Forecasting as a Problem of Cognitive Search: Experimental Evidence from Forecasting Tournaments in the Context of the Automotive Industry

Rahul Kapoor
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104
kapoorr@wharton.upenn.edu

Daniel Wilde
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104
danwilde@wharton.upenn.edu

Version – November 1, 2022

ABSTRACT

Managers and entrepreneurs rely on forecasting as a means to sense opportunities and threats in an evolving industry, and to identify the appropriate course of action. Having superior industry foresight has been deemed as an important enabler for effective decision making. We explore the antecedents of superior industry foresight by conceptualizing the forecasting process as a problem-solving process. Individuals face the problem of forecasting the specific industry context under conditions of significant uncertainty and limited information by searching for relevant information and developing conjectures about the specific industry outcome. We argue that an individual's ability to forecast accurately will depend on the problem's complexity and structure. Forecast accuracy would be highest for low-complexity well-structured problems, lowest for high-complexity ill-structured problems, and intermediate for high-complexity well-structured and low-complexity ill-structured problems respectively. The data for the study were collected from two successive year-long forecasting tournaments conducted between 2017 and 2019, focusing on the evolution of the automotive industry shaped by the emergence of electric and autonomous vehicles. Evidence from over 14,779 forecasts made by nearly 1,395 individuals who participate in a leading global forecasting platform, offers support for our arguments. We also explore how individuals may improve their forecasting by updating their beliefs, and for which types of forecasting problems the belief updating process is likely to be more effective.

Authorship is alphabetical; both authors contributed equally. We gratefully acknowledge our automobile industry expert and collaborator John Paul MacDuffie who was central in helping design and maintain the forecasting tournaments employed in this study. We are also grateful to Dan Levinthal and Phil Tetlock for their valuable guidance on the research. We appreciate the invaluable assistance from Luis Enrique Urtubey De Césaris, Amory Bennett, Philip Decesaris, and their colleagues at Good Judgment Inc. for running the forecasting tournaments. We thank seminar participants at Ohio State University, and Bocconi, and the reviewers for the 2021 Academy of Management conference for helpful reactions and suggestions. Lastly, we would like to gratefully acknowledge the financial support provided by the Mack Institute for Innovation Management at the Wharton School of the University of Pennsylvania and from the Strategy Research Foundation of the Strategic Management Society.

INTRODUCTION

Forecasting the industry context is a key activity performed by managers and entrepreneurs as a basis for sensing the opportunities and threats in an evolving industry, and acting accordingly. The importance of forecasting towards managerial and entrepreneurial decision-making and firm performance has long been underscored in a variety of established strategy perspectives (Porter, 1980; Barney, 1986; Venkataraman, 1997; Eckhardt & Shane, 2003; Teece, 2007; Csaszar & Laureiro-Martínez, 2018; Camuffo et al., 2020). However, the extant literature has underexplored the actual process of individual-level forecasting in an evolving industry context (Peterson & Wu, 2021; Kapoor & Wilde, 2022). This has made it somewhat difficult to link the agency of managers with that of firms' strategy and performance (e.g., Denrell et al., 2003; Gavetti & Lecuona Torras, 2021). In this study, we offer an in-depth examination of the individual-level forecasting process in an evolving industry context.

We consider forecasting as an individual-level problem of cognitive search in which individuals attend to and evaluate relevant information and develop conjectures about future industry outcomes. We draw on Simon's view of problem solving according to which boundedly rational individuals employ cognitive processes to search for problem solutions (Simon, 1973; Fernandes & Simon, 1999). Strategy scholars have built on this conceptualization to explore problem solving at the firm-level as it relates to entrepreneurial opportunity discovery and innovative search (Nickerson & Zenger, 2004; Macher, 2006; Hsieh et al., 2007). We explore problem solving at the individual-level as it relates to forecasting the industry context. Forecasting problems can vary along two key dimensions that have been considered in the extant literature. First, problems can vary in their complexity, ranging from lower complexity with a small number of interconnected variables determining the solution and higher complexity with a large number of interconnected variables (Simon, 1962; Funke, 1991; Jonassen, 2004). Second, they can vary in their structure, ranging from well-structured problems with an unambiguous set of determining variables to ill-structured problems with ambiguous variables (Simon, 1973; Fernandes & Simon, 1999).[1]

[1] Strictly speaking, Simon considers most problems within a representative managerial context to be somewhat ill-structured in nature. However, he outlines how well-structured problems do exist in such contexts when the problem

We consider that solving forecasting problems poses various cognitive challenges to the individual. Forecasting problems of high complexity require an individual to engage with various mental activities including attention and interpretation in order to account for multiple interconnected variables underlying the solution (Simon, 1962; Funke, 1991; e.g., Felin & Zenger, 2014). This cognitive strain can inhibit an individual's ability to accurately solve the forecasting problem. Additionally, forecasting ill-structured problems require an individual to rely heavily on their own perceptions because of the associated ambiguity of the underlying variables and relevant information (Simon, 1973; Fernandes & Simon, 1999; Santos & Eisenhardt, 2009; Laureiro-Martínez & Brusoni, 2018). Perceptions are prone to cognitive bias, which can significantly reduce an individual's effectiveness in solving a forecasting problem (Kahneman & Lovallo, 1993; Kapoor & Wilde, 2022). Taken together, we suggest that the effectiveness with which individuals could forecast a given industry context would depend on the joint consideration of problem complexity and problem structure. High complexity ill-structured problems should have the lowest forecastability, low complexity well-structured problems should have the highest forecastability, and low-complexity ill-structured and high complexity well-structured problems should have intermediate forecastability.

To explore these arguments, we employ an experimental design using a novel method of forecasting tournaments. Forecasting tournaments include carefully designed questions bounded by time on a specific topic (e.g., will a nuclear fusion reactor be in operation in the U.S. by the end of 2025?) on which individuals provide probabilistic forecasts (e.g., 70% "Yes") on possible answers with the goal of being the most accurate. Forecasting tournaments have been successfully used as a methodology to assess foresight primarily in geopolitical contexts (Atanasov et al., 2017; Chang et al., 2016; Moore et al., 2017). We use forecasting tournaments to study forecasting as a problem-solving process in order to assess the forecastability of key industry outcomes in a representative managerial context. The data for the study

is formalized and the problem solver is provided with unambiguous underlying variables necessary to solve the problem. Consistent with this approach, in our research design, we only designate problems as well-structured in cases in which there is strong indication that the underlying variables are reasonably unambiguous for the individual.

were collected from two different forecasting tournaments conducted between 2017 and 2019. The tournaments focused on the evolution of the automotive industry surrounding the emergence of electric and autonomous vehicles. We assembled a unique dataset of 14,779 forecasts made by 1,395 individuals to examine how the complexity and structure of forecasting problems affect their forecastability. These individuals are volunteers motivated by assessing and honing their forecasting skills, typically have several years of work experience, and involvement in decision making across a variety of organizations and industry contexts.

We find strong evidence to suggest that individuals tend to be more accurate in forecasting low complexity well-structured problems, less accurate in forecasting high complexity ill-structured problems, and moderately accurate in forecasting high complexity well-structured and lower complexity ill-structured problems. Additionally, in a series of post-hoc analyses we explore how differences in terms of whether and how individuals update their beliefs based on new information over time may impact their effectiveness at solving certain forecasting problems.

We consider that individuals form initial beliefs (a forecast) regarding a forecasting problem and that they can update their beliefs based on new information that becomes available over time. Belief updating is cognitively demanding in terms of attending to new information, determining its relevance, and incorporating it into an updated belief. We find that on average updating beliefs improves forecasting accuracy for well-structured problems, particularly complex well-structured problems, but not for ill-structured problems. Further, in exploring how individuals update, we find that a belief updating process that is consistent with a Bayesian belief updating helps improve forecasting accuracy for all types of forecasting problems. Additionally, we observe a significant penalty from belief updating that does not align with a Bayesian process for high complexity ill-structured problems.

These findings offer a novel perspective of how forecasting can be viewed as an individual-level problem of cognitive search (Simon, 1973; Fernandes & Simon, 1999; Jonassen, 2004). In so doing, they contribute to the emerging literature stream on how entrepreneurs and managers evaluate the industry context as a basis for strategic decision-making (Felin & Zenger, 2017; Camuffo et al., 2020; Coali et al.,

2022; Ehrig & Schmidt, 2022). This stream has underscored the benefit of theory-based conjectures in developing strategies and the importance of learning and experimentation to systematically adapt strategies over time. We identify how the nature of the strategy problem by itself may shape the relative effectiveness of theory-based conjectures and the importance of learning and experimentation. For example, theory-based conjectures are likely to be less accurate for high complexity well-structured problems than for low complexity well-structured problems. The findings also contribute to the nascent literature exploring antecedents to entrepreneurial and managerial foresight (Gavetti & Menon, 2016; Csaszar & Laureiro-Martínez, 2018; Peterson & Wu, 2021; Kapoor & Wilde, 2022), in terms of when foresight may be more or less likely, and how can it be improved through learning. Further, the study also contributes to the problem solving perspective in strategy (Nickerson & Zenger, 2004; Leiblein & Macher, 2009) by extending the validity of this perspective to individual-level cognitive processes that were the initial basis for Simon's conceptualization (Simon, 1973; Fernandes & Simon, 1999). Finally, by deploying a novel method of forecasting tournaments, the study highlights the benefits of such an experimental research design in developing and testing theories that explain how managers and entrepreneurs may develop conjectures about the industry context as a basis for strategic decision-making and firm performance.

LITERATURE AND HYPOTHESES

Managers and entrepreneurs engage in forecasting to navigate through an evolving industry environment (Porter, 1980; Barney, 1986; Venkataraman, 1997; Eckhardt & Shane, 2003; Teece, 2007). Individuals forecasting the industry context engage in a cognitive process that fundamentally entails problem solving in an evolving industry context with significant uncertainty and limited information. They develop forecasts based on attending to, perceiving, and interpreting available information in conjunction with their cognition. Their cognition consists of their beliefs and simplified mental representations of the world (Walsh, 1995). Mental representations or "mental models" comprise of concepts and relationships regarding how things work in the real world and are the foundations of an

individual's beliefs about alternative states of the future (Holland et al., 1989; Csaszar & Levinthal, 2016; Gavetti & Menon, 2016).

We draw on Simon's original conceptualization of problem-solving premised on individuals facing problems with distinct characteristics (Simon, 1973; Fernandes & Simon, 1999). Under this view, boundedly rational individuals search for solutions to a problem by engaging in various cognitive processes. This view has evolved into the problem solving perspective centered on organizational level search processes (cf. Leiblein & Macher, 2009). Specifically, scholars in this literature have largely focused on heterogeneity of the structure of problems in terms of how differences in organizational forms may be best suited to firm-level search and problem solving (Nickerson & Zenger, 2004; Macher, 2006; e.g., Hsieh et al., 2007). While some scholars have suggested the importance of forward-looking cognitive search processes (e.g., Gavetti & Levinthal, 2000), extant literature has yet to explicitly consider forecasting through the lens of problem solving. Yet, forecasting can be viewed as an individual-level problem solving process in which managers and entrepreneurs search for relevant information and develop a conjecture about future industry outcomes.

The problem-solving perspective offers important distinctions in terms of the nature of questions managers may face. First, problems can vary in terms of their complexity (Funke, 1991; e.g., Felin & Zenger, 2014). Low complexity problems are associated with fewer underlying variables to consider in solving the problem and these variables tend to be less interconnected in terms of how they impact the solution. In contrast, high complexity problems tend to have more underlying variables and have a higher degree of interconnectivity among these variables, like complex systems "made up of a large number of parts that interact in a nonsimple way" (Simon, 1962, p. 468). For example, a lower complexity problem would be solving a common textbook math problem with few interconnected variables and a higher complexity problem would be solving an international political problem involving multiple interconnected parties (Jonassen, 2004).[2]

[2] These conceptions of low versus high complexity are also consistent with notions of P-type (simple) problems that require at most a set of independent variables to solve the problem, and NP-type (hard) problems that require a more

Second, problems vary in their structure, ranging from well-structured problems to ill-structured problems (Simon, 1973; Fernandes & Simon, 1999; Jonassen, 2004). Well-structured problems have clear underlying variables such as relevant information associated with finding the solution and ill-structured problems have poorly defined variables associated with solving the problem. Additionally, while well-structured problems have well-understood problem solving approaches, ill-structured problems tend to be associated with ambiguous solution approaches (Laureiro-Martínez & Brusoni, 2018). To illustrate, a well-structured problem would be making a single move in chess in which the rules and variables to win are well-defined, and an ill-structured problem would be designing a house in which the underlying variables to successfully do so are ambiguous (Simon, 1973).

Collectively, these dimensions engender a wide array of problems that individuals may face ranging from low complexity well-structured problems to high complexity ill-structured problems. While low complexity problems are often well-structured and high complexity problems are often ill-structured, problems can also be low complexity ill-structured such as selecting an outfit to wear and well-structured but complex such as playing an advanced video game (Jonassen, 2004). Thus, in this paper we develop a general framework around the joint structure and complexity of forecasting problems and how their interaction may shape the forecastability of problems, meaning how challenging the problem is to solve.

Forecasting problems and forecastability

In viewing forecasting as a problem-solving exercise, we can consider clear distinctions between forecasting outcomes that embody problems with different degrees of structure and complexity. First, well-structured low complexity forecasting problems are those that have relatively few underlying variables with low interconnectivity (Simon, 1962; e.g., Funke, 1991; Felin & Zenger, 2014). Additionally, these variables are relatively clear and unambiguous to the problem solver (Klein, 1998; Laureiro-Martínez & Brusoni, 2018). To illustrate, consider a CEO of a handset manufacturing company trying to estimate the future sales of the Apple iPhone, a competing product. She would have ample

complex function than P-type (e.g., interactions of variables, exponential functions) to solve, respectively (Moldoveanu, 2009).

access to historical data from which to extrapolate a range of future sales. The underlying variables such as average consumer disposable income and phone features are few in number and quite independent in terms of their effect on future sales. For instance, phone features have very little to do with disposable income and neither of these variables should impact their effect on phone sales.

Second, low complexity ill-structured problems are similarly complex but differ in terms of the ambiguity of underlying variables associated with the problem (Fernandes & Simon, 1999). In other words, relevant information associated with such outcomes may be noisy (Silver, 2012; Kahneman et al., 2021). This is because such forecasting problems are often associated with nascent industry outcomes (Santos & Eisenhardt, 2009), in which there are no clear historical references to draw upon, resulting in significant variation in terms of individuals' mental representations of the underlying variables.[3] For example, consider the aforementioned CEO attempting to project the probability that an emerging handset with unique augmented reality (AR) capabilities will become mainstream in the near future. In this case, there are no clear historical reference points or clear drivers underlying the future outcome from which to form a forecast.

Third, high complexity well-structured forecasting problems have many, well-identified underlying variables that are strongly interconnected in terms of their effect on the forecasted outcome. For instance, consider the previously mentioned CEO is trying to predict whether the U.S. Congress will pass a law that would mandate advanced cybersecurity hardware to be placed on all mobile devices, driving a significant increase in the cost structure of their products. Key underlying variables for this outcome may include (1) the quantity of major cyberattacks caused by breaches in mobile devices, (2) the quantity of users of mobile devices using a particular platform (e.g., Android), and (3) the quantity of firms lobbying the legislature to pursue such legislation, among others. These variables would be quite

[3] We would expect that forecasting problems would generally transition from ill-structured to well-structured over time as the predictive variables and relevant information sources becomes less equivocal. In fact, if the forecasting question regarding iPhone sales were posed during the emerging stages of the smartphone industry in which relevant information regarding iPhone sales was ambiguous, this question would be a good example of in ill-structured low complexity problem. Indeed, in our research design we pose the same forecasting question over a period of years and find that the question often shifts from an ill-structured to well-structured problem over time.

interconnected. For instance, as the number of mobile devices on a given platform increases, the incentive for hackers to pursue cyberattacks in that platform increases and this provides political pressure for Congress to act as they hear complaints from their constituents. Additionally, as cyberattacks increase, a growing number of firms are likely to lobby for the focal legislation because they have been negatively impacted by the cyberattacks, driving further pressure on Congress to pass the law.

Lastly, high complexity ill-structured forecasting problems are those associated with many interconnected but ambiguous variables. These problems are similar to their well-structured high complexity counterparts, but are associated with events with little or no precedence such as technological breakthroughs, regulatory policy of emerging technology. For instance, consider the previous example regarding cybersecurity but suppose the forecast question was posed during an election year or during the nascent stages of the industry in which the underlying variables associated with the regulatory environment was quite ambiguous. In such a setting, the underlying variables would still be numerous and interconnected, but these variables would be generally ambiguous at the time of forecast.

Solving forecasting problems can pose various cognitive challenges for individuals. In terms of ill-structured forecasting problems, individuals must rely heavily on individual perceptions in order to make sense of the inherently ambiguous available variables and information (Klein 1998; Fernandes & Simon 1999; Laureiro-Martinez & Brusoni 2018). Such perceptions are prone to cognitive biases, which should increase the difficulty in solving these types of forecasting problems (Kahneman & Lovallo, 1993; Kapoor & Wilde, 2022). In contrast, well-structured forecasting problems are associated with relatively unambiguous underlying variables and are less reliant on the individual's perceptions because evidence associated with forming a forecast is less equivocal.

In terms of complexity, higher complex problems have a higher level of "landscape ruggedness," and the solution search is cognitively more demanding in order to attend to and interpret the multiple interconnected variables (Gavetti & Levinthal, 2000; Jonassen, 2004; Helfat & Peteraf, 2015). These high cognitive requirements pose challenges to boundedly rational individuals in terms of their ability to effectively solve the problem. Contrastingly, lower complex forecasting problems require the individual

to account for fewer and independent variables, which should impose a much lower cognitive load on the individual.

The joint variables of structure and complexity should impact the general forecastability of a forecasting problem. Low complexity well-structured forecasting problems should be somewhat straightforward to solve because the central task involved in these problems is to pay attention to relatively unambiguous information pertinent to the industry outcome in order to formulate a forecast of the future. In other words, if an individual can attend to new pertinent information associated with the forecast, they can essentially "connect the dots" to get a sense of a likely future outcome. In contrast, high complexity ill-structured forecasting problems should be particularly challenging because they require both high cognitive demands in terms of engaging in mental activities such as attention and interpretation and also rely heavily on individual bias-prone perceptions. Lastly, high complexity well-structured and low complexity ill-structured problems should be moderately challenging given the unambiguous but cognitively demanding task and the ambiguous but cognitively simple task, respectively, associated with these problems. Accordingly, we predict:

> *Hypothesis: The forecastability associated with an uncertain industry outcome would depend on the joint structure and complexity of the forecasting problem: lowest with high complexity ill-structured, greatest with low complexity well-structured, and intermediate in the case of high complexity well-structured and low complexity ill-structured, respectively.*

METHODS

Research setting and experimental design

To test our hypotheses, we employ an experimental design through a novel method of forecasting tournaments, which include carefully designed questions bounded by time on a specific theme in which individuals provide probabilistic forecasts on possible answers with the goal of being the most accurate. Forecasting tournaments have emerged as an effective means of studying issues of cognition and forecasting over time within a realistic business context (Kapoor & Wilde, 2022).

The research setting for the forecasting tournament is the evolution of the automotive industry, stemming from the emergence of electric and autonomous vehicles. The data for the study comes from

two successive forecasting tournaments – first from July 2017 to August 2018, and second from November 2018 to December 2019.[4] Each tournament was designed and managed through collaboration with Good Judgment Inc., an organization that excels in designing and running such tournaments (Atanasov et al., 2017; Moore et al., 2017) and is hosted on the Good Judgment Open (GJ Open) platform, www.gjopen.com.[5] Each of the two tournaments includes a set of carefully selected questions that relate to technological and commercialization progress, and government policies, whose resolution would have a significant impact on the emergence of electric and autonomous vehicles. We went through a rigorous process to generate questions for each of the two tournaments to make sure that each question is relevant to the ongoing transformation of the automotive industry, that each question would be resolvable to the satisfaction of all forecasters, and that all question options could feasibly occur within the question window.

In order to ensure the relevance of each question, we had discussions with several industry experts and performed an extensive review of the literature and the popular press. Additionally, as part of the 2018-2019 tournament we followed a more rigorous question-generation process in which we assembled a panel of experts and the most accurate forecasters from the previous tournament to suggest the most important forecasting indicators toward the long-term evolution of the automotive industry. We then surveyed the panel members to rate each indicator in terms of how much its outcome would impact the long-term trend of the industry. The most informative indicators then served as the basis of the questions in the tournament.[6] Importantly, most of our highly important indicators from this process were

[4] We also ran a shorter pilot tournament during 2016 to learn about this experimental design and its utility for the purpose of our research. We did not include data from this tournament because the GJ Open's question-generation process was in the early stage of development. Nonetheless, as an additional exploration, we incorporated these data in our analysis and found qualitatively consistent results.

[5] Since its inception in 2015, the GJ Open platform has hosted dozens of forecasting tournaments covering geopolitical and other current affairs. The platform has attracted thousands of forecasters with a goal of being the most accurate.

[6] First, the panel anonymously completed a survey submitting potential events (i.e., indicators) over a 1-year horizon, the outcome of which would be indicative of the long-term outlook (7 years out) with respect to EVs and AVs. This process was motivated by Tetlock's recently developed Full Inference Cycle Design for Forecasting Tournaments (Tetlock, 2017).

consistent with questions within the 2017 -2018 tournament, suggesting that our questions have strong relevance to the transformation of the automotive industry. We also had extensive consultations with Good Judgment Inc. to ensure that each of the questions are feasible within the specific time-period and resolvable using publicly-available data.

Given the significant cognitive demands and time commitment associated with engaging in forecasting tournaments, it is important to motivate individuals to participate. The overarching motivation for individuals participating in these tournaments is to be more accurate than their peers, and to improve their forecasting skills in a real industry environment. There are several ways that forecasters can receive recognition on the GJ Open platform including digital certificates (badges) prominently displayed on their profile for being the most accurate within a tournament, and having their names displayed on leaderboards for best accuracy scores on a specific question or the entire tournament. Additionally, if they consistently illustrate superior performance they can earn the prestigious distinction of being a "Superforecaster," which provides new professional opportunities through Good Judgment, Inc.

Dependent variable

Evaluating the forecastability of a problem is difficult to measure, but our research design provides a systematic way of doing so. We assess the challenge of forecasting a given question by measuring the forecast accuracy of each forecast made based on the difference between the forecasted probability (e.g., 80% "yes") and the outcome of the question (e.g., 100% if the event occurred). It is also the same measure that is used to provide feedback to each forecaster on the accuracy of their forecasts relative to their peers on the GJ Open forecasting tournament platform and is driven by prior research (Jose et al., 2009; Tetlock et al., 2014).

Questions within the forecasting tournaments come in two basic forms: (a) *unordered* questions in which there is no distinct order or rank in the option set (e.g., binomial questions such as A = "Yes", B = "No", or multinomial unordered such as A = "Yes, manufacturing without a Chinese partner", B = "Yes, manufacturing with a Chinese partner", C = "No"), and (b) *ordered* questions in which there is clear order or rank in the option set (e.g., A = "Less than 10,000", B = "10,001 - 15,000", C = "15,001 -

20,000", D = "20,001 - 25,000", E= "More than 25,000"). Unordered questions tend to be associated with forecasting of specific ill-structured problems whereas ordered questions are always associated with forecasting of specific well-structured problems.

For unordered questions, the forecast accuracy is calculated as the sum of squared errors commonly called the probability (or Brier) score (Brier, 1950). Formally, the *forecast accuracy* for forecaster *i* on question *q* at time *t* is:

$$forecast\ accuracy = \sum_{i=1}^{r}(f_{ijqt} - o_{jq})$$

where f_{ijqt} is forecaster *i*'s probability forecast for question option *j* for question q at instance *t*, *r* is the number of possible choices in which the event can fall (e.g., "Yes – with condition [A]", "Yes – with condition [B]", "No" would have $r = 3$), and o_{jq} the actual outcome of the question choice *j* of question *q* (equals 0 if it does not happen and 1 if it does happen). For ordered questions, the forecast accuracy is calculated in a very similar manner but we assign partial credit for near-misses, meaning inaccurate forecasts that were close to the resolved answers (Jose, Nau, & Winkler, 2009). [7] The measure can range from 0 to 2, with lower values indicating greater accuracy. See Appendix 2 for details on the forecast accuracy calculations as well as illustrative examples.

Independent variables

Establishing the structure of a forecasting problem is a challenging empirical task. The central distinction between a well- and ill-structured forecasting problem is whether the underlying variables to consider in solving the problem are relatively unambiguous or ambiguous, respectively (Simon, 1973; Fernandes & Simon, 1999). Accordingly, we went through a three-pronged approach to help distinguish between forecast questions that embody more of a well-structured problem versus an ill-structured

[7] This distinction is important because the forecast accuracy calculation outlined above would not distinguish between less and very inaccurate predictions. Consider the question "How many public DC Fast Charge electric vehicle charging stations will be available in the United States?" with options A = "less than 2,400", B = "between 2,400 and 2,700 inclusive", C = "between 2,701 and 3,100 inclusive", and D = "More than 3,100". If the question resolved with A, the method used on unordered questions would not distinguish between probabilities in B, C, or D, but clearly D was much more incorrect than B or even C.

problem. First, we considered that each question within the forecasting tournament platform offered background information for all forecasters. Importantly, experts at GJ Open curated the background information for each question from publicly-available information to ensure that any extant relevant information associated with a question would be included. Thus, we considered that the available background information for a well-structured problem would include unambiguous variables to solve the problem. In contrast, we considered that the available background information for an ill-structured problem would not include clear variables to consider, making the underlying variables to solve the problem somewhat ambiguous.

For example, consider two forecast questions below, question A and questions B. The publicly-available background information associated with question A included data (i.e. historical sales) that serves as an unambiguous underlying variable to help solve the forecast question, consistent with a well-structured problem. In contrast, the only publicly available background information associated with question B was the announcement that motivated the question in the first place, leaving the underlying variables quite ambiguous, consistent with an ill-structured problem:

> Question A (well-structured)
>
> FORECAST QUESTION: "How many Mirais will Toyota sell or lease between January 2018 and June 2018, inclusive?"
>
> BACKGROUND INFORMATION**:** The Mirai is Toyota's first commercially launched fuel cell-powered electric vehicle (Toyata, Forbes, Toyota). Since 2015, Toyota has sold or leased over 3,000 Mirais in the US, all in California (The Drive). Toyota Mirai sales data can be tracked here.
>
> Question B (ill-structured)
>
> FORECAST QUESTION: "A second question was "Before 1 April 2018, will General Motors test an autonomous vehicle in New York City?"
>
> BACKGROUND INFORMATION: Cruise Automation, the self-driving unit of General Motors, recently announced its intention to test autonomous Chevy Bolts in New York City a high regulation, high density location.

Second, we explored structural differences across questions by analyzing voluntary comments that forecasters submitted as part of their forecasts. For instance, comments associated with one forecast

question may reference very similar variables across forecasters implying a more well-structured problem and another question may reference very different variables across forecasters implying a more ill-structured problem. We combed through thousands of forecasts and found clear differences between questions for which the aggregate comments by forecasters converged to a common set of variables (e.g., sales volume last quarter) and those for which the aggregate comments by forecasters tended to vary widely in terms of variables considered across forecasters. Importantly, these two approaches provided very consistent differences across questions in terms of aligning with well- or ill-structured forecasting problems.

Finally, as a third approach, we recruited 121 English-speaking adults based out of the United States through Prolific, a leading research participant recruiting platform that verifies and monitors participants, to independently categorize questions in terms of their structure. To do so, we administered a survey that first provided definitions and examples of ill-structured and well-structured problems based on our theory. Then, we asked the individual to categorize a single question that was randomly selected from four questions, two of which were well-structured and two were ill-structured based on the first two approaches. The vast majority of respondents (109 out of 120) categorized the assigned question consistent with the previous two approaches. In terms of quantifying the reliability of our categorizations, these results corresponded to a Cohen's kappa statistic of 0.91, which is considered almost perfect agreement or reliability (McHugh, 2012). See Appendix 4 for further details. In total, 19 questions were categorized as well-structured and 17 questions were categorized as ill-structured.

Designating the complexity of a forecast question is another difficult empirical challenge (Moldoveanu, 2009). We conceptualize the complexity of a forecast question as a function of the number of variables associated with the question, and the extent to which these variables interconnect (Simon, 1962). In order to capture complexity of a forecast question, we first considered that forecasters can voluntarily offer comments to justify each forecast. In total, our sample includes 4,351 comments

across the 36 questions.[8] Such comments can reasonably represent a forecaster's cognitive representation of the forecasting problem (Csaszar & Laureiro-Martínez, 2018). A central component of an individual's cognitive representation is the perceived complexity of the problem (Gary & Wood, 2011; Csaszar & Levinthal, 2016; Martignoni et al., 2016).

Drawing on natural language processing, we ascertain the complexity of each comment by identifying the variables being considered and the relationship between them (Karvetski et al., 2022; Tetlock et al., 2014). Specifically, we measure a construct called integrative complexity of each comment through a tool called autoIC as the basis for assessing complexity of each comment (Conway et al., 2014; Houck et al., 2014). Integrative complexity (IC) is a composite measure that considers both the degree of differentiation, meaning the number of distinctions made among mentioned variables within the text, (e.g., "on the other hand" indicates at least two underlying variables discussed) and (2) integration, meaning the extent of connections between these differentiated variables (e.g., "in conjunction with" indicates one factor is connected to another). To calculate IC, the program first collects all words or phrases within the text that are found in a dictionary of indicators for differentiation and separately found in a dictionary of indicators for integration. Next, each of these words or phrases are assigned a probability that it indicates either differentiation or integration.[9] The software then provides a score for each differentiation indicator, if any, recording the highest score. If there were differentiation indicators, it provides a score for integration indicators, noting the highest score. These scores are then added together to offer a total IC score ranging from 1 (total lack of differentiation or integration), to 2 - 3 (levels of differentiation), to 4 - 7 (differentiation plus integration). This general approach is consistent with recent research capturing complexity of CEOs' mental models (Graf-Vlachy et al., 2020; Malhotra & Harrison,

[8] This represents approximately 28% of all forecasts when including forecasts removed for individual fixed-effects analysis.
[9] These probabilities are based on training data from expert human scorers

2022), however, this approach has the additional benefit of incorporating phrases instead of just keywords and utilizes probabilities for categorization based on expert coders and trained datasets.[10]

Below is an example of a comment that was used to categorize a question associated with legal precedence for self-driving vehicles as that of low complexity (1.75 out of 7):

> i don't really think there are **enough self driving vehicles on the road** right now for there to be an accident before the alloted time the only crash i could find on a precursory search was one involving uber and though they had **disabled one of the safety features** to prevent erratic behavior from the car neither uber nor the mfgr were liable in the crash this would set a precendent for case law in arizona if another crash happens in arizona i would assume the victim would need to take the case higher than state level.

Notice the relatively few underlying variables driving the outcome (e.g., low quantity of self-driving vehicles on the road, and disabling safety features) and a lack of interconnectedness. In contrast, consider the comment below that was used to categorize a question associated with average industry-wide battery cost as that of high complexity (6 out of 7):

> considering a lot of **capacity in battery production** ia planned but not yet on the market i assume there will come down as soon as there is more **competition** however from this side i dont expect much of a **pressure** until 2020 the other side is the **cost side** combined with the law of more **technical development capability** vs cost and **eco of scale** this will push the **existing players to defend their positions** ahead of new entrants so my best estimate currently is less than 215 is absolutely in the reach i am curious about other's opinion and findings

We see many more underlying variables (e.g., “technical development capability”, “existing players”, “eco of scale”, “cost side”) and evidence of their interconnectedness (e.g., “the cost side” is connected to “technical development capacity”).

This approach to measuring complexity is very consistent with Simon's (1962) two-pronged conception of complexity we theorize. Indeed, the notion of differentiation is a strong indication of the number of variables underlying a problem and the notion of integration is indicative of the interconnectedness of these variables. This methodology has also been used on forecaster comments within forecasting tournaments as a means to measure complexity (Karvetski et al., 2022).

[10] An alternative operationalization of mental representation complexity is based on the number of elements mentioned in a survey (e.g., number of competitors or number of strategies identified in a market) (Mcnamara et al., 2002). However, while this is approach captures differentiation, it does not consider issues of integration consistent with our theory.

We assume that if the average integrative complexity within comments of a forecasting question is high (low), the overall complexity of the forecasting problem is likely high (low) as well. Therefore, we calculated the mean complexity score for each question and if this score was higher than the median score of all questions (i.e., 1.69) we designated the question as high complexity, and as low complexity otherwise. Given the low median score, for robustness we incorporate an alternative operationalize of high complexity based on the percentage of total comments in a question with complexity scores of 4 to 7 out of 7 (Model 6), finding consistent results. Figure 1 outlines the distribution of forecasts and questions that fall into each of the four quadrants, with Quadrant 1 as high complexity ill-structured problems, Quadrant 2 as high complexity well-structured problems, Quadrant 3 as low complexity well-structured problems, and Quadrant 4 as low complexity ill-structured problems.

Insert Figure 1 Here

Appendix 1 includes the list of questions for the two tournaments, with specific question-level details of its structure and complexity (e.g., a high complexity, ill-structured problem is indicated by quadrant "1"), when the question was launched, when it was closed, how it was resolved, how many individuals participated, and the total number of forecasts. The hypotheses are tested on a dataset of 14,779 forecasts made by 1,395 forecasters on 36 questions. These forecasters come from various professional backgrounds that encompass entrepreneurial, managerial, and technical roles. A common characteristic among them is their many years of work experience and their involvement in decision making across a variety of industry settings.[11]

Control variables

We control for a number of individual- and question-level covariates that may affect the accuracy of the focal forecast. First, we consider that forecasters are able to update their beliefs (forecasts) on a

[11] For reasons of privacy, we are unable to determine the specific background and the demographic characteristics of each of the forecasters. However, a small number of forecasters (1,015) do disclose their backgrounds in their user profile on the GJ Open platform, and we used a matching process to identify additional information on 601 forecasters on publicly available sources such as LinkedIn. When controlling differences in experience, we found consistent results to our main findings.

given question over time. Updated forecasts may reflect new information resulting in superior forecast accuracy. We, therefore, operationalize the updating of beliefs through a simple proxy (1/0) variable *belief updated* equaling one if the forecast is an updated forecast on the focal question and zero if not. We also account for differences in uncertainty, meaning the general unpredictability of a specific real world issue (Koopmans, 1957; Packard et al., 2017), faced by the forecaster when she makes her forecast. Higher uncertainty should generally negatively affect forecast accuracy.[12] We measure *uncertainty* based on the forecast accuracy of the daily consensus forecast.[13] The variable is operationalized as the mean of the forecast accuracy of the consensus forecast for each of the three days surrounding the date of the focal forecast.[14]

We account for differences in terms of the forecasters engagement within the forecasting platform during the focal question, which may be effecting the accuracy of the forecast. We do so by including the variable *days active*, which is the log of the number of days the forecaster logged on the GJ Open platform during the months the focal question was available for forecasting on the platform. The variable *questions platform* is the log of the number of total questions answered by the forecaster on the platform during the time that the focal question is available for forecasting. The variable *questions*

[12] Note that ambiguity is distinct from uncertainty. While ambiguity is associated with a lack of clarity in the underlying drivers causing outcomes, uncertainty pertains to the general unpredictability of outcomes (Santos & Eisenhardt, 2009). Thus, an ill-structured problem is by definition ambiguous, but may vary in terms of its uncertainty.

[13] The daily consensus forecast is calculated based on an algorithm developed by Good Judgment Inc. It is calculated by first creating a set of the most recent forecast for each forecaster on the given question as of that day. From this set of forecasts, a subset of the larger of (a) the most recent 40% of the forecasts, or (b) all forecasts from the last 72 hours is generated. The consensus is calculated as the median of this subset of forecasts. The 40% criteria provides a good mix of recent activity and historical perspective from forecasters. The 72-hour criteria is useful during periods of high forecasting activity such as when a new question is launched. During such periods, the use of the 40% criteria would only capture a few hours of forecasting thereby limiting the representativeness of the consensus. The median is used to mitigate the effects of outliers.

[14] An alternative operationalization of uncertainty could be based on the distribution of forecasts made by forecasters on a given question. However, using such distribution-based measures of uncertainty in the forecasting tournament where each forecaster provides probability estimates bound between 0% and 100% with respect to different choices may be problematic. To illustrate, consider a binomial ("Yes"-"No") question in which the collection of forecasts have a very low variance, but in which the mean of forecasts is close to 50% "Yes." In this case, the low variance may suggest low uncertainty as described above, but the central mean in probability may indicate high uncertainty because forecasters are unitedly uncertain whether the question will resolve closer to 0% or 100%.

domain is the count of questions forecasted by a forecaster in a tournament within the same domain (i.e., electric vehicles or autonomous vehicles).

We also control for the number of peer forecasters on a given question through the variable *forecaster count*, which measures the count of other forecasters who forecasted the focal question during the three day window [-1,+1] between one day before and one day after the focal forecast. Higher count of forecasters may provide more information for forecasting, and may help improve the forecaster's accuracy. To test for robustness, we also expand the time windows for forecaster count and uncertainty to five days [-2, +2] and seven days [-3, +3]. Additionally, we control for whether a forecaster submitted a comment with her forecast, which may reflect a deeper analysis underpinning the forecast and therefore forecast accuracy, through the dichotomous (1/0) variable *commented*. We also account for the fact that some forecasters participated in multiple automotive industry forecasting tournaments that we had designed through the dichotomous (1/0) variable *multiple tournaments*. Finally, because individuals can forecast at different times and earlier forecasts may be associated with lower accuracy, we include the variable *days-to-end* as the number of days between when the forecast was made and when the question was closed.

Statistical analysis

We control for unobserved time-invariant differences across individuals, and we estimate the following equation using fixed effects ordinary least squares regression, employing the reghdfe procedure in Stata:

$$y_{iqt} = \beta_0 + \mathbf{X}_{iqt}\beta + \mathbf{F}_i\delta + u_{iqt}$$

where y_{iqt} is forecaster *i*'s forecast accuracy on question *q* at time *t*, $\mathbf{X}_{iqt}$ is the vector of independent and control variables, $\mathbf{F}_i$ is the vector of dichotomous categorical variables for each individual**,** and u_{iqt} is the error term.

Insert Table 1 Here

RESULTS

The descriptive statistics and pairwise correlations for the variables are displayed in Table 1. The results from the regression analyses are reported in Table 2. Note that higher values of the dependent variable imply lower accuracy of the forecast. Model 1 is the baseline model and includes all the control variables. On average, individuals tend to have higher forecast accuracy (lower value) when the forecast is an updated belief (forecast), when uncertainty is lower, and when the individual is more active on the platform. Additionally, we see evidence that forecast accuracy tends to be lower when the individual forecasts more questions within the platform and forecasts more questions in the same domain as the focal question, when there are more individuals forecasting around the forecast, when a comment is included, and when the resolution of the question is more days away.

Model 2 is the full model and includes dichotomous variables for three of the four quadrants of forecasting problems, with quadrant 3 as the omitted variable. The coefficient of *High complexity ill-structured* forecast questions is positive and significant ($p<0.000$), suggesting high complexity ill-structured forecasting problems are significantly more challenging on average than low-complexity well-structured questions. Additionally, the coefficient of *Low complexity ill-structured* and *High complexity well-structured* forecast questions are positive and significant at $p<0.000$, meaning both quadrants are also on average more challenging to forecast than low-complexity well-structured questions. These two estimated coefficients are also lower in magnitude than the coefficient for high complexity ill-structured forecast questions, and Wald tests of the equality of these coefficients with that for high complexity ill-structured forecast questions were both rejected at $p=0.000$ (Table 3). Collectively, these findings strongly support our Hypothesis that the forecastability of forecasting problems would be lowest for problems in quadrant 1 (ill-structured high complexity), highest in quadrant 3 (i.e. well-specified low complexity), and intermediate in both quadrant 2 (well-structured high complexity) and quadrant 4 (ill-structured low complexity). We also observe the coefficient for quadrant 4 is higher in magnitude than quadrant 2 and find that this difference is marginally significant ($p=0.092$), suggesting that forecasting challenges associated with ill-structuredness may be marginally higher than challenges associated with

complexity.

Insert Tables 2 and 3 Here

Robustness Checks

Table 4 outlines a series of additional analyses we conducted to assess the robustness of our results. Table 5 displays robustness results related to the Hypothesis. We first considered that our results may be driven by the fact that most questions that are categorized as well-structured tend to be multinomial ordered in which the forecast accuracy is calculated slightly differently than unordered questions. Model 3 includes a dichotomous (1/0) variable indicating whether the focal question is multinomial-ordered. We also consider that we may find significantly different results using different time windows for calculating the *uncertainty* and *forecasters count* variables. Model 4 applies a five-day window [-2,+2] comprising of two days before and two days after the focal day, and Model 5 applies a seven-day window [-3,+3] comprising of three days before and three days after the focal day. Finally, Model 6 incorporates an alterative classification of high vs. low complexity questions based on the percentage of total comments in a question that has significantly high complexity (i.e., complexity score of 4 - 7 out of 7). The estimates from all of these analyses and differences between their subsample model counterparts are qualitatively similar to our main findings and continue to offer strong statistical support for our predictions.

Insert Tables 4 and 5 Here

Post-hoc Analysis: Exploring ways to improve industry foresight

We have theorized and shown that the forecastability associated with an uncertain industry outcome depends on both the complexity and the structure of the forecasting problem such that the lowest forecastability would be for high complexity ill-structured, the highest would be low complexity well-structured, and intermediate forecastability would be for both high complexity well-structured and low complexity ill-structured. Yet, we do not theorize how individuals may improve their forecasting of

different types of forecasting problems. As a post-hoc analysis, we utilize the unique structure of our forecast data to explore heterogeneity among individuals in terms of the forecasting behavior and the forecasting accuracy.

Specifically, we focus on the heterogeneity among forecasters in terms of whether and how they update their beliefs (forecasts) for a given question. We found evidence that belief updating, in general, is associated with superior forecasting accuracy. Updating of beliefs is a cognitively demanding task as individuals need to attend to new information, interpret it, and incorporate it into an updated belief. The updated belief that is premised on new information is likely to be more reflective of the industry's evolutionary trajectory. Additionally, how individuals update their beliefs may also affect their forecast accuracy. For example, individuals may follow the representativeness heuristic in which they consider new evidence to be representative of the new state of the world regardless of prior states of the world (Kahneman & Tversky, 1973). Accordingly, the updated belief may correspond to an "overreaction" to new information. On the other hand, individuals may follow the anchoring heuristic in which they believe the prior information is representative of the real world regardless of the new evidence (Kahneman et al., 1982). In this case, the updated belief may correspond to an "underreaction" to new information. Belief updating guided by Bayes rule (i.e., Bayesian belief updating) can mitigate issues of under- or overreaction in forecasting (Kahneman & Lovallo, 1993; Kapoor & Wilde, 2022). This is because such a mental process strikes a balance between considering information underpinning prior forecasts and new information when updating one's forecasts (DeGroot, 1970).

We evaluate the effectiveness of whether and how an individual updates their beliefs for the different types of forecasting problems. For the subset of observations that represent updated beliefs on a given question, we draw on recent research by Augenblick & Rabin (2021) to identify characteristics that are aligned with a Bayesian belief updating process and those which diverge from such an updating process (see Appendix 3 for details). We utilize this approach to operationalize forecasting behavior through three dichotomous variables (1/0)—*Bayesian belief updated* equaling one if the updated forecast corresponded to a Bayesian process, *non-Bayesian belief updated* equaling one if the updated forecast did

not correspond to a Bayesian process, and *First forecast* (with or without subsequent updating) as the omitted category.

Insert Table 6 Here

The results are reported in Table 6. Model 7 replicates our main model and explores whether and how an individual updates their beliefs impacts the forecast accuracy. Model 8 includes interaction terms between problem types and belief updating. As compared to forecasts that do not capture a belief updating process, those that capture a Bayesian belief updating process are associated with higher forecast accuracy, and those that capture non-Bayesian belief updating are statistically indistinguishable in terms of forecast accuracy. This finding points to the overall effectiveness of Bayesian belief updating process with respect to forecast accuracy. In exploring the interactions between belief updating and different types of forecasting problems, the non-Bayesian belief updating seems to be particularly detrimental to high-complexity ill-structured forecasting problem, whereas the Bayesian belief updating does not seem to have any significant interaction with the different types of forecasting problems. These findings suggest that while different types of forecasting problems within an industry context are subject to different degrees of forecastability as we had predicted, a Bayesian belief updating process can help improve forecast accuracy for all types of forecasting problems. Moreover, belief updating that does not conform to a Bayesian process can actually be detrimental for ill-structured and high complexity forecasting problem.

DISCUSSION

Managers and entrepreneurs engage in forecasting the industry context as a means of anticipating and acting on opportunities and threats within an evolving industry (Porter, 1980; Barney, 1986; Eckhardt & Shane, 2003; Teece, 2007; Csaszar & Laureiro-Martínez, 2018). In this study, we propose that forecasting can be viewed as a problem-solving process in which managers and entrepreneurs search for relevant information and develop a conjecture about future industry outcomes.

We develop a general framework around the forecastability or general difficulty of forecasting problems contingent upon the joint complexity and structure of the problem (Simon, 1973; Fernandes & Simon, 1999; Jonassen, 2004). We consider that forecasting problems of high complexity imposes significant cognitive demands in terms of attention and interpretation in order to manage the multiple interconnected variables underlying the problem solution (Simon, 1962; Funke, 1991; e.g., Felin & Zenger, 2014). This cognitive strain can inhibit an individual's ability to accurately solve the forecasting problem. We further consider that ill-structured problems require individuals to rely heavily on their perceptions to make sense of the inherent ambiguity associated with the problem (Simon, 1973; Fernandes & Simon, 1999; Laureiro-Martínez & Brusoni, 2018). Perceptions are prone to cognitive bias, which should reduce the effectiveness with which an individual solves the forecasting problem (Kahneman & Lovallo, 1993; Kapoor & Wilde, 2022). Collectively, we argue that the forecastability of a forecasting problem should depend on both problem complexity and problem structure, the lowest for high complexity ill-structured problems, the highest for low complexity well-structured problems, and intermediate for low-complexity ill-structured and high complexity well-structured problems.

We explore these arguments by employing a novel experimental design of forecasting tournaments, focusing on the evolution of the automotive industry during 2017-2019 shaped by the emergence of electric and autonomous vehicles. Drawing on a dataset of 14,779 forecasts made by 1,395 individuals who participate in a leading global forecasting platform, we find strong and robust support for our arguments.

These findings extend Simon's original conceptualization of problem-solving (Simon, 1973; Fernandes & Simon, 1999) by showcasing how forecasting can be viewed as a problem solving process in which managers and entrepreneurs engage in cognitive search for relevant information to develop conjectures about future industry outcomes. They highlight how forecasting problems may impose varying degrees of cognitive demands based on the complexity and structure of the problem. Thus, the paper offers a framework for evaluating the general difficulty of various forecasting problems and how

managers and entrepreneurs might improve their effectiveness in forecasting key industry outcomes depending on the characteristics of the problem.

Further, in a post-hoc analysis, we explore different forecasting behaviors that individuals may employ in order to improve their effectiveness in forecasting problems within an industry context. We find evidence to suggest that a Bayesian belief updating process can help improve forecast accuracy for all types of forecasting problems. Further, we observe that belief updating that does not confirm to a Bayesian process is detrimental for high complexity ill-structured forecasting problems.

In so doing, the study contributes to the emerging theory-based perspective centered on how entrepreneurs and managers evaluate the industry context as a basis for strategic decision-making (Felin & Zenger, 2017; Camuffo et al., 2020; Coali et al., 2022; Ehrig & Schmidt, 2022). This literature stream has highlighted the benefits of making theory-driven conjectures in developing strategies and the importance of learning and experimentation in order to adapt strategies over time. We identify how the nature of the strategy problem itself may shape the relative effectiveness of theory-based conjectures and the importance of learning and experimentation. For example, theory-based conjectures are likely to be less accurate for high complexity well-structured problems than for low complexity well-structured problems. Further, a learning-based process that aligns with a Bayesian process is likely to improve forecast accuracy regardless of the question type. These insights may help explain recent evidence that entrepreneurs of early-stage ventures who engage in theory-driven decision making consistent with Bayesian belief updating make superior decisions and reach superior revenue performance (Messinese, 2022).

The study also sheds important light on the nascent literature stream on entrepreneurial and managerial foresight (Gavetti & Menon, 2016; Csaszar & Laureiro-Martínez, 2018; Peterson & Wu, 2021; Kapoor & Wilde, 2022). It complements the extant literature that has largely focused on the subjective representations of uncertain outcomes by highlighting the importance of considering intrinsic characteristics of forecasting problems when evaluating the foresight of individuals or firms. Additionally, by invoking a problem-solving sensibility into the discourse, these findings help illuminate

specific cognitive challenges associated with particular types of forecasting problems faced by managers and entrepreneurs.

Lastly, the study offers several methodological contributions. First, it utilizes a novel experimental design of forecasting tournaments to identify the forecastability of key industry outcomes faced my managers and entrepreneurs. Second, it offers a template for assessing the effectiveness of various forecasting behaviors. Finally, it provides a methodological guide for ascertaining the underlying complexity and structure of forecasting problems in a realistic managerial context.

The study has a number of limitations. First, the findings are specific to shifts taking place in the automobile industry, and exploring other industry environments would help to evaluate the generalizability of our problem-solving framework. Second, our measurement of problem complexity and problem structure relies on voluntary comments and judgements by individuals, operationalizations that may not fully reflect the inherent nature of the problem. We hope that our approach can help guide future research on measuring this important source of problem heterogeneity for managers and entrepreneurs in a given business context. Despite these and other limitations, we hope that the study has contributed to our understanding of the managerial and entrepreneurial problem-solving processes associated with forecasting and shed light on when deliberate predictive strategies may be more or less effective.

REFERENCES

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P. E., Ungar, L., & Mellers, B. (2017). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Management Science*, *63*(3), 691–706.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. E. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Augenblick, N., & Rabin, M. (2021). Belief Movement, Uncertainty Reduction, and Rational Updating. *The Quarterly Journal of Economics*, *136*(2), 933–985. https://doi.org/10.1093/qje/qjaa043

Barney, J. B. (1986). Strategic factor markets: Expectations, luck, and business strategy. *Management Science*, *32*(10), 1231–1241.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Camuffo, A., Cordova, A., Gambardella, A., & Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science*, *66*(2), 564–586. https://doi.org/10.1287/mnsc.2018.3249

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509–526.

Coali, A., Gambardella, A., & Novelli, E. (2022). Understanding Probabilistic Reasoning in Innovation. *Academy of Management Proceedings*.

Conway, L. G., Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated Integrative Complexity. *Political Psychology*, *35*(5), 603–624. https://doi.org/10.1111/pops.12021

Csaszar, F. A., & Laureiro-Martínez, D. (2018). Individual and Organizational Antecedents of Strategic Foresight: A Representational Approach. *Strategy Science*, *3*(3), 513–532.

Csaszar, F. A., & Levinthal, D. A. (2016). Mental representation and the discovery of new strategies. *Strategic Management Journal*, *37*(10), 2031–2049. https://doi.org/10.1002/smj.2440

Denrell, J., Fang, C., & Winter, S. G. (2003). The economics of strategic opportunity. *Strategic Management Journal*, *24*(10), 977–990.

Eckhardt, J. T., & Shane, S. A. (2003). Opportunities and entrepreneurship. *Journal of Management*, *29*(3), 333–349.

Ehrig, T., & Schmidt, J. (2022). Theory-based learning and experimentation: How strategists can systematically generate knowledge at the edge between the known and the unknown. *Strategic Management Journal*, *43*(7), 1287–1318. https://doi.org/10.1002/smj.3381

Felin, T., & Zenger, T. R. (2014). Closed or open innovation? Problem solving and the governance choice. *Research Policy*, *43*(5), 914–925.

Felin, T., & Zenger, T. R. (2017). The theory-based view: Economic actors as theorists. *Strategy Science*, *2*(4), 258–271.

Fernandes, R., & Simon, H. A. (1999). A study of how individuals solve complex and ill-structured problems. *Policy Sciences*, *32*(3), 225–245.

Funke, J. (1991). Solving Complex Problems : Exploration and Control of Complex Systems. *Solving Complex Problems:Exploration and Control of Complex Systems*, 223.

Gary, M. S., & Wood, R. E. (2011). Mental models, decision rules, and performance heterogeneity. *Strategic Management Journal*, *32*(6), 569–594. https://doi.org/10.1002/smj.899

Gavetti, G., & Lecuona Torras, J. R. (2021). A Neo-Carnegie Approach to the Agency Question: Bridging the Evolutionary and Cognitive Views of Strategy. *Strategy Science*, *December*. https://doi.org/10.1287/stsc.2021.0149

Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly*, *45*(1), 113–137.

Gavetti, G., & Menon, A. (2016). Evolution Cum Agency: Toward a Model of Strategic Foresight.

Strategy Science, *1*(3), 207–233. https://doi.org/10.1287/stsc.2016.0018
Graf-Vlachy, L., Bundy, J., & Hambrick, D. C. (2020). Effects of an advancing tenure on CEO cognitive complexity. *Organization Science*, *31*(4), 936–959.
Helfat, C. E., & Peteraf, M. A. (2015). Managerial cognitive capabilities and the microfoundations of dynamic capabilities. *Strategic Management Journal*, *36*(6), 831–850.
Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. MIT press.
Houck, S. C., Conway, L. G., & Gornick, L. J. (2014). Automated Integrative Complexity: Current Challenges and Future Directions. *Political Psychology*, *35*(5), 647–659. https://doi.org/10.1111/pops.12209
Hsieh, C., Nickerson, J. A., & Zenger, T. R. (2007). Opportunity discovery, problem solving and a theory of the entrepreneurial firm. *Journal of Management Studies*, *44*(7), 1255–1277. https://doi.org/10.1111/j.1467-6486.2007.00725.x
Jonassen, D. H. (2004). *Learning to solve problems: An instructional design guide* (Vol. 6). John Wiley & Sons.
Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science*, *55*(4), 582–590.
Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*(1), 17–31.
Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown.
Kapoor, R., & Wilde, D. (2022). Peering into a Crystal Ball: Forecasting Behavior and Industry Foresight. *Strategic Management Journal, forthcoming*. https://doi.org/10.1002/smj.3450
Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2022). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting*, *38*(2), 688–704. https://doi.org/10.1016/j.ijforecast.2021.09.003
Klein, G. A. (1998). *Sources of power: How people make decisions*. MIT Press.
Koopmans, T. C. (1957). *Three essays on the state of economic science*.
Laureiro-Martínez, D., & Brusoni, S. (2018). Cognitive flexibility and adaptive decision-making: Evidence from a laboratory study of expert decision makers. *Strategic Management Journal*, *39*(4), 1031–1058.
Leiblein, M. J., & Macher, J. T. (2009). The problem solving perspective: A strategic approach to understanding environment and organization. In *Economic institutions of strategy*. Emerald Group Publishing Limited.
Macher, J. T. (2006). Technological development and the boundaries of the firm: A knowledge-based examination in semiconductor manufacturing. *Management Science*, *52*(6), 826–843. https://doi.org/10.1287/mnsc.1060.0511
Malhotra, S., & Harrison, J. S. (2022). *A blessing and a curse : How chief executive officer cognitive complexity influences firm performance under varying industry conditions*. *April*, 1–20. https://doi.org/10.1002/smj.3415
Martignoni, D., Menon, A., & Siggelkow, N. (2016). Consequences of misspecified mental models: Contrasting effects and the role of cognitive fit. *Strategic Management Journal*, *37*(13), 2545–2568. https://doi.org/10.1002/smj.2479
McHugh, M. L. (2012). Lessons in biostatistics interrater reliability : the kappa statistic. *Biochemica Medica*, *22*(3), 276–282. https://hrcak.srce.hr/89395
Mcnamara, G. M., Luce, R. A., & Tompson, G. H. (2002). Examining the effect of complexity in strategic group knowledge structures on firm performance. *Strategic Management Journal*, *23*(2), 153–170. https://doi.org/10.1002/smj.211
Messinese, D. (2022). *Information acquisition with predictive and non-predictive strategies*.
Moldoveanu, M. (2009). Thinking strategically about thinking strategically: the computational structure and dynamics of managerial problem selection and formulation. *Strategic Management Journal*, *30*(7), 737–763. https://doi.org/10.1002/smj.757

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P. E., Yang, H. H. J., & Tenney, E. R. (2017). Confidence Calibration in a Multiyear Geopolitical Forecasting Competition. *Management Science*, *63*(11), 3552–3565.

Nickerson, J. A., & Zenger, T. R. (2004). A Knowledge-Based Theory of the Firm—The Problem-Solving Perspective. *Organization Science*, *15*(6), 617–632. https://doi.org/10.1287/orsc.1040.0093

Packard, M., Clark, B., & Klein, P. (2017). Uncertainty Types and Transitions in the Entrepreneurial Process. *Organization Science*, *28*(5), 840–856.

Peterson, A., & Wu, A. (2021). Entrepreneurial learning and strategic foresight. *Strategic Management Journal*, *July 2020*, 1–32. https://doi.org/10.1002/smj.3327

Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press.

Santos, F., & Eisenhardt, K. (2009). Constructing Markets and Shaping Boundaries: Entrepreneurial Power in Nascent Fields. *The Academy of Management Journal*, *52*(4), 643–671.

Silver, N. (2012). *The signal and the noise: why so many predictions fail--but some don't*. Penguin.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*, 467–482.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, *4*(3–4), 181–201.

Teece, D. J. (2007). Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*(13), 1319–1350.

Tetlock, P. E. (2017). *Full-Inference-Cycle Tournaments: The Quality of Our Questions Matters As Much As the Accuracy of Our Answers*.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, *23*(4), 290–295.

Tetlock, P. E., Metz, S. E., Scott, S. E., & Suedfeld, P. (2014). Integrative Complexity Coding Raises Integratively Complex Issues. *Political Psychology*, *35*(5), 625–634. https://doi.org/10.1111/pops.12207

Venkataraman, S. (1997). The distinctive domain of entrepreneurship research: An editor's perspective. J. Katz, R. Brockhaus, eds. Advances in Entrepreneurship, Firm Emergence and Growth, Vol. 3. *Greenwich, CT: JAI Press Vickers, New York: Basic Books*, *3*, 119–138.

Walsh, J. P. (1995). Managerial and organizational cognition: Notes from a trip down memory lane. *Organization Science*, *6*(3), 280–321.

Figure 1. A framework for analyzing industry forecasting problems

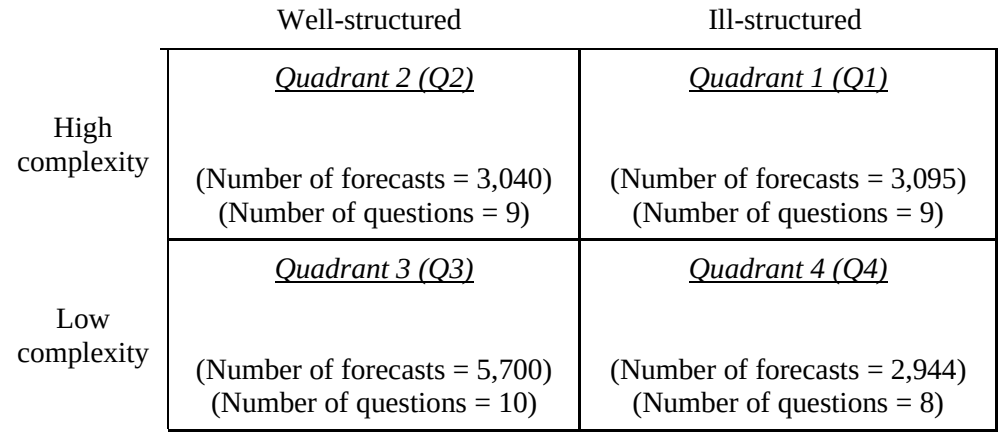

Table 1. Descriptive statistics and bivariate correlation matrix

	Variables	Mean	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Forecast accuracy	0.44	0.54	0.00	2.00	1.00												
2	High complexity ill-structured	0.21	0.41	0.00	1.00	0.08	1.00											
3	Low complexity ill-structured	0.20	0.40	0.00	1.00	0.07	-0.26	1.00										
4	High complexity well-structured	0.21	0.40	0.00	1.00	0.01	-0.26	-0.25	1.00									
5	Belief updated	0.52	0.50	0.00	1.00	-0.21	-0.06	-0.14	0.11	1.00								
6	Uncertainty	0.35	0.39	0.00	2.00	0.53	-0.03	0.08	0.09	-0.11	1.00							
7	Days active	3.33	2.02	0.00	6.40	-0.20	-0.06	-0.12	0.13	0.70	-0.07	1.00						
8	Questions platform	3.71	1.19	0.00	6.09	-0.10	-0.01	-0.05	0.04	0.42	-0.04	0.68	1.00					
9	Questions domain	3.96	2.77	0.00	10.00	-0.03	0.01	-0.01	0.08	0.26	0.03	0.40	0.67	1.00				
10	Forecaster count	10.99	9.64	1.00	76.00	0.12	0.00	0.07	-0.15	-0.28	0.07	-0.33	-0.21	-0.20	1.00			
11	Commented	0.27	0.44	0.00	1.00	-0.01	0.00	-0.01	0.03	0.07	0.00	0.12	-0.03	-0.02	0.02	1.00		
12	Multiple tournaments	0.33	0.47	0.00	1.00	-0.05	0.02	-0.04	0.08	0.29	0.01	0.41	0.35	0.26	-0.17	0.04	1.00	
13	Days-to-end	154.00	101.90	8.00	411	0.169	-0.01	-0.05	0.02	-0.24	0.18	-0.10	-0.07	-0.03	0.10	0.03	0.13	1.00

Values above 0.01 and below -0.01 indicate significant at $p<0.01$. N = 14,779

Table 2. Forecast level coefficient estimates from fixed effects OLS regression
(Note: Lower values of the dependent variable imply higher accuracy)

Dependent variable: Forecast accuracy	(1)	(2)
High complexity Ill-structured (Q1)		0.128***
		(0.012)
Low complexity Ill-structured (Q4)		0.074***
		(0.012)
High complexity well-structured (Q2)		0.053***
		(0.009)
Belief updated	-0.031**	-0.020+
	(0.011)	(0.011)
Uncertainty	0.682***	0.676***
	(0.013)	(0.013)
Days active	-0.044**	-0.041*
	(0.016)	(0.016)
Questions platform	0.042*	0.056**
	(0.018)	(0.018)
Questions domain	0.010**	0.012**
	(0.004)	(0.004)
Forecaster count	0.002**	0.002**
	(0.001)	(0.001)
Commented	0.027*	0.023*
	(0.011)	(0.011)
Multiple tournaments	0.029+	-0.012
	(0.017)	(0.017)
Days-to-end	0.000***	0.000***
	(0.000)	(0.000)
Observations	14,779	14,779
R-squared	0.448	0.454
Individual FE	YES	YES

Robust standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10
Omitted category: Low complexity well-structured (Q3)

Table 3. Difference between coefficient estimates using Wald tests

Null hypothesis	F stat	Prob > F
Q1 = Q2	33.89	<0.000
Q1 = Q4	13.38	0.0003
Q2 = Q4	2.84	0.0921

Table 4. Summary of robustness checks

Model	Robustness check	Rationale
3	Control for multinomial-ordered questions	Given the high correlation between well-structured questions and multinomial-ordered questions, the results may be driven by systematic differences in how the brier score is calculated for those questions
4, 5	Applied five-day and seven-day windows surrounding the forecast to variable calculations	There may be idiosyncratic effects during a three-day measurement period explaining results
6	Alternative classification of high/low complexity questions based on the percentage of total comments in a question with significantly high complexity (i.e., complexity score of 4-7 out of 7)	Average complexity of comments may not reflect the complexity of a question as much as considering only highly complex comments

Table 5. Robustness checks for full sample

(Note: Lower values of the dependent variable imply higher accuracy)

Dependent variable: Forecast accuracy	(3) Control for Mult. ordered questions	(4) Window = 5 days	(5) Window = 7 days	(6) Complexity based on % of high complex comments only
High complexity ill-structured (Q1)	0.150*** (0.039)	0.127*** (0.012)	0.127*** (0.012)	0.126*** (0.012)
Low complexity ill-structured (Q4)	0.098* (0.043)	0.073*** (0.012)	0.073*** (0.012)	0.066*** (0.011)
High complexity well-structured (Q2)	0.054*** (0.009)	0.049*** (0.009)	0.049*** (0.009)	0.041*** (0.009)
Belief updated	-0.020+ (0.011)	-0.021+ (0.011)	-0.021+ (0.011)	-0.018 (0.011)
Uncertainty	0.674*** (0.014)	0.684*** (0.013)	0.684*** (0.013)	0.680*** (0.013)
Days active	-0.040* (0.016)	-0.039* (0.016)	-0.039* (0.016)	-0.040* (0.016)
Questions platform	0.057** (0.018)	0.055** (0.018)	0.055** (0.018)	0.044* (0.018)
Questions domain	0.012** (0.004)	0.012** (0.004)	0.012** (0.004)	0.010** (0.004)
Forecaster count	0.002** (0.001)	0.001 (0.000)	0.000 (0.000)	0.002** (0.001)
Commented	0.023* (0.011)	0.024* (0.011)	0.024* (0.011)	0.025* (0.011)
Multiple tournaments	-0.014 (0.017)			
Days-to-end	0.000*** (0.000)			
Multinomial ordered question	0.023 (0.041)			
Observations	14,779	14,784	14,784	14,779
R-squared	0.454	0.457	0.457	0.454
Individual FE	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.001, ** p<0.01, * p<0.05, + p<0.10,
Omitted category: Low complexity well-structured (Q3)

Table 6. Effects of belief updating on forecast accuracy by quadrant
(Note: Lower values of the dependent variable imply higher accuracy)

Dependent variable: Forecast accuracy	(7)	(8)
High complexity ill-structured (Q1) X Bayesian belief updated		-0.012
		(0.026)
High complexity ill-structured (Q1) X non-Bayesian belief updated		0.098***
		(0.028)
Low complexity ill-structured (Q4) X Bayesian belief updated		-0.003
		(0.024)
Low complexity ill-structured (Q4) X non-Bayesian belief updated		0.046
		(0.036)
High complexity well-structured (Q2) X Bayesian belief updated		0.009
		(0.022)
High complexity well-structured (Q2) X non-Bayesian belief updated		-0.036+
		(0.021)
Bayesian belief updated	-0.048***	-0.045**
	(0.014)	(0.016)
Non-Bayesian belief updated	0.005	-0.010
	(0.014)	(0.016)
High complexity ill-structured (Q1)	0.129***	0.107***
	(0.012)	(0.019)
Low complexity ill-structured (Q4)	0.083***	0.076***
	(0.012)	(0.018)
High complexity well-structured (Q2)	0.049***	0.067***
	(0.009)	(0.018)
Uncertainty	0.672***	0.669***
	(0.013)	(0.013)
Days active	-0.046**	-0.045**
	(0.016)	(0.016)
Questions platform	0.058**	0.060**
	(0.018)	(0.019)
Questions domain	0.012**	0.012**
	(0.004)	(0.004)
Forecaster count	0.002**	0.002**
	(0.001)	(0.001)
Commented	0.024*	0.023*
	(0.011)	(0.011)
Multiple tournaments	-0.017	-0.019
	(0.017)	(0.017)
Days-to-end	0.000***	0.000***
	(0.000)	(0.000)
Observations	14,779	14,779
R-squared	0.455	0.457
Individual FE	YES	YES

Robust standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10
Omitted categories: Low complexity well-structured (Q3), First forecast

Appendix 1: Forecast Tournament Questions

Tournament	Quadrant	Question	Launch Date	Closed Date	Resolution	Forecasters	Forecasts
2018-19	4	Before 1 January 2020, will the U.S. President sign legislation increasing the number of exemptions for autonomous vehicles allowed per manufacturer by the Federal Motor Vehicle Safety Standards?	11/16/2018	1/1/2020	No	214	475
2018-19	2	Between 1 October 2018 and 31 December 2019, how many Model 3 cars will Tesla deliver to customers?	11/16/2018	1/1/2020	Between 330,000 and 380,000, inclusive	233	917
2018-19	3	What will be the 2019 industry-wide average cost of Li-ion batteries used in battery-powered electric vehicles?	11/16/2018	1/1/2020	More than $155 but less than $170 per kWh	191	492
2018-19	2	Between 1 January 2019 and 31 December 2019, how many reports of traffic accidents involving an autonomous vehicle will the California Department of Motor Vehicles receive?	11/16/2018	1/1/2020	Between 95 and 115, inclusive	221	824
2018-19	1	Before 1 January 2020, will General Motors launch a ride-hailing service open to the public in the U.S. which uses autonomous vehicles?	12/14/2018	1/1/2020	No	289	553
2018-19	3	On 10 December 2019, how many total locations with Combined Charging System (CCS) fast chargers will be installed in the European area?	12/14/2018	12/10/2019	Between 8,000 and 8,999, inclusive	172	468
2018-19	4	Before 1 January 2020, will Tesla release an Autopilot feature designed to navigate traffic lights?	1/18/2019	1/1/2020	No	279	588
2018-19	4	Will legislation eliminating the unit limit per manufacturer for the U.S. federal electric vehicle tax credit become law before 1 January 2020?	1/18/2019	1/1/2020	No	148	348
2018-19	2	What will annual sales of new energy vehicles (NEVs) be in China in 2019?	1/18/2019	1/1/2020	Less than 1.25 million	135	492
2018-19	4	Before 1 January 2020, will the registration deadline for Germany's ownership tax exemption for fully-electric vehicles be extended beyond 2020?	2/15/2019	1/1/2020	No	159	364
2018-19	1	Before 1 July 2019, will AT&T, Sprint, T-Mobile, or Verizon offer 5G smartphones to US customers?	2/15/2019	7/1/2019	Yes, all 4 of the companies	159	397
2018-19	1	Before 1 January 2020, will EV Rater list a full-electric vehicle with a range of 420 miles or more?	3/8/2019	1/1/2020	Yes	254	541
2018-19	3	Between 8 March and 31 December 2019, how many accidents involving a self-driving vehicle operating in autonomous mode in the U.S. will result in a fatality?	3/8/2019	12/31/2019	0 accidents	325	666
2018-19	1	Before 1 January 2020, will Velodyne announce the release of a LiDAR unit with a maximum range of 400 meters or more?	3/8/2019	1/1/2020	No	124	324
2018-19	3	What will be the price of regular gasoline in the U.S. per gallon on 30 December 2019?	4/26/2019	12/31/2019	Between $2.40 and $2.650, inclusive	157	592
2018-19	1	Before 1 January 2020, will a firm or paid backup driver operating a self-driving vehicle face criminal charges in relation to an accident involving a self-driving vehicle in the U.S.?	4/26/2019	1/1/2020	No	220	444
2018-19	2	As of 31 December 2019, how many current and on target for production future full-electric vehicle models will EV Rater list?	6/7/2019	12/30/2019	More than 38 but less than 45	59	198
2018-19	1	Before 1 January 2020, will the Federal Communications Commission (FCC) vote in favor of granting a waiver petition [GN Docket No. 18-357] to allow for the further deployment of Cellular Vehicle-to-Everything (C-V2X) technology?	6/21/2019	12/31/2019	No	59	156
2018-19	3	How many Plug-in Hybrid Electric Vehicles (PHEVs) will be registered in the UK in 2019, according to the Society of Motor Manufacturers and Traders (SMMT)?	9/20/2019	1/1/2020	More than 27,000	52	149
2018-19	1	Before 31 December 2019, will a Model 3 produced in Tesla's Shanghai Gigafactory 3 be delivered to a customer?	9/20/2019	12/31/2019	Yes	105	367

Appendix 1: Forecast Tournament Questions (continued)

Tournament	Quadrant	Question	Launch Date	Closed Date	Resolution	Forecasters	Forecasts
2017-18	NA[15]	Between 21 July 2017 and 20 July 2018, a date after which they will sell only electric or hybrid vehicles?	7/21/2017	7/20/2018	No	593	1243
2017-18	3	Before 1 July 2018, how many Model 3 cars will Tesla deliver to customers?	7/21/2017	7/1/2018	Less than 50,000	518	1312
2017-18	1	Before 1 July 2018, will Uber, or any of its subsidiaries, agree to a settlement or be found liable for trade secrets violations in the case brought by Waymo in the Northern District of California?	7/21/2017	7/1/2018	Yes	309	633
2017-18	4	Before 1 January 2018, will the U.S. President sign legislation increasing the number of autonomous vehicle exemptions allowed by the Federal Motor Vehicle Safety Standards?	7/21/2017	1/1/2018	No	226	421
2017-18	2	What will be the 2017 industry-wide average cost of Li-ion batteries used in battery-powered electric vehicles?	7/21/2017	1/1/2018	>$230, but < $245 kWh	178	404
2017-18	3	How many Chevrolet Bolt EV's will be sold between January and June 2018?	8/18/2017	7/1/2018	Less than 10,000	355	909
2017-18	3	On 29 June 2018, how many public DC Fast Charge electric vehicle charging stations will be available in the United States?	9/8/2017	6/29/2018	Less than 2,400	332	850
2017-18	3	On 30 March 2018, how many GitHub forks will Baidu's Apollo autonomous driving software have?	9/8/2017	3/30/2018	Between 2,001 and 3,000, inclusive	241	849
2017-18	1	Before 1 July 2018, will Tesla announce that it will build a factory to manufacture electric vehicles in China?	10/20/2017	7/1/2018	No	520	947
2017-18	4	Before 1 July 2018, will Waymo launch a driverless transportation service open to the public?	11/17/2017	7/1/2018	Yes	370	547
2017-18	4	Before 1 April 2018, will General Motors test an autonomous vehicle in New York City?	11/17/2017	4/1/2018	No	409	730
2017-18	3	Between 1 January 2018 and 30 June 2018, how many reports of traffic accidents involving an autonomous vehicle will the California Department of Motor Vehicles receive?	12/8/2017	7/1/2018	Between 20 and 29, inclusive	496	1072
2017-18	4	Before 20 July 2018, will Audi sell or lease a motor vehicle with Traffic Jam Pilot?	1/24/2018	7/20/2018	No	312	506
2017-18	2	On 29 June 2018, how many public hydrogen fueling stations will be available in the United States?	2/28/2018	6/29/2018	Fewer than 45	195	435
2017-18	2	How many Mirais will Toyota sell or lease between January 2018 and June 2018, inclusive?	2/28/2018	7/1/2018	Less than 900	100	265
2017-18	2	As of 1 July 2018, how many manufacturers will hold permits for driverless testing of autonomous vehicles in California?	4/18/2018	7/1/2018	Zero	154	312
2017-18	2	Between 1 July 2017 and 1 July 2018, how many Model 3 cars will Tesla deliver to customers?	4/18/2018	7/1/2018	29,000 or less	107	391

[15] In the middle of the question window for this question, Good Judgment Inc made an important clarification in what the question was asking and how the question would be resolved. Accordingly, forecasts made before this clarification are unreliable for inference and we removed the question from analysis.

Appendix 2: Calculation details for forecast accuracy

The *forecast accuracy* measure is the forecasting error, which is based on quadratic scoring rules (QSR) that measure the squared error between the predicted probabilities and the ultimate resolution of a question. For *unordered* questions including binary and unordered multinomial questions, the QSR is the sum of squared errors commonly known as the probability (or Brier) score (Brier, 1950). Formally, the daily forecasting error for unordered questions is calculated as follows:

$$error_{iqt} = \sum_{i=1}^{r} (f_{ijqt} - o_{jq})^2$$

where f_{ijqt} is forecaster *i*'s probability forecast for question option *j* for question *q* at time *t*, *r* is the number of possible answer options in which the event can fall (e.g., "Yes – with condition A", "Yes – with condition B", "No" would have $r = 3$), o_{jq} the actual outcome of the answer option *j* of question *q* (equals 0 if it does not happen and 1 if it does happen). Errors range from 0 to 2, with lower errors indicating higher accuracy.

To illustrate, consider a forecaster submits a forecast for the binomial question "Will annual sales of electric vehicles in China reach 500,000 in 2016?" with 80% for A = "Yes" and 20% for B = "No". If the resolution of the question was A = "Yes", then we would set the resolution for A as 1, and the resolution for B (and all other answer options if the question was unordered multinomial) as 0. For each answer option, we then calculate the difference between the forecast and the respective resolutions, square the differences, and sum these together, thus arriving at $(1 - 0.8)^2 + (0 - 0.2)^2 = 0.08$. Given the forecast for the answer option that occurred was high at 80%, the forecasting error is quite low. For an unordered multinomial question with answer options A, B, and C, in which a forecaster's forecast were A = 60%, B = 10%, C = 30% and option A occurred, the forecasting error would be $(1 - 0.6)^2 + (0 - 0.1)^2 + (0 - 0.3)^2 = 0.26$.

For *ordered* questions, we follow (Jose et al., 2009) to apply a "sensitive-to-distance" QSR that assigns partial credit for near-misses (i.e., incorrect but close answers to the resolved answers). For the 2016 tournament there were no ordered questions, so this second QSR is relevant for the 2017 - 2018 and 2018 - 2019 tournaments. We calculate the error for ordered questions in the following steps. First, we generate a set of pairs of answer options by systematically splitting the *r* answer options into two groups where the threshold between the two groups shifts upward from between A and B to between B and C and so forth. Ultimately, this will generate *r* - 1 pairs such as A vs. BCD, AB vs. CD, and ABC vs. D for an ordered question with $r = 4$ answer options. Second, with these pairs in mind, we sum the forecaster's probabilities associated with the answer options in each group. We also set the question group containing the resolved answer option to 1 and set the other group in the pair to 0. Third, we calculate the sum of squared errors for each of these pairs by squaring the difference between each group's summed probabilities and the outcome (i.e., 1 or 0). Finally, we take the mean across these errors. Formally, the $error_{iqt}$ for ordered questions is calculated as follows:

$$error_{iqt} = \frac{1}{r-1}\left[\sum_{c=1}^{r-1}\left[\left(\sum_{j=1}^{c} f_{ijqt} - \sum_{j=1}^{c} o_{jq}\right)^2 + \left(\sum_{j=c+1}^{r} f_{ijqt} - \sum_{j=c+1}^{r} o_{jq}\right)^2\right]\right]$$

where f_{ijqt} is forecaster *i*'s probability forecast for answer options *j* for question *q* at time *t*, o_{jq} the actual outcome of the question *q* of answer option *j* at time *t* (i.e., equals 0 if the answer option does not occur and 1 if it does occur), *r* is the total number of answer options, and *c* is the ordered index number associated with an answer option (e.g., A = 1, B = 2, C = 3). Just as with unordered scores, ordered scores can range from 0 to 2 with lower scores indicating higher accuracy. However, for ordered questions in which the resolution is not on one of the polar end choices (e.g., not answer option A or the highest letter

option in the question such as C where $r = 3$), the score will range from 0 to less than 2. Nonetheless, due to question fixed effects in our analysis, this potentially lower upper bound is not a concern.

To illustrate this calculation, consider the ordered multinomial question "On 29 June 2018, how many public DC Fast Charge electric vehicle charging stations will be available in the United States?" with options A = "less than 2,400", B = "between 2,400 and 2,700 inclusive", C = "between 2,701 and 3,100 inclusive", and D = "More than 3,100". Further consider a forecast of A = 25%, B = 25%, C = 50%, D = 0% and that option B occurred. We would first divide the four answer options of A-B-C-D into three binary pairs as follows: A versus BCD, AB versus CD, and ABC versus D. In the A vs BCD pair, the sum of forecasts for A is 0.25 and the sum of forecasts for BCD is 0.75. Also, because answer option B occurred, the outcome for BCD is 1 and the outcome for A is 0. We then arrive at the following score for the A vs. BCD binary pair: $(0.25 – 0)^2 + (0.75 − 1)^2 = 0.125$. We can repeat this process for the other binary pairs: AB vs CD: $(0.5 − 1)^2 + (0.5 − 0)^2 = 0.50$, ABC vs D: $(1 − 1)^2 + (0 − 0)^2 = 0$. Finally, we would average these errors to arrive at 0.208. If we were to employ the simpler forecasting rule used for unordered questions, this error would have been much worse at 0.875, not giving partial credit for predicting "near misses" of the A and C answer options that were near the resolution of B.

Appendix 3: Calculating Bayesian belief updating

We draw on Augenblick & Rabin (2021) to consider that forecasting behavior consistent with Bayesian belief updating can be identified by considering all forecasts made by the individual for a given question. Specifically, Bayesian belief updating can be identified by observing an individual's initial forecast and the magnitude of subsequent changes in forecast over time. A Bayesian belief updating process is such that the weaker an initial forecast (i.e., closer to 50%), the larger the subsequent updates should be. This approach is premised on individuals developing stronger beliefs (i.e., closer to 0% or 100%) over time because relevant new information should make them more informed. Through this method, we can also uncover forecasting behavior that deviates from a Bayesian process. For instance, making a strong initial forecast followed by large updates is associated with overreaction and making a weak initial forecast followed by small updates is associated with underreaction (Augenblick & Rabin, 2021). We follow Augenblick & Rabin (2021) to operationalize the relative shift in forecast according to the following calculation:

$$= \sum_{t=1}^{T} \sum_{j=1}^{r} (f_{iqjt} - f_{iqjt-1})^2 - \sum_{j=1}^{r} f_{iqj0}(1 - f_{iqj0})$$

Where f_{ijqt} is the forecaster i's probability forecast for answer option j for question q at time t (i.e., each day beginning with the forecaster's initial forecast of the question and over all days T until the question closes), f_{ijq0} is the initial forecast of the focal question, and r is the number of possible answer options in which the event can fall (e.g., "Yes – with condition A", "Yes – with condition B", "No" would have r = 3). A value of zero would indicate perfect Bayesian cognitive process, and the larger the deviation from zero, the greater is the extent of under-reaction (negative values) and over-reaction (positive values).

We categorize forecasting behavior as Bayesian belief updating for observations between the 25th and 75th percentile of the distribution for each question, which has a median very close to zero, and 0 for observation that deviate significantly from the median (less than 25th percentile and greater than 75th percentile). This methodology has recently been applied to explore the extent of Bayesian belief updating in forecasting tournaments (Atanasov et al., 2020; Kapoor & Wilde, 2022). Finally, within our main dataset we operationalize each updated forecast as *Bayesian belief updated* equaling one if the forecast is associated with a Bayesian belief updating process at the individual-question level as indicated above, and zero otherwise. Similarly, we operationalize updated forecasts as *non-Bayesian belief updated* if the forecast is not associated with a Bayesian belief updating process, and zero otherwise.

Appendix 4: Question structure categorization survey

To gain further comfort around our categorization of question structure, we recruited 121 individuals from within Prolific, a leading research participant recruiting platform that verifies and monitors participants, to independently categorize questions via survey. These individuals were English-speaking adults based in the United States. The survey first provided information outlining the difference between the two types of structures. Then, the survey provided a single forecast question from our tournament and asked the individual to categorize the question's structure. This question was randomly selected from a set of four questions that represent the quadrants within our framework (see table below for list of questions and associated quadrants).

A total of 109 (92%) responses aligned with our categorizations, with 29 of 33 (88%) agreeing with our categorization of the quadrant 1 question, 25 of 26 (96%) agreeing with the quadrant 2 question, 27 of 29 (93%) agreeing with the quadrant 3 question, and 28 of 31 (90%) agreeing with the quadrant 4 question. The Cohen Kappa comparing our classifications with those of the respondents was 0.91, which is considered almost perfect agreement (McHugh, 2012).

We took several precautions to ensure the quality of these results. First, we considered that the words "ill" or "well" have strong connotations that may bias individuals' responses. We, therefore, used "Type 1" and "Type 2" in place of "well-structured" and "ill-structured," respectfully. Second, we considered that asking respondents to categorize multiple questions could bias their categorizations if they compared questions (e.g., some questions have longer background information than others, that may prime individuals to categorize them as well-structured). Moreover, the cognitive demanding nature of categorizing multiple questions could make a respondent's first categorization higher quality than their subsequent categorizations. Accordingly, we imposed a conservative design of a single categorization task per individual. Third, we considered that individuals may give spurious responses by either using survey bots for their responses or by not paying attention during the survey. Accordingly, we required respondents to both explain the reason behind their categorization and also answer an attention check question that an attentive human would correctly answer. Two individuals did not pass these assessments and were excluded from our sample. Finally, we considered that providing examples of well- and ill-structured problems with different length of background information could bias the individual in their categorizations. Therefore, we provided examples with similar length. Below is the core content of the survey.

Section 1 of 1: Explaining Question Structure

As part of our forecasting research, we asked individuals to make predictions regarding several important industry outcomes (forecast questions) dealing with the global automotive industry.

Each question varied in terms of its structure based on its provided background information, ranging from Type 1 to Type 2. More specifically, **Type 1** questions provided background information that clearly guided the individual with key factors, such as historical trend information, to consider when forecasting. In contrast, **Type 2** questions provided background information that did *not* offer clear guidance as to the key factors or historical trend information to consider when forecasting.

Let's illustrate what we mean with two examples:

Example Type 1 forecast question

FORECAST QUESTION: "How many Mirais will Toyota sell or lease between January 2018 and June 2018, inclusive?"

BACKGROUND INFORMATION**:** The Mirai is Toyota's first commercially launched fuel cell-powered electric vehicle (Toyata, Forbes, Toyota). Since 2015, Toyota has sold or leased over 3,000 Mirais in the US, all in California (The Drive). Toyota Mirai sales data can be tracked here.

This was a Type 1 problem because the background information provided clear relevant information (e.g., the historical data) from which you could make a reasonable forecast.

Example Type 2 forecasting question

FORECAST QUESTION: "A second question was "Before 1 April 2018, will General Motors test an autonomous vehicle in New York City?"

BACKGROUND INFORMATION: Cruise Automation, the self-driving unit of General Motors, recently announced its intention to test autonomous Chevy Bolts in New York City a high regulation, high density location.

This was a Type 2 problem because the background information did not provide clear relevant information (e.g., historical data) from which you could make a reasonable forecast.

How would you describe the structure of the forecast question below, given the provided background information?

[Randomly assigned forecast question. See below for questions]

- Type 1 forecast question
- Type 2 forecast question

Please explain the reason behind your answer above (15 words or less):

_____

The vehicle test you are about to take part in is very simple, when asked for your favorite vehicle you must select "Dodge Viper". This is an attention check.

Based on the text you read above, what is your favorite vehicle?

- Dodge Viper
- Tesla Model 3
- Maserati MC20
- Chevrolet Corvette
- Bugatti Divo

END OF SURVEY

Representative questions from each quadrant used in the categorization exercise survey

Quadrant	Forecast question	Background Information
1	By the end of the year, will a firm or paid backup driver operating a self-driving vehicle face criminal charges in relation to an accident involving a self-driving vehicle in the U.S.?	The legal implications surrounding self-driving vehicles are of special interest to many stakeholders of the auto industry (Reuters, NY Times, The Atlantic, CNN, Popular Science).
2	Will Nissan sell more than 15,000 units of the LEAF in the US in 2016?	Nissan LEAF is the world's all-time best selling battery-powered electric car, with global sales of over 200,000 units since its launch in 2010 (Nissan). More than 30,000 LEAFS were sold in the US during 2014, up from 22,610 in 2013. However, there was a sharp decline in 2015 with total sales in the US amounting to only 17,269. The newly-released 2016 LEAF has an upgraded 30 kWh battery option which will yield an estimated 107-mile range, a 27% improvement over the 24 kWh option offered in the previous model (Car and Driver). During the first quarter of this year, 2,931 units of LEAF were sold in the US.
3	On 10 December 2019, how many total locations with Combined Charging System (CCS) fast chargers will be installed in the European area?	The adoption of electric vehicles is subject to a "chicken and egg" problem where potential consumers want a more extensive charging network, but businesses want more electric cars on the roads to justify building those new charging stations (Reuters, Inside EVs). Europe has gone from zero Combined Charging System (CCS) fast charger locations in 2014 to 5,712 as of 14 December 2018. This question will be resolved using the total installed charger locations listed on 10 December 2019 from the "CCS Charge Map – Europe" website.
4	Before the end of the year, will Tesla release an Autopilot feature designed to navigate traffic lights?	Tesla CEO Elon Musk has expressed interest in adding a feature to navigate city streets, including traffic lights (Ars Technica, The Drive).