Intelligent Management of Intelligence Agencies

Beyond Accountability Ping-Pong

Philip E. Tetlock and Barbara A. Mellers
University of Pennsylvania

The intelligence community (IC) is asked to predict outcomes that may often be inherently unpredictable—and is blamed for the inevitable forecasting failures, be they false positives or false negatives. To move beyond blame games of accountability ping-pong that incentivize bureaucratic symbolism over substantive reform, it is necessary to reach bipartisan agreements on performance indicators that are transparent enough to reassure clashing elites (to whom the IC must answer) that estimates have not been politicized. Establishing such transideological credibility requires (a) developing accuracy metrics for decoupling probability and value judgments; (b) using the resulting metrics as criterion variables in validity tests of the IC's selection, training, and incentive systems; and (c) institutionalizing adversarial collaborations that conduct level-playing-field tests of clashing perspectives.

Keywords: expert judgment, accountability, intelligence analysis

Intelligence agencies are under intense pressure to predict the arguably unpredictable. The United States government does not spend tens of billions of dollars on its sprawling network of intelligence agencies just to be told their best guess is that history will stay on its current trajectory—and the best predictor of the future is probably the present. The core consumers of intelligence analysis, policy makers in the executive and legislative branches, want the intelligence community (IC) to tell them something they didn't already know. And they want guidance most when they suspect the world is transitioning out of a geopolitical equilibrium (they think they have figured out) into turbulence (they think might surprise them).

From this vantage point, intelligence agencies—and the thousands of analysts who work inside them—are set up for failure. Research on experts in nonclassified settings suggests that the outcomes that governments ask intelligence analysts to forecast range from the very difficult to predict (Jervis, 2010) to the virtually impossible to predict (Taleb, 2007). In periods of stability, experts are hard pressed to out-predict simple extrapolation algorithms, and in periods of turbulence, experts are hard pressed to out-predict random guessing strategies (Armstrong, 2005; Tetlock, 2005). Among macroeconomic forecasters, few in 1980 anticipated the extraordinary three-decade expansion of the Chinese economy; few in 1990 anticipated the two-decade stagnation of the Japanese economy; and only a few came close to anticipating the financial crises that rocked the global economy in 2008. In geopolitics, the record is no more impressive. Few predicted the glasnost policies of Mikhail Gorbachev in the late 1980s, or the disintegration of the Soviet Union in 1991, or the rise of fundamentalist Islam. And, of course, no one outside Al Qaeda anticipated the attacks of September 11, 2001—indeed, if someone in an intelligence agency had possessed actionable data, the attacks would not have occurred, an example of the self-negating prophecies that arise in intelligence analysis. Examining the uninspiring collective track record of expert predictions (Gardner, 2010; Tetlock, 2005), it is tempting to dust off an old Marxist aphorism: When the train of history hits a curve, the intellectuals (including the Marxists) fall off.

When the IC cannot satisfy often unreasonable performance expectations, the blame game ensues. Critics point accusatory fingers at policy makers who then point at the IC. We hear angry demands for greater accountability followed by indignant denials of incompetence followed by sharp scrutiny of the denials (Betts, 2009; Jervis, 2010; Posner, 2005; Tetlock, 2000; Zegart, 2007). This blame game has been evident in the last decade's exchanges over the failures to predict the 9/11 attacks (an error often attributed to a failure to connect the dots) and to verify beyond doubt the existence of weapons of mass destruction (WMD) in Iraq (an error often attributed to overconnecting the dots; Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction, 2005; National Commission on Terrorist Attacks Upon the United States, 2004; Office of the Director of National Intelligence, 2008, 2009).

"Accountability" is, however, more mantra than panacea. The frantic post-9/11 effort to redesign accountability systems, in response to Presidential-Commission, post-mortem promptings for more collaboration among agencies, has underscored both how hard it is to overcome

This article was published Online First August 8, 2011.

Philip E. Tetlock and Barbara A. Mellers, Department of Psychology and Wharton School of Business, University of Pennsylvania.

We appreciate the helpful insights of Jonathan Baron, Baruch Fischhoff, Daniel Kahneman, Don Moore, Roxy Silver, and Nassim Taleb—but the views expressed here are, of course, ours alone.

Correspondence concerning this article should be addressed to Philip E. Tetlock or Barbara A. Mellers, Department of Psychology, University of Pennsylvania, Solomon Labs, 3720 Walnut Street, Philadelphia, PA 19104. E-mail: tetlock@wharton.upenn.edu or mellers@wharton.upenn.edu

Vol. 66, No. 6, 542–554 DOI: 10.1037/a0024285

Philip E. Tetlock

institutional inertia and how impossible it is to find trade-off-free solutions (National Research Council, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security, 2011). Congressional acceptance of the Presidential Commission's advice led to the formation of the Office of the Director of National Intelligence, which has mandates to promote cross-agency collaboration and data sharing but still confronts the same knotty trade-offs that existed prior to 9/11, each one creating new opportunities for hindsight-tainted second-guessing (Posner, 2005; Tetlock & Mellers, 2011). Consider two of the most vexing trade-offs:

1. Centrally coordinated data sharing and accountability make it easier to overcome bureaucratic problems of silo-ization that impede the free flow of information and the connecting of dots, but decentralized data-collection operations make it easier to launch creative searches for dots worth connecting. Decentralized need-to-know operations also reduce the risk that, if there are security breaches, the effects will ramify rapidly through the system—a risk illustrated by the 2010 Wikileaks fiasco, in which Pfc. Bradley Manning, stationed in a remote outpost in Iraq, surreptitiously downloaded from SIPRNet, a shared cross-agency network, thousands of confidential American national-security documents and released them to an Australian activist. The irony is that the most trumpeted post-9/11 solution to silo-ization—inducing "stove-piped" agencies to talk to each other—led to SIPRNet, which transformed a small leak into a massive hemorrhaging of embarrassing information. The Central Intelligence Agency (CIA) decision not to join SIPRNet now looks prescient, but the CIA was initially criticized for failing to internalize the post-9/11 reformist spirit. Once again, the blame game took an abrupt turn.

2. Although post-mortem commissions worry about the politicization of intelligence analysis (Betts, 2009), most recently in light of the nonexistence of WMD in Iraq, no commission has offered a clear definition of politicization, much less spelled out how to manage the trade-off between the clashing institutional-design goals of responsiveness and independence. Responsiveness requires that the IC honor all legitimate requests of elected leaders, whereas independence requires insulating the IC from illegitimate manipulation by their policy masters. There is, however, a fine line between legitimate requests to take a second look and illegitimate requests to "politicize" intelligence by skewing interpretations to justify preferred course of action. How will we react the next time high-level officials pay serious attention to low-level intelligence analysts: deplore their intrusion or applaud their due diligence? The safest prediction is that reactions will divide along the usual partisan lines.

Given the intricacy of these trade-offs and the crudeness of the second-guessing, one need not be a chronic pessimist to worry that the IC has become the ball in a game of accountability ping-pong: One set of critics slams agencies for false-positive errors and then another set slams agencies for false negatives. Drawing on signal detection theory (McClelland, 2011; Swets, Dawes, & Monahan, 2000), we call this oscillation the beta-shift cycle in which agencies respond to superficial accountability demands with superficial adjustments of their response thresholds for warnings. For instance, a cycle might begin with the IC reacting to criticism of a false-negative error by lowering its threshold for alerts, setting itself up for false-positive errors, which give it a reputation for "crying wolf," a reputation it tries to neutralize by raising both its threshold and its willingness to incur another false-negative error.

Sociological work in the neo-institutionalist tradition suggests that this ping-pong type of accountability is likely to produce a shell-shocked, blame-averse organizational culture that tries to shield itself from a capricious environment by creating buffer bureaucracies that symbolize the organization's commitment to shared values but accomplish little else (Meyer & Rowan, 1977). Accountability ping-pong is also likely to encourage short-term forms of thinking (do what it takes to get them off our backs) that takes a cynical view of programmatic efforts to improve accuracy on transparent performance metrics (McGraw, Todorov, & Kunreuther, 2011; Tetlock & Mellers, 2011). Indeed, accountability ping-pong may push organizations in the opposite direction toward obfuscation of track records: Better to reduce than to increase the targets for the second-guessers.

We recognize that, in an intensely competitive, pluralistic democracy, it may be impossible to completely extricate intelligence analysis from this blame-game predicament. Nonetheless, the central premise of this article is that the potential benefits are large enough that even a modest probability of partial success yields an expected return that justifies the gamble. The IC does not have to lower the probability of multibillion-dollar fiascoes by much to recoup a multimillion-dollar investment. In this

Barbara A. Mellers

sense, our prescriptions are in the spirit of Taleb's (2011) arguments for robustness as a guiding precept in social-system design in low-predictability environments: containing the IC's downside vulnerability to reactive leadership while retaining its upside capacity to thrive under proactive leadership.

We devote the remainder of this article to proposing a three-step extrication process:

1. Brokering bipartisan agreement that current hindsight-tainted "methods" of evaluating IC performance are problematic and that relentlessly partisan second-guessing can transform a tough problem into an impossible one;

2. Developing and institutionalizing metrics for gauging the accuracy of forecasts—metrics the IC actually uses as criterion variables in validity tests of its selection, training, performance-appraisal, and aggregation systems, and validity tests the IC uses in deciding which practices add or subtract value. The more transparent this process is to the legislative and executive branches of government, the more credible the IC's commitment will be to moving beyond accountability ping-pong;

3. Acknowledging that controversies over intelligence estimates are driven not just by factual disputes but also by competing values, such as the relative importance of avoiding false-positive-versus-negative errors. Building trans-ideological credibility requires admitting that although perfect value neutrality is a noble ideal, there is inevitably an ideological component in intelligence analysis. It is silly to pretend to be something that is implausible, if not impossible: a collection of people who have mysteriously transcended the political allegiances that cloud the judgments of ordinary mortals. The analytical-transparency agenda needs to be supplemented by an adversarial-collaboration agenda that encourages policy elites to engage rather than second-guess the process—and advance testable hypotheses that, they agree in advance, will be subjected to level-playing-field tests with the potential to change their minds.

Step 1: Agreeing on the Feasibility and Desirability of Escaping the Blame Game

Intelligence agencies are supposed to steer clear of domestic politics, but there is no reason why the IC cannot periodically step back to reflect on the parameters of the blame-game predicament in which it finds itself enmeshed. Indeed, the IC has an obligation to do so for two intertwined reasons. First, the IC has a formal mandate to provide timely and accurate intelligence estimates to the policy community. Second, the IC's accountability relationship to the policy community can affect accuracy via its influence on how the IC processes information and communicates assessments. Ping-pong accountability pressures the IC to focus its collective analytic resources on defensive bolstering (justifying previous judgments) and strategic attitude shifting (providing the answers policy makers want; Lerner & Tetlock, 1999). By contrast, the collaborative, hypothesis-testing forms of accountability proposed here are designed to redirect these resources into constructive self-criticism focused on exploring methods of measuring and improving overall performance (Tetlock & Mellers, 2011).

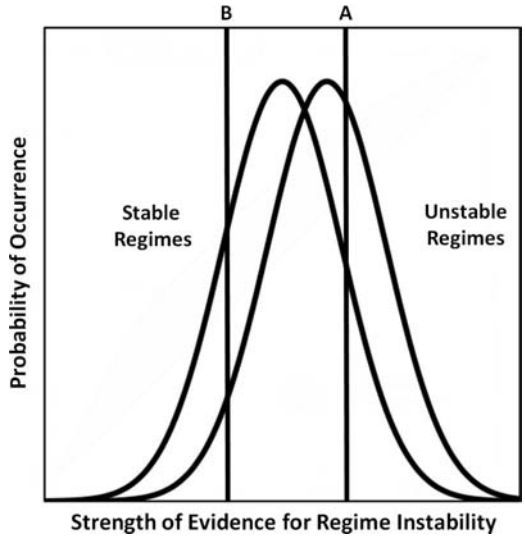
This subtle but crucial distinction—between what signal detection theorists call beta and d prime—will tax the patience of busy elites. But there is no substitute for a solid conceptual foundation for reform. Our starting point is an idea on which virtually everyone agrees: Forecasting performance is constrained by the difficulty of the task.

Figure 1 builds on this simple insight. The task is to distinguish stable from unstable authoritarian regimes. The left distribution captures the information profile for stable regimes, and the right distribution captures the profile for unstable ones. In some cases, analysts can easily infer whether an observation has been sampled from the stable or unstable distribution. But many cases fall in the overlapping zone, where the evidence is ambiguous (coin toss as to which type of country) or even misleading (stability looks like instability and vice versa).

Note that analysts cannot avoid errors when evidence falls in this zone, no matter how much we press them. But analysts can choose which types of errors they will make. Imagine two forecasters who are equally skilled at discriminating instability from stability but, in response to being bashed in different rounds of accountability ping-pong, have different error-avoidance priorities. Forecaster B sets his threshold to tolerate many false alarms to avoid just one miss—and Forecaster A sets his threshold to tolerate many misses to avoid just one false alarm.

Figure 2 plots the resulting hit rates [p("instability prediction" | instability in reality)] and false alarm rates [p("instability prediction" | stability in reality)], in a world equally populated with stable and unstable regimes. A

Figure 1
Two Hypothetical Probability Distributions of Evidence, One for Stable and the Other for Unstable Regimes



Note. Greater overlap means a harder forecasting task.

hypothetical perfect forecaster would reside in the upper left corner, achieving a 100% hit rate at 0% cost in false alarms. Dart-throwing-chimp forecasters would fall along the chance–accuracy diagonal. The two forecasters from Figure 1 fall on the same iso-accuracy curve because of their identical ability to extract signals about regime stability, but they fall in different places on the curve because of their differential distaste for false positives and false negatives. The cross-hatched area between the curve and the diagonal represents the value the forecasters add beyond chance.

Figure 3 captures the potential opportunity costs of accountability ping-pong. When Forecasters A and B shift from ping-pong to transparent-metrics accountability, they move from the lower to the higher iso-accuracy function (see arrows). Note that they do not change their error-aversion priorities. But both can now deliver more hits with fewer false alarms, a Pareto improvement that should offend no rational faction. The shaded area between the higher and lower curves represents one hypothesis about the magnitude of the opportunity costs of sticking with the status quo and losing the boosts to accuracy from exploring methods of improving long-run accuracy. The dotted curve represents one hypothesis about the location of the optimal forecasting frontier, a function that captures maximum hypothetical performance and is bounded only by the irreducible indeterminacy of history (captured in a signal detection framework by the degree of overlap of the functions in Figure 1).

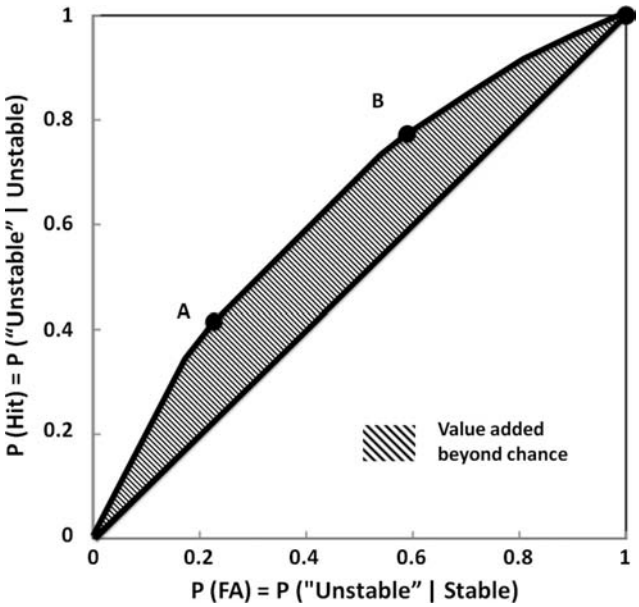
All of the functions in Figure 3, save the diagonal, are purely speculative. And the IC should begin its conversation with the policy community on this note: by stressing how profoundly ignorant we currently are. No one knows which iso-accuracy curves best capture IC performance—or how much room for improvement exists before we hit the optimal frontier. (also a function of the overlap of functions in Figure 1). Moreover, we will lack even rough answers to these fundamental questions as long as accountability ping-pong remains the dominant method of evaluating IC performance.

To avoid setting off a blame game over who is responsible for the blame game, this initial conversation should unfold in closed meetings that make it easier to discuss why it is so easy to start, and hard to stop, playing accountability ping-pong. A full explanation would highlight six mutually-reinforcing psychopolitical processes, each difficult but not impossible to check:

1. *Hindsight distortions.* People who learn an event has occurred often exaggerate the degree to which they saw it coming all along (Arkes, 2001; Fischhoff, 1975; Wohlstetter, 1962). One implication is that it is likely to be extraordinarily hard for consumers of intelligence products, once contaminated by outcome knowledge such as Pearl Harbor or the 9/11 attacks, to recall their before-the-fact probabilities after the fact (Hawkins & Hastie, 1990). Once history has connected the dots for us, we easily forget how nonobvious the connections were.

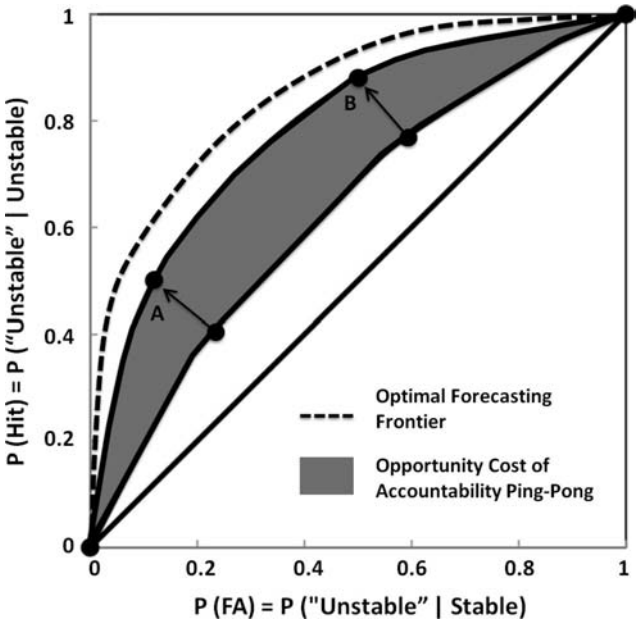
2. *Outcome bias.* In a related vein, people often judge decision makers by how well their decisions worked out

Figure 2
Trade-Offs That Forecasters Might Strike Between Hits and False Alarms



Note. A and B are equally adept at distinguishing stable from unstable regimes but have differential distaste for false alarms and misses. Accountability ping-pong can move A and B up or down the iso-accuracy curve (constant accuracy) but not to a higher function.

Figure 3
Hypothesized Opportunity Costs of Accountability Ping-Pong: The Shaded Area Between the Two Iso-Accuracy Functions



Note. When they move to the higher function, Forecasters A and B do not change their error-aversion priorities, but they can deliver more hits with fewer false alarms. The dotted curve is the hypothetical upper bound on accuracy.

rather than by how rigorously the decisions were made (Baron & Hershey, 1988). This implies that, no matter how impressive the underlying analytics, intelligence agencies are more likely to be blamed for "getting it wrong" than praised for rigorously thinking through the intricacies of the problem.

3. *Motivated reasoning.* Political partisans often do not recognize when they are applying lenient standards of evidence and proof to ideologically congenial claims (sometimes as lenient as the "can-I-believe this" test) and demanding standards for ideologically discomfiting claims (sometimes as demanding as the "must-I-believe this" test; Kunda, 1990; Gilovich, 1991).

4. *Naïve realism.* Even smart people tend to be egocentric epistemologists who make unwarranted assumptions about the veridicality of their own perceptions, which lead them to make strong inferences about the rationality and good faith of those who see the world as they do (Griffin & Ross, 1991) and about the irrationality and bad faith of those who do not. Naïve realism can convince us there is little value in trusting or cooperating with those with different views (Robinson, Keltner, Ward, & Ross, 1995), which can set the stage for a biased-perception conflict spiral that hardens each side's inherent-bad-faith suspicions of the other (Jervis, 1976; Kelman & Bloom, 1973).

5. *Miscalibration of confidence.* Falsificationist models of science call on us to disprove—not prove—our hypotheses. But that mode of thinking does not come naturally (Gilbert, 1991)—and people often display a confirmation bias in hypothesis testing (Klayman, 1995; Nickerson, 1998; Tetlock, 2005), which can fuel overconfidence (Fischhoff, Slovic, & Lichtenstein, 1977). In a typical overconfidence study, participants answer difficult true–false questions, such as "The population of London is greater than that of Paris." If completely unsure of their answers, they should say "50%," and if absolutely sure, they should say "100%." Average confidence across challenging questions usually substantially exceeds the percentage correct (Fischhoff et al., 1977). Overconfidence has also been documented in real-world tasks among political scientists (Tetlock, 2005), physicians (Christensen-Szalanski & Bushyhead, 1981), clinical psychologists (Oskamp 1965), lawyers (Goodman-Delahunty, Granhag, Hartwig, & Loftus, 2010), negotiators (Neale & Bazerman, 1985), scientists (Bruine De Bruin, Fischhoff, Brilliant, & Caruso, 2006; Henrion & Fischhoff, 1986), security analysts (Stael von Holstein, 1972), and eyewitnesses (Sporer, Penrod, Read, & Cutler, 1995). Only a few groups—such as meteorologists (Murphy & Winkler, 1984) and racetrack handicappers (Ceci & Liker, 1986)—have proven well calibrated. And their superior performance offers valuable hints for improving intelligence analysis.

6. *Amplification of biases under threat.* The threat-rigidity literature suggests that the effects listed above become more pronounced under time pressure and cognitive load (Staw, Sandelands, & Dutton, 1981; Suedfeld & Tetlock, 1977)—as well as under postdecisional accountability (Tetlock, 1992). The grounds for pessimism grow still stronger when we couple this work with evidence from political science that competition between the parties has grown more intense and voting patterns in Congress more polarized (Theriault, 2006, 2008). The more polarized the dispute, the lower the chance is of crafting integrative frameworks within which elites can explore why they disagree and how they might reconcile their points of view.

Bundled together, these process obstacles shed light on why elites find it so tempting to blame the IC for forecasting failures and invoke incompetence or malevolence as explanations for estimates that contradict their preconceptions. And bundled together, these obstacles look so formidable that there seems little chance that even inspired bipartisan leadership can overcome them. However, there are also grounds for optimism. Kahneman and Klein's (2009) effort to synthesize work on naturalistic decision making with work on judgmental biases suggests that those analysts who get timely accuracy feedback on recurring problems should develop more bias-resistant intuitions. There are repeated opportunities for learning how to direct Predator-drone attacks on enemy camps or coordinate Navy Seal assaults on Al Qaeda leaders, but not for predicting the outcome of the quirky third-generation dynastic succession in North Korea. And experimental work on debiasing exercises also offers some hope: Asking people to explain the opposite of the observed outcome can

check hindsight bias (Arkes, Faust, Guilmette, & Hart, 1988; Sanna & Schwarz, 2003) and outcome bias (Tetlock, Mitchell, & Anastasopoulos, 2011); and obtaining explicit probability judgments and providing timely feedback can check overconfidence (McGraw, Mellers, & Ritov, 2004) and promote accuracy (Kluger & DeNisi, 1996).

Candor about the difficulty of the task is, however, crucial for inoculating participants against the inevitable setbacks ahead. It is asking a lot of competitive political elites to refrain from impugning intelligence analyses of inevitably uncertain quality when those analyses undercut their preferred policy postures. There is, inevitably, opacity shrouding the creation of intelligence estimates: The analytical process largely remains an impenetrable black box into which unknown sources enter unspecified inputs that are integrated in a mysterious fashion into authoritative-sounding national intelligence estimates that can tip high-stakes debates on whether to use military force, impose economic sanctions, or withdraw support from repressive but friendly regimes. The norm of reciprocity dictates that if the IC is going to ask for painful restraint from the policy community, it will have to make its own painful, identity-transformative concessions that reduce, if not eliminate, suspicions of analytical dice-loading.

None of the essential steps for establishing epistemic trust will be easy for the IC to institutionalize (at the risk of sounding like economists, we suspect that if it were easy to do something this valuable, someone would have already done it). Each step will require overriding deeply ingrained self-protective, organizational reflexes—secrecy reflexes, hedge-your-bets reflexes, obfuscation reflexes, declare-self-above-politics reflexes, autonomy-preservation reflexes, and budget-protection reflexes—that have evolved as adaptive responses to decades of accountability ping-pong. Escaping the blame game will require high-risk, bet-the-farm initiatives that should make prudent executives reconsider how much they want to escape.

Allaying suspicions about politicized intelligence estimates and power grabs will require vastly more transparency from the IC about *exactly* how it prevents its estimates from becoming covert forms of political advocacy. Operationally, that means spelling out how they construct uncertainty bands around intelligence estimates, how they measure their predictive track records, and how those records stack up against open-source competitors, such as prediction markets. It will also require the IC to ensure the representation in the analytical process of politically influential perspectives that feel slighted—always subject to the uncompromising caveat that intelligence analysis is not a game of give-and-take and all perspectives must be treated as hypotheses to be tested on level forecasting fields. Here work on procedural justice and the value of the voice option is of foundational relevance (Lind, Tyler, & Huo, 1997): Elites will exercise more restraint in criticizing dissonant intelligence estimates when they sense their views are being tested fairly inside the IC, and it will be much easier to convince them the tests are fair when the evidentiary ground rules are laid out ex ante.

Even if the Director of National Intelligence (DNI) accepted these recommendations, it remains an open question what the reaction would be to a DNI who tried to engage key players in the Congressional and executive branches in a conversation about the blame game in intelligence analysis. It is instructive to entertain three scenarios about what might unfold:

1. The most upbeat scenario is that the political elites who relentlessly second-guess intelligence analyses are, like fish in the sea, oblivious to the blame-game predicament and its suboptimal properties. But they could be induced to see the need for reform if the IC made a determined case for it.

2. A more plausible scenario is the emergence of a wide range of individual differences in reactions, from open-mindedness to dismissiveness. Some elites are already quite aware of the blame-game predicament and of many of the psychological processes fueling it, such as the hindsight bias (Monday-morning quarterback) and motivated reasoning (the everyday hypocrisies of partisanship). But most elites still have blind spots and, like ordinary mortals, are prone to the naïve-realism illusion that deviations from rationality are far more common on the other side than on their own. In addition, research on cognitive styles warns us to expect that there will be pockets of dogmatism populated by elites who will view any DNI interest in jumpstarting a navel-gazing conversation as a sign of dangerous naïveté—a sign the DNI does not understand how addle-brained those at the other end of policy spectrum are (Tetlock, 2000, 2005).

3. The last scenario offers the grimmest prognosis: Political elites view intelligence analysis as a useful tool when they are in power and as a useful target for criticism when they are out of power. Proposals to escape the blame game strike these observers as evidence that the proposers just do not understand that intelligence analysis is an integral part of the struggle for power, not an exercise in applied science (Fuller, 2010). In this view, the current article is grounded in a misplaced neopositivist faith that there is a capital-T truth lurking beneath the rhetorical posturing of contending factions.

We will never know the truth until a DNI accepts the challenge, but we suspect the more one believes in Hypothesis 3, the more likely one has already stopped reading. To our few remaining readers who lean toward the third hypothesis, we can offer this reassurance: It is politically awkward even for the most Machiavellian elites to be totally dismissive of efforts to create an IC culture that displays an operational, not just rhetorical, commitment to accuracy. Even hardball partisans are supposed to feign an interest in the capital-T truth.

Step 2: Developing Transparent Methods for Evaluating Analytic Performance

From a game-theory perspective, it would be odd if the intelligence and policy communities were not already locked into complementary equilibrium-sustaining strate-

gies for coping with the blame game, strategies that each side has been fine-tuning for decades and strategies that, by definition, would be irrational for either side to change unilaterally (Bueno de Mesquita, 2009). In a blame-game, stochastic world, we should expect a rational IC to gravitate toward a cynical modus vivendi with the policy community that, in essence, takes the following politically comfortable but epistemically indefensible form: If the policy community is going to hold the IC capriciously accountable for unpredictable outcomes, the IC should protect itself by drafting national intelligence estimates that are as informative as it can honestly offer but also as qualification-laden, open-ended, and nonfalsifiable as it can get away with.

From this standpoint, private- and public-sector prognosticators alike must walk the same tightrope. Their livelihoods require sounding as though they are offering bold fresh insights not readily available off the street. But their livelihoods also require never being linked to flat-out mistakes, which would happen if they were rash enough to follow our advice and start making falsifiable forecasts in a stochastic world in which the optimal forecasting frontier may not be much higher than extrapolation algorithms or even random guessing (Armstrong, 2005; Tetlock, 2010). Long-term survival requires mastering the art of appearing to go out on a limb without actually going out on one.

Although high-level officials have occasionally been rash enough to offer falsifiable predictions—such as former CIA director George Tenet's infamous "slamdunk" assurance to the Bush administration about WMD in Iraq—such foolhardiness is the exception, not the rule. The vast majority of intelligence estimates rely on vague-to-the-point-of-nebulous verbal characterizations of the likelihood of outcomes.

Consider this hypothetical but not atypical example: "Although North Korea will *quite likely* continue its policy of threatening rhetoric, punctuated by actual violence, to extort aid, the leadership succession process has added *new elements of uncertainty*, raising the *possibility* of sharp policy shifts in either a confrontational or conciliatory direction. The new leadership will *probably* try to establish its credibility early on because that is when the *risk* of a coup will peak." If nothing else, this statement covers all the scenario bases: It could be spun as prescient if the North Koreans did nothing different or something radically different—anything from launching a nuclear attack on Seoul to launching Chinese-style, economic reform—or if the leadership were overthrown.

Compounding the ambiguity, analysts are adept at attaching open-ended qualifying clauses to already vague probabilistic claims: For instance, "China *might eventually* fissure into regional fiefdoms, *but only if* the leadership fails to manage its security, growth, and legitimacy trade-offs deftly and *only if* global economic growth stalls for a protracted period." This statement sounds informative, but it says no more than that the likelihood of an underdefined scenario materializing—regional fiefdoms—is an unknown function of a series of other underdefined scenarios materializing.

Psychological research has repeatedly shown that these vague "possibility" forecasts set the conceptual stage for big misunderstandings. When researchers ask readers to translate uncertainty language, such as "might," "may," "possibly," and "likely," into probability metrics, they discover a wide array of interpretations (Budescu & Wallsten, 1995). From a decision-theory point of view, these differences are consequential: It matters whether one is assigning a 20% or 80% probability to an outcome that "could" occur—or a split-the-difference 50%. Historically, however, the IC has resisted assigning numerical probabilities to well-defined exclusive and exhaustive scenarios that pass the clairvoyance test (scenarios that one could, in principle, turn over to a genuine clairvoyant who could tell you, with no need for clarification, whether the predicted events will transpire).

From a game-theory perspective, this resistance is a prudent response to the perverse incentives of the blame game: Why put one's forecasting head on the chopping block? And from a public-relations perspective, it is unsurprising that this self-protective explanation is not among those offered by the IC. Official explanations fall into three categories: (a) assigning numerical ratings would imply more rigor than we intend; (b) analytic judgments are not certain so we use (verbal) probabilistic language to reflect the uncertainty; and (c) numerical probabilities would be inappropriate because the IC must grapple with unique events to which probabilities cannot be applied (see National Research Council, 2011, Chapter 2).

The first two arguments collapse under scrutiny. Readers of national intelligence estimates would need to be mind-readers to figure out whether the intended meaning of "quite possible" is one-in-ten or seven-in-ten. If the authors intended such a wide range, they should have said so. If they intended a narrower or even wider range, they should have said that. Given how familiar IC professionals already are with work on the ambiguities of verbal quantifiers of uncertainty, and how long this insight has been percolating at high levels within the IC (Kent, 1951), it is hard not to view this resistance as obfuscation: retreating behind opaque verbiage that, as we shall see, makes it impossible to track relative predictive performance and impossible to tease apart factual and value judgments in intelligence assessments.

The third objection cannot, however, be dispatched so easily. It correctly notes that the IC does not typically deal with recurring events that lend themselves to classification and tabulation in contingency tables (Jervis, 2010). Confronted by sui generis occurrences such as the collapse of the Soviet Union, the emergence of Al Qaeda, or the wobbly third-generation dynastic succession in North Korea, analysts lack the luxury of relying on actuarial tables of conditional probabilities derived from historical base rates of "similar events." And it correctly implies that when events assigned 90% probabilities fail to occur or events assigned 10% probabilities do occur, we do not have a logical warrant for concluding that the assessments were wrong. Unlikely events sometimes happen and likely events sometimes do not. Only those brave or rash enough

to assign zero or 1.0 are at risk of falsification at the individual-case level.

The third objection ultimately rests, however, on a narrow frequentist conception of probability that discounts the value of a subjectivist or Bayesian conception of uncertainty (Bayarri & Berger, 2004) and the benefits of imperfect quantification. The point of translating intelligence estimates into probability scales is not to create a façade of pseudorigor that covers up the imprecision of our knowledge; it is to explore how imprecise our knowledge is—and gauge which training or other interventions make things better or worse. We carve the opportunity cost of forgoing quantification into two broad categories, foregone learning opportunities and foregone opportunities to demonstrate good faith efforts to separate factual from value judgments, each crucial parts of building epistemic trust with the policy community.

The first category of opportunity costs requires passing on the chance to harness the combined power of outcome feedback and the law of large numbers in facilitating organizational learning—and learning how to learn. Quantification allows us to assess how well calibrated and discriminating various clusters of intelligence analysts are when they make repeated judgments of large numbers of well-specified events over long stretches of time. We can then answer previously unanswerable questions. For instance, key calibration questions include the following: When do events that analysts assign, say, 80% likelihood occur 80% of the time, and when less often, as work on overconfidence suggests (Barber & Odean, 2001; Lichtenstein, Fischhoff, & Phillips, 1982) and when more often, as work on underconfidence suggests (Erev, Wallsten, & Budescu, 1994; Moore & Cain, 2007)? Can we identify contextual and individual-difference moderators that predict when various analysts will be better or worse calibrated? Key discrimination questions include the following: How good are analysts at assigning higher probabilities to outcomes that do versus do not occur—and can we identify useful moderators? Are there calibration–discrimination trade-offs? Do some well-calibrated forecasters degrade their discrimination scores by never straying far from minor shades of maybe or far from reference-class, base rates? Can we spot analysts who give us the best of both worlds: good calibration and good discrimination?

When querulous skeptics in Congress or the White House demand to know how accurate the source of a dissonant intelligence estimate is, agencies that quantify expert judgment can do more than engage in hand waving about the professionalism of their staff. They can offer approximate odds ratios, with uncertainty brackets, that tell us how much they would change their minds if various patterns of evidence emerged. And they can invite their interlocutors to offer their own estimates and indicate when even they might change their minds, ideally transforming adversarial encounters into collaborative, Bayesian, problem-solving sessions.

An IC that embraced probability scoring of analysts' judgments would be well positioned to use this metric in correlational and experimental studies that transform bureaucratic dogma into testable hypotheses about the drivers of accuracy. For instance, the IC is sitting on perhaps the world's largest database for assessing linkages between managerial ratings of judgment processes and the accuracy of those judgments. The IC has tacitly placed a massive institutional bet on the validity of its home-grown theory of good judgment: namely, that accuracy should be a positive function of how well analysts conform to the process standards embodied in its performance-management guidelines (designed, in part, to check biases such as overconfidence). If the official theory about how to promote good judgment via process accountability to these guidelines is correct, there should be positive correlations between how favorably supervisors assess analysts' process performance and how accurate analysts' subsequent judgments are.

It would be flattering to the IC if its official theory were validated. But there is no guarantee that it is right or, if right, that it has been implemented effectively. There could be cognitive biases in how managers rate the cognitive performance of analysts. Perhaps they give too favorable process ratings to analysts with whose opinions they agree and too unfavorable process ratings to analysts with whom they disagree. Regardless of outcome, however, it should redound to the epistemological credit of the IC that it voluntarily subjected its core policies to scientific evaluation. The message is this: We care about what works and truth trumps pride. And we hope those whom we advise will be equally open minded about the fallibility of their judgment.

The IC could also use its new accuracy metrics as dependent variables in field experiments. For instance, the IC has invested over many years in a wide range of training systems, collectively known as structured analytical techniques, aimed at checking cognitive biases. These techniques, usually developed by former analysts familiar with the demands of the work and the psychological literature, have face validity (Heuer, 1999; Heuer & Pherson, 2010). But face validity is not construct validity. There is no evidence that structured analytical techniques have the beneficial effects assumed in courses required of all trainees. The Office of the DNI, thus, has another opportunity to signal how committed it is to accuracy—committed to the point of risking the cardinal bureaucratic sin of embarrassing itself to enhance the training of its analysts.

The second class of opportunity costs of resisting metrics requires our passing on the chance to develop analytical methods that can calm the darkest suspicions the policy community harbors of the IC: the power-grab suspicion that analysts are, consciously or unconsciously, smuggling their political values into ostensibly purely factual analyses by inflating the vague verbal probabilities of consequences that cut in favor of policies they prefer (Betts, 2009). Liberal elites often suspect conservative-leaning analysts of inflating the likelihood of WMD in Iraq—to justify going to war—whereas conservative elites often suspect liberal-leaning analysts of deflating the likelihoods of the North Korean and Iranian nuclear programs' achieving key benchmarks—to bolster a dove-ish, pro-negotiation stance.

The IC can never eliminate these suspicions, but it can reduce them by institutionalizing accuracy metrics that are plainly value neutral (Winkler, 1994) and incentivizing analysts to maximize these scores. Probability scoring is value neutral in that it assigns equal weight to underprediction (assigning too low probabilities to events that occur) and overprediction (assigning too high probabilities to nonoccurrences)—and "punishes" those forecasters who smuggle value judgments into their assessments by using likelihood scales to highlight threats they fear policy makers would otherwise miss (Jervis, 2010; Tetlock, 2005).

A transparent, probability-scoring system would flush such epistemic corruption into the open and reinforce the traditional division of labor between the IC (which is supposed to focus exclusively on factual/probabilistic issues) and the policy community (which is supposed to have the final say on values). It is possible, however, to get too much of a good thing. The elected overseers of the IC may not welcome being handed the hot-potato task of making taboo or tragic trade-offs between under- and overpredicting hypersensitive criterion variables such as nuclear proliferation (Tetlock, Kristel, Elson, Green, & Lerner, 2000). Few politicians want to be caught attaching anything other than zero tolerance for underpredicting a radiological bomb attack on an American city—a stance that, taken literally, would translate into the utterly impractical directive to the IC to have infinite tolerance for false alarms. The net result would be to put politicians in a no-win situation—a foreseeable result that some game theorists might note should cause farsighted politicians to want to preserve the politically-stable-albeit-scientifically-indefensible blame game.

A solution may, however, still be possible. Balancing the IC's commitment to analytical transparency and the policy community's understandable aversion to toxic trade-offs may stimulate the invention of new methods of communicating acceptable risk in democracies. Probability scoring can be adjusted to yield value-weighted accuracy metrics that reflect the varying importance that clashing elites attach to avoiding false-positive and false-negative errors (e.g., Christoffersen & Diebold, 2009; Granger & Machina, 2006; Tetlock, 2005).

Intelligence agencies could still maintain a stance of value neutrality but present their overseers with choices among value-weighted probability scores. If elites insisted that even a 1% chance of a false-negative would be intolerable—as Susskind (2006) depicts Vice President Cheney's views on WMD in Iraq—they could take their case to the public but not press analysts for higher estimates. The job of the IC would end with presenting the trade-offs, leaving the elites to thrash out the right trade-off weights in the court of public opinion—a process that, for the most incendiary trade-offs, might paradoxically incentivize elites to exercise restraint in criticizing each other, a domestic-political version of the mutually-assured-destruction equilibrium in the Cold War. The taboo option for elites in a technocracy-assisted democracy would be to iterate back and pressure the IC to alter its assessments. An IC committed to preserving its epistemic integrity and political viability should respond resolutely to such pressure with its always-open invitation to adversarial collaboration, to which we now turn.

Step 3: Embracing Adversarial Collaboration

Thus far, we have sketched a technocratic, neopositivist solution to the problem of establishing transideological credibility. The working assumption has been that, the easier it is for outsiders to see how committed the IC is to measuring and incentivizing value-neutral accuracy goals, the harder it is for outsiders to dismiss dissonant intelligence estimates. That approach is probably not, however, sufficient to allay ideological suspicions of the IC in highly polarized environments. The nature of intelligence analysis means that it can never be completely transparent or reducible to readily reproducible algorithms. There will always be wiggle room for skeptical elites to suspect the worst. In our view, the neopositivist approach, which focuses on rigorous metrics and level-playing-field tests of clashing ideas, needs to be supplemented by a procedural-justice approach, which focuses on reassuring elites who feel slighted that their perspectives are indeed being given a fair hearing.[1]

Of course, a "fair hearing" to the ears of hardball Machiavellians means prevailing. No intelligence agency can guarantee that outcome. But agencies can probably persuade a reasonably wide spectrum of elite opinion that their points of view will be treated respectfully and will be subjected to the same evidentiary ground rules of the adversarial-collaboration process, with the same rights of appeal.

The concept of adversarial collaboration was originally developed by Daniel Kahneman as a superior method of resolving disputes with both friends and foes who were taking aim at various prongs of his multipronged research program on judgment and choice (e.g., Mellers, Hertwig, & Kahneman, 2001; Ariely, Kahneman, & Loewenstein, 2000; Gilovich, Medvec, & Kahneman, 1998). The core idea was that the field would advance faster if, instead of the usual point–counterpoint format of scientific exchanges, each side made a good-faith effort to understand the other's position and reach pre-data-collection agreements on research designs with the potential to change minds. It is unclear how well this approach will scale from scientific to political disputes (see Tetlock & Mitchell, 2009a, 2009b, on how difficult it is to achieve adversarial collaboration even inside scientific psychology), but it is

[1] The two prongs of our approach, the technocratic/neopositivist and adversarial collaboration, correspond to Thibaut and Walker's (1975) distinction between inquisitorial and adversarial methods of resolving disputes. They recommended the inquisitorial method for truth conflicts—differences of factual opinion in which the parties have shared objectives and values and want to discover the optimal approach for pursuing them. They recommended the adversarial approach for conflicts of interest that are not amenable to integrative solutions. The goal shifts from finding the truth to providing a fair procedure for resolving the conflict. Unfortunately, the relationships between the IC and the policy community do not map neatly onto this tidy dichotomy. Most, if not all, of these relationships involve shifting mixes of disagreements over facts and values.

crucial for procedural-justice reasons that the IC conveys a determination to treat all major perspectives on national-security issues in a democratic polity as hypotheses that are worth testing.

The goal of inviting political sparring partners into the IC is to induce them to play by the epistemic norms of the transparency regime rather than by the "street-fight" norms of public campaigning. The invitation should be attractive because it creates an opportunity for outsiders to win greater legitimacy by performing well in adversarial-collaboration tournaments inside the IC. But the invitation comes with a price tag: One must be ready to translate vague hunches about geopolitical trends into probability metrics that can be scored for accuracy and ready to specify the types of evidence that would induce one to change one's mind.

This sounds easy in principle but, in practice, it is not. Imagine a classic deterrence-versus-conflict-spiral dispute over Chinese geopolitical intentions (Jervis, 1976). The deterrence camp sees evidence that China is ready to pursue an expansionist agenda vis-à-vis India, Vietnam, Taiwan, South Korea, and Japan and that the United States should sell sophisticated weapons systems to these nations, invest more heavily in high-technology weaponry itself, and so forth. The conflict spiral camp sees Chinese intentions as essentially defensive and warns that the Chinese government might be driven into a more offensive posture if it senses an American-led conspiracy to encircle China by creating the South and East Asian equivalent of NATO. As we should have learned from the American–Soviet Cold War, these positions are capable of mimicking each other's predictions for extended periods (with conflict-spiral theorists dismissing "offensive" acts as really defensive and deterrence theorists dismissing "defensive" acts as tactical pauses designed to lull us into complacency; Jervis, 1976; Tetlock, 1983).

The challenge confronting adversarial-collaboration coordinators inside the IC is, thus, a higher-stakes version of the challenge confronting such coordinators in the wider scientific community: Require each side to resist the temptation of reducing the other side to a "strawperson," and induce each side to articulate distinctive ex ante testable expectations that, if disconfirmed, would cause it to lower to some degree its confidence in its preferred ideological framework. The ex ante specification is critical because, without it, each side will be free to exercise its demonstrated ex post capacity to explain away dissonant findings.

For instance, if between now and 2014 the Chinese Navy were to act more aggressively in claiming islands and oil rights in the South China Sea than an administration dominated by conflict-spiral theorists expected, those "theorists" would be under a logical obligation—from which it would be embarrassing to renege—to lower their confidence in their assessments of Chinese geopolitical intentions by a Bayesian-specified amount (e.g., .9 to .75). Conversely, deterrence "theorists" would be under a reciprocal obligation to change their minds about the correctness of administration policy if China and its neighbors were to settle oil-rights disputes via a mixture of multilateral negotiations and international arbitration.

Of course, this process is unlikely to produce a miraculous convergence in which doves become hawks and hawks, doves. Defenders of conflict-spiral-informed administration policy could always argue—post hoc—that the unexpected outcome should be attributed to a failure in policy implementation or to not taking conflict-spiral logic far enough. And deterrence-theory critics of the policy could always argue—post hoc—that the unexpected outcome should be credited not to administration policy but rather to unrelated forces. Expert political observers always have the option of retreating into dissonance-reducing historical counterfactuals (Tetlock, 1998, 2005). But such post hoc-ery is embarrassing—and the desire to avoid repeatedly retreating into that defensive crouch should promote a thoughtful, problem-solving focus in conversations across camps. We should not expect the prospect of gradual falsification and slow pundit-career death to have the same wonderful, mind-concentrating powers that Samuel Johnson attributed to the gallows, but even a nudge toward greater realism would be welcome.

Closing Observations

It would be wrong to imply that the IC has merely been the ball in accountability ping-pong—and never tried to extricate itself from the blame game. It has institutionalized many rigorously self-critical practices (National Research Council, 2011). It conducts intensive retrospective assessments of its forecasting failures (Jervis, 2010) and successes (National Research Council, 2011). It makes concerted efforts to hold its professional staff accountable to performance-management guidelines that focus on avoiding inferential biases documented by behavioral scientists (Tetlock & Mellers, 2011). In one sense, we are proposing that the IC follow through further on its avowed commitments to scientific intelligence analysis.

Many insiders are, however, likely to see our proposals as ridiculously naïve, not as incremental adjustments. We need to understand why. The core problem is that the blame game is a systemic predicament and not the fault of any one interest group. It is, thus, easy to make a strong pragmatic case against being a first mover. One risk of embracing transparency and adversarial collaboration is that, rather than scoring credibility points for adopting a rigorously self-critical stance toward its own procedures, the IC will lose points for revealing its mistakes. A second risk is that the probability scores of intelligence analysts will be unimpressive when compared against other forecast-generation mechanisms, such as prediction markets and game-theory models (Bueno de Mesquita, 2009; Wolfers & Zitzewitz, 2004). Yet a third risk is that when the IC invites its critics into collaborations, the critics will often win. The result of these blows to the IC's reputation could be marginalization: a policy community that decides to ignore it. Prudent leaders do not gamble their organization's future on academic speculation.

This analysis helps us to understand the naïveté critique and why escaping the blame game will require leaders

willing to take reputational risks to break out of a suboptimal-equilibrium trap (analogous to breaking out of the defect–defect cell in the Prisoners' Dilemma). But why, skeptics ask, would any rational leader want to do this? We can imagine many reasons, some more and others less noble. Nobility requires leaders ready to be guided by the logic of obligatory rather than consequential action (March & Olsen, 1989), the moral imperative of promoting the public good by raising our collective intelligence. Self-interest requires less: leaders willing to brave the storm for the ego-gratifying prospect of leaving their historical mark.

But it does not matter to our argument whether leaders do the right thing for the "right" or "wrong" reasons. What matters is that, once we institutionalize level-playing-field metrics, we will have a framework for gradually learning how well or poorly each side can predict the consequences of contested policies. Over time, this should make it easier to depolarize previously intractable disputes—easier because, in theory, moderates on each side will be incentivized to bring their probability judgments into alignment with real-world trends and extremists who resist these incentives will accumulate ever-lengthening track records of predictive failures, which will make it ever more implausible to insist that their estimates be taken seriously (people living in glass forecasting houses should not throw stones).

One long-term consequence should, again in theory, be a gradual shift in the balance of power, away from ideologues with weaker forecasting records and toward centrist pragmatists (Tetlock, 2005, 2009). Of course, like anything else, such a trend could go too far. But the core virtue of the proposed system is its capacity for self-correction via iterative hypothesis-testing—as opposed to endless rounds of accountability ping-pong. Politically viable ideologues are not likely always to be wrong and, when they are right, the adversarial-collaboration process should adjust by assigning greater weights to their views.

Finally, it is instructive to compare the credibility-management challenges confronting the IC to those confronting the behavioral-science community when it becomes entangled in policy debates. The similarities are numerous. Both communities take professional pride in their analytical skills and willingness to "speak truth to power"—and are offended when accused of using the mask of objectivity to disguise a value-laden agenda. Both communities recognize, in their reflective moments, that they cannot be 100% sure they are innocent of charges of politicization. Neither can plausibly claim to have an absolute-zero, value-neutrality point against which it can gauge the validity of external critiques of its knowledge-generation practices. And although the first instinct of loyal members of each community is to rally to their professional flag and dismiss outsiders who refuse to admit the error of their ways, it stretches credulity to suppose the insiders are always right. As noted earlier, it is notoriously difficult to define politicization in real-world settings in a perspective-independent fashion, so we should not be surprised that no one has invented an objective method for tallying when an epistemic community has violated canonical norms of objectivity.

These similarities run deep enough that it is worth exploring whether our prescriptions for the IC could be usefully adapted by psychologists in the nonclassified world when they become entangled in policy controversies. Exploring these parallels would require another article, but two recent reports by the National Research Council (National Research Council, Committee on Identifying the Needs of the Forensic Sciences Community, 2009; National Research Council, 2011)—the former on forensic science, the latter on intelligence analysis—point to exportable lessons. Our prescriptions—accuracy metrics for judgment, transparency in hypothesis testing, and invitations to adversarial collaboration—are in the spirit of these reports as well as the Supreme Court's 1993 Daubert (Daubert et al. v. Merrell Dow Pharmaceuticals, 1993) guidelines for distinguishing real from junk science in federal courts. Uncontroversial though the guidelines sound, they have the potential to transform the professional practices of both psychologists in legal-policy disputes and intelligence analysts in national-security debates.

We close by offering a testable sociology-of-science proposition: The closer scientists come to applying their favorite abstractions to real-world problems, the harder it becomes to keep track of the inevitably numerous moderator variables and to resist premature closure on desired conclusions. If true, there is a prescriptive corollary: The more ambiguous and important the applied problem, the more pressing the need for offsetting institutionalized checks, accountability systems of organized skepticism (Merton, 1973), that are committed to transparently balanced standards of evidence and proof. In this view, the struggle of the IC to extricate the blame game is but a special case of a fundamental challenge confronting all forms of applied behavioral and social science.

REFERENCES

Ariely, D., Kahneman, D., & Loewenstein, G. (2000). Joint comment on "When does duration matter in judgment and decision making?" (Ariely & Loewenstein, 2000). *Journal of Experimental Psychology: General, 129,* 524–529. doi:10.1037/0096-3445.129.4.524

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 495–516). Norwell, MA: Kluwer Academic.

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology, 73,* 305–307. doi:10.1037/0021-9010.73.2.305

Armstrong, J. S. (2005). *Principles of forecasting: A handbook for researchers and practitioners.* Boston, MA: Kluwer Academic.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics, 116,* 261–292. doi:10.1162/003355301556400

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54,* 569–579. doi:10.1037/0022-3514.54.4.569

Bayarri, M., & Berger, J. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19,* 58–80. doi:10.1214/088342304000000116

Betts, R. K. (2009). *Enemies of intelligence: Knowledge and power in American national security.* New York, NY: Columbia University Press.

Bruine De Bruin, W., Fischhoff, B., Brilliant, L., & Caruso, D. (2006). Expert judgments of pandemic influenza. *Global Public Health, 1,* 178–193. doi:10.1080/1744169060067394

Budescu, D., & Wallsten, T. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation, 32,* 275–318. doi:10.1016/S0079-7421(08)60313-8

Bueno de Mesquita, B. (2009). *The predictioneer's game: Using the logic of brazen self-interest to see and shape the future.* New York, NY: Random House.

Ceci, S., & Liker, J. (1986). A day at the races: A study of IQ, expertise, and cognitive complexity. *Journal of Experimental Psychology: General, 115,* 255–266. doi:10.1037/0096-3445.115.3.255

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928–935. doi:10.1037/0096-1523.7.4.928

Christoffersen, P. F., & Diebold, F. X. (1997). *Optimal prediction under asymmetric loss. Economic Theory, 13,* 808–817.

Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction. (2005, March 31). *Report to the President of the United States.* Retrieved from www.gpoaccess.gov/wmd/pdf/full_wmd_report.pdf

Daubert, W., et al., Petitioners v. Merrell Dow Pharmaceuticals, Inc. (1993). Supreme Court of the United States. 509 U.S. 579; 113 S. Ct 2786; 125 L. Ed. 2d 469. March 30. 1993, ArguEd. June 28, 1993, DecidEd.

Erev, I., Wallsten, T., & Budescu, D. (1994). Simultaneous over and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527. doi:10.1037/0033-295X.101.3.519

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 288–299. doi:10.1037/0096-1523.1.3.288

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564. doi:10.1037/0096-1523.3.4.552

Fuller, S. (2010). Thinking the unthinkable as a radical scientific project. *Critical Review: A Journal of Politics and Society, 22,* 397–413. doi:10.1080/08913811.2010.541691

Gardner, D. (2010). *Future babble: Why expert predictions fail—and why we believe them anyway.* New York, NY: Penguin Books.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46,* 107–119. doi:10.1037/0003-066X.46.2.107

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York, NY: Free Press.

Gilovich, T., Medvec, V. H., & Kahneman, D. (1998). Varieties of regret: A debate and partial resolution. *Psychological Review, 105,* 602–605. doi:10.1037/0033-295X.105.3.602

Goodman-Delahunty, J., Granhag, P. A., Hartwig, M., & Loftus, E. F. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy, and Law, 16,* 133–157. doi:10.1037/a0019060

Granger, C. W. J., & Machina, M. J. (2006). Forecasting and decision theory. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 81–98). Oxford, England: Elsevier.

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 24, pp. 319–359). San Diego, CA: Academic Press.

Hawkins, S., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107,* 311–327. doi:10.1037/0033-2909.107.3.311

Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics, 54,* 791–798. doi:10.1119/1.14447

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis.* Washington, DC: Central Intelligence Agency, Center for the Study of Intelligence.

Heuer, R. J., Jr., & Pherson, R. H. (2010). *Structured analytic techniques for intelligence analysis.* Washington, DC: CQ Press.

Jervis, R. (1976). *Perception and misperception in international politics.* Princeton, NJ: Princeton University Press.

Jervis, R. (2010). *Why intelligence fails: Lessons from the Iranian revolution and the Iraq war.* Ithaca, NY: Cornell University Press.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515–526. doi:10.1037/a0016755

Kelman, H. C., & Bloom, A. (1973). Assumptive frameworks in international politics. In J. N. Knutson (Ed.), *Handbook of political psychology* (pp. 261–295). San Francisco, CA: Jossey-Bass.

Kent, S. (1951). *Strategic intelligence for American world policy* (2nd printing). Princeton, NJ: Princeton University Press.

Klayman, J. (1995). Varieties of confirmation bias. In J. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of learning and motivation: Vol. 32. Decision making from a cognitive perspective* (pp. 365–418). New York, NY: Academic Press.

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284. doi:10.1037/0033-2909.119.2.254

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. doi:10.1037/0033-2909.108.3.480

Lerner, J. S., & Tetlock, P. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125,* 255–275. doi:10.1037/0033-2909.125.2.255

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York, NY: Cambridge University Press.

Lind, A., Tyler, T., & Huo, Y. (1997). Procedural context and culture: Variation in the antecedents of procedural justice judgments. *Journal of Personality and Social Psychology, 73,* 767–780. doi:10.1037/0022-3514.73.4.767

March, J., & Olsen, J. P. (1989). *Rediscovering institutions.* New York, NY: Free Press.

McClelland, G. (2011). Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. In B. Fischhoff & C. Chauvin (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 83–100). Washington, DC: National Academies Press.

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making, 17,* 281–295. doi:10.1002/bdm.472

McGraw, A. P., Todorov, A., & Kunreuther, H. (2011). A policy maker's dilemma: Preventing terrorism or preventing blame. *Organizational Behavior and Human Decision Processes, 115,* 25–34. doi:10.1016/j.obhdp.2011.01.004

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects: An exercise in adversarial collaboration. *Psychological Science, 12,* 269–275. doi:10.1111/1467-9280.00350

Merton, R. K. (1973). *The sociology of science.* Chicago, IL: University of Chicago Press.

Meyer, J., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *The American Journal of Sociology, 83,* 340–363. doi:10.1086/226550

Moore, D., & Cain, D. (2007). Over and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes, 103,* 197–213. doi:10.1016/j.obhdp.2006.09.002

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79,* 489–500. doi:10.2307/2288395

National Commission on Terrorist Attacks Upon the United States. (2004). *The 9/11 Commission report.* New York, NY: Norton. Retrieved from http://www.9-11commission.gov/report/911Report.pdf

National Research Council, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security. (2011). *Scientific intelligence analysis.* Washington, DC: National Academies Press.

National Research Council, Committee on Identifying the Needs of the Forensic Sciences Community. (2009). *Strengthening forensic science*

in the United States: A path forward. Washington, DC: National Academies Press.
Neale, M. A., & Bazerman, M. H. (1985). The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. *Academy of Management Journal, 28,* 34–49. doi:10.2307/256060
Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220. doi:10.1037/1089-2680.2.2.175
Office of the Director of National Intelligence. (2008). *United States intelligence community information sharing strategy.* Retrieved from http://www.dni.gov/reports/IC_Information_Sharing_Strategy.pdf
Office of the Director of National Intelligence. (2009). *An overview of the United States intelligence community for the 111th Congress.* Retrieved from http://www.dni.gov/overview.pdf
Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29,* 261–265. doi:10.1037/h0022125
Posner, R. A. (2005). *Remaking domestic intelligence.* Stanford, CA: Hoover Institution Press.
Robinson, R., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology, 68,* 404–417. doi:10.1037/0022-3514.68.3.404
Sanna, L. J., & Schwarz, N. (2003). Debiasing the hindsight bias: The role of accessibility experiences and (mis)attributions. *Journal of Experimental Social Psychology, 39,* 287–295.
Sporer, S., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118,* 315–327. doi:10.1037/0033-2909.118.3.315
Stael von Holstein, C. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance, 8,* 139–158. doi:10.1016/0030-5073(72)90041-4
Staw, B. M., Sandelands, L. E., & Dutton, J. E. (1981). Threat rigidity effects in organizational behavior: A multilevel analysis. *Administrative Science Quarterly, 26,* 501–524. doi:10.2307/2392337
Suedfeld, P., & Tetlock, P. E. (1977). Integrative complexity of communications in international crises. *Journal of Conflict Resolution, 21,* 169–184.
Susskind, R. (2006). *One percent doctrine.* New York, NY: Simon and Schuster.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26. doi:10.1111/1529-1006.001
Taleb, N. N. (2007). *The black swan: The impact of the highly improbable.* New York, NY: Random House.
Taleb, N. N. (2011). *Anti-fragility: How to live in a world we don't understand.* New York, NY: Random House.
Tetlock, P. E. (1983). Policy-makers' images of international conflict. *Journal of Social Issues, 39,* 67–86.
Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 331–376). New York, NY: Academic Press.
Tetlock, P. E. (1998). Close-call counterfactuals and belief system defenses: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology, 75,* 639–652. doi:10.1037/0022-3514.75.3.639
Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the ideological beholder? *Administrative Science Quarterly, 45,* 293–326. doi:10.2307/2667073
Tetlock, P. E. (2005). *Expert local judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.
Tetlock, P. E. (2009, September–October). Reading tarot on K Street. *The National Interest,* 57–67.
Tetlock, P. E. (2010). Second thoughts about expert political judgment: Reply to the symposium. *Critical Review, 22.*
Tetlock, P. E., Kristel, O., Elson, B., Green, M., & Lerner, J. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78,* 853–870. doi:10.1037/0022-3514.78.5.853
Tetlock, P. E., & Mellers, B. A. (2011). Structuring accountability systems in organizations: Key trade-offs and critical unknowns. In Committee on Behavioral and Social Science Research to Improve Intelligence Analysis for National Security (Eds.), *Intelligence analysis: Behavioral and social scientific foundations* (pp. 249–270). Washington, DC: National Academies Press.
Tetlock, P. E., & Mitchell, G. (2009a). Adversarial collaboration aborted, but our offer still stands. In B. M. Staw & A. Brief (Eds.), *Research in organizational behavior* (Vol. 29, pp. 77–79). New York, NY: Elsevier.
Tetlock, P. E., & Mitchell, G. (2009b). Implicit bias and accountability systems: What must organizations do to prevent discrimination? In B. M. Staw & A. Brief (Eds.), *Research in organizational behavior* (Vol. 29, pp. 3–38). New York, NY: Elsevier.
Tetlock, P. E., Mitchell, G., & Anastasopoulos, J. (2011). *Embracing or rejecting emerging mind-reading technologies: Intuitive prosecutors play ideological favorites but most stop when caught.* Unpublished manuscript, University of Pennsylvania.
Theriault, S. M. (2006). Party polarization in the US Congress. *Party Politics, 12,* 483–503. doi:10.1177/1354068806064730
Theriault, S. (2008). Party polarization and the rise of partisan voting in U.S. House elections. *American Politics Research, 36,* 62–84.
Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis.* Hillsdale, NJ: Erlbaum.
Winkler, R. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science, 40,* 1395–1405. doi:10.1287/mnsc.40.11.1395
Wohlstetter, R. (1962). *Pearl Harbor: Warning and decision.* Stanford, CA: Stanford University Press.
Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives, 18,* 107–126. doi:10.1257/0895330041371321
Zegart, A. (2007). *Spying blind: The CIA, the FBI, and the origins of 9/11.* Princeton, NJ: Princeton University Press.