I think, therefore I deter. *World Politics* 41(2): 208–224.
Levy JS (2003) Political psychology and foreign policy. In: Sears DO, Huddy L and Jervis R (eds) *Oxford Handbook of Political Psychology*. Oxford: Oxford University Press, pp. 253–284.
Lieberman MD, Gaunt R, Gilbert DT, et al. (2002) Reflexion and reflection: a social cognitive neuroscience approach to attributional inference. *Advances in Experimental Social Psychology* 34: 199–249.

Lohmann S (1992) Optimal commitments in monetary policy: credibility versus flexibility. *American Economic Review* 82(1): 273–286.
Luce D and Raiffa H (1957) *Games and Decisions*. New York: Wiley.
Lumsdaine DH (1993) *Moral Vision in International Politics: The Foreign Aid Regime, 1949-1989*. Princeton, NJ: Princeton University Press.
Lurquin JH and Miyake A (2017) Challenges to ego-depletion research go beyond the replication crisis: a need for tackling the conceptual crisis. *Frontiers in Psychology* 8(568): 1–5.
McCloskey DN and Ziliak ST (1996) The standard error of regression. *Journal of Economic Literature* 34(1): 97–114.
McDermott R, Fowler JH and Smirnov O (2008) On the evolutionary origins of prospect theory preferences. *Journal of Politics* 70(2): 335–350.
Masters D and Alexander RM (2008) Prospecting for war. 9/11 and selling the Iraq war. *Contemporary Security Policy* 29(3): 434–454.
May ER (1984) *Knowing One's Enemies: Intelligence Assessment before the Two World Wars*. Princeton, NJ: Princeton University Press.
Mercier H and Sperber D (2011) Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences* 34(2): 57–74.
Milner HV (1991) The assumption of anarchy in international relations theory: a critique. *Review of International Studies* 17(1): 67–85.
Milner HV (1997) *Interests, Institutions, and Information*. Princeton, NJ: Princeton University Press.
Mintz A (2007) The forum. Behavioral IR as a subfield of international relations. *International Studies Review* 9(1): 157–172.
Muth JF (1961) Rational expectations and the theory of price movements. *Econometrica* 29(3): 315–335.
Neale MA and Bazerman MH (1991) *Cognition and Rationality in Negotiation*. New York: Free Press.
Nincic M (1997) Loss aversion and the domestic context of military intervention. *Political Research Quarterly* 50(1): 97–120.
Nisbett RE and Ross L (1980) *Human Inference. Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, NJ: Prentice-Hall.
Ostrom E (2003) How types of goods and property rights jointly affect collective action. *Journal of Theoretical Politics* 15(3): 239–270.
Pennycook G, Fugelsang JA and Koehler DJ (2015) What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology* 80: 34–72.
Pollack MA (1997) Delegation, agency, and agenda setting in the European Community. *International Organization* 51(1): 99–134.
Posner RA (1997) Rational choice, behavioral economics, and the law. *Stanford Law Review* 50: 1551–1575.
Potoski M and Prakash A (2009) Information asymmetries as trade barriers: ISO 9000 increases international commerce. *Journal of Policy Analysis and Management* 28(2): 221–238.
Powell R (1999) *In the Shadow of Power: States and Strategies in International Politics*. Princeton, NJ: Princeton University Press.
Powell R (2006) War as a commitment problem. *International Organization* 60(1): 169–203.
Powell R (2017) Research bets and behavioral IR. *International Organization* 71(S1): S265–S277.
Quackenbush SL (2004) The rationality of rational choice theory. *International Interactions* 30(2): 87–107.
Rabin M (1993) Incorporating fairness into game theory and economics. *American Economic Review* 83(5): 1281–1302.

Rathbun BC (2007) Uncertainty about uncertainty: understanding the multiple meanings of a critical concept in international relations theory. *International Studies Quarterly* 51(3): 533–557.
Rathbun BC (2008) A rose by any other name: neoclassical realism as the logical and necessary extension of structural realism. *Security Studies* 17(2): 294–321.
Rathbun BC (2019) *Reasoning of State. Realists, Romantics and Rationality in International Relations*. Cambridge: Cambridge University Press.
Rathbun BC, Kertzer JD and Paradis M (2017) Homo diplomaticus: mixed-method evidence of variation in strategic rationality. *International Organization* 71(S1): 33–60.
Rich N (1973–1974) *Hitler's War Aims*. 2 volumes. New York: W.W. Norton & Co.
Richardson JL (1988) New perspectives on appeasement: some implications for international relations. *World Politics* 49(3): 289–316.
Rosenau J (1961) Pre-theories and theories of foreign policy. In: Farrell RB (ed.) *Approaches to Comparative and International Politics*. Evanston, IL: Northwestern University Press, pp. 29–92.
Rothschild M and Stiglitz JE (1976) Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90(4): 629–649.
Samuelson PA (1965) Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6(2): 41–49.
Saunders E (2017) No substitute for experience: presidents, advisors, and information in group decision making. *International Organization* 71(S1): S219–S247.
Schelling T (1966) *Arms and Influence*. New Haven, CT: Yale University Press.
Selten R (1978) The chain store paradox. *Theory and Decision* 9(2): 127–159.
Shapiro C and Stiglitz JD (1984) Equilibrium unemployment as a worker discipline device. *American Economic Review* 74(3): 433–444.
Shiffrin RM and Schneider W (1977) Controlled and automatic human information processing: II. perceptual learning, automatic attending, and a general theory. *Psychological Review* 84(2): 609–643.
Simon HA (1957) *Models of Man: Social and Rational*. New York: Wiley.
Simon HA (1976) From substantive to procedural rationality. In: Kastelein TJ, Kuipers SK, Nijenhuis WA, et al. (eds) *25 Years of Economic Theory*. Leiden: Martinus Nijhoff, pp. 65–86.
Sinayev A (2016) *Dual-system theories of decision making: analytic approaches and empirical tests*. Doctoral Dissertation. Available at: http://rave.ohiolink.edu/etdc/view?acc_num=osu1471296200 (accessed 10 October 2020).
Singer JD (1961) The levels of analysis problem in international relations theory. *World Politics* 14(1): 77–92.
Sloman SA (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119(1): 3–22.
Stanovich KE and West RF (2000) Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences* 23(5): 645–665.
Stein JG (2017) The micro-foundations of international relations theory: psychology and behavioral economics. *International Organization* 71(S1): S249–S263.
Stigler G (1961) The economics of information. *Journal of Political Economy* 69(3): 213–225.
Stiglitz JE (1979) Equilibrium in product markets with imperfect information. *American Economics Review* 69(2): 339–345.
Taliaferro JW (2000) Security seeking under anarchy: defensive realism revisited. *International Security* 25(3): 128–161.
Teichman D and Zamir E (2020) Nudge goes international. *European Journal of International Law* 30(4): 1263–1279.

Thaler RH and Sunstein CR (2009) *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York: Penguin Books.
Thaler RH, Sunstein CR and Balz JP (2012) Choice architecture. In Shafir E (ed.) *The Behavioral Foundations of Public Policy*. Princeton, NJ: Princeton University Press, pp. 428–439.
Thompson JR, Williams EE and Findlay III, MC (2003) *Models for Investor in Real World Markets*. Hoboken, NJ: Wiley.
Trachtenberg M (1989) Strategic thought in America, 1952-1966. *Political Science Quarterly* 104(2): 301–344.
Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice. *Science*: 211452–211458.
Tversky A and Kahneman D (1984) Choices, values, and frames. *American Psychologist* 39: 341–350.
Tversky A and Kahneman D (1986) Rational choice and the framing of decisions. *Journal of Business* 59(4): S251–275.
van Aaken A and Broude T (2019) Symposium: the psychology of international law. *European Journal of International Law* 30(4): 1225–1357.
van den Assem MJ, van Dolder D and Thaler RH (2012) Split or steal? Cooperative behavior when the stakes are large. *Management Science* 58(1): 2–20.
von Neumann J and Morgenstern O (1944) *The Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
Walker JL (1966) Ballot forms and voter fatigue: an analysis of the office bloc and party column ballots. *Midwest Journal of Political Science* 10(4): 448–463.
Waltz K (1959) *Man, the State, and War*. New York: Columbia University Press.
Waltz K (1979) *Theory of International Politics*. New York: Random House.
Weeks JL (2014) *Dictators at War and Peace*. Ithaca, NY: Cornell University Press.
Wolfers A (1962) *Discord and Collaboration: Essays on International Politics*. Baltimore, MD: Johns Hopkins University Press.
Zermelo E (2010 [1913]) Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels (On an application of set theory to the theory of the game of chess). In: Ebbingaus H-D, Fraser CD and Kanamori A (eds) *Ernst Zermelo—Collected Works/Gesammelte Werke. Schriften der Mathematisch-naturwissenschaftlichen Klasse der Heidelberger Akademie der Wissenschaften*, vol. 21. Berlin: Springer Verlag, pp. 260–273.

Author biography

James W. Davis is Professor of International Relations in the Department of Political Science at the University of St. Gallen, Switzerland. A former Associate Editor of *European Journal of International Relations* and *Security Studies*, he holds a PhD from Columbia University. His research focusses on the psychology of foreign policy decision-making.