Received: 8 October 2022 | Revised: 17 November 2022 | Accepted: 17 November 2022

DOI: 10.1002/arcp.1089

REVIEW

Predicting the future with humans and AI

Barbara A. Mellers[1] | Louise Lu[2] | John P. McCoy[1]

[1]Department of Marketing, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Department of Marketing, Stanford University, Stanford, California, USA

Correspondence
Barbara A. Mellers, Solomon Labs, University of Pennsylvania, 3720 Walnut St., Philadelphia, PA 19104, USA.
Email: mellers@wharton.upenn.edu

Abstract

We review the classic clinical versus statistical prediction debate as well as related modern work on humans versus. algorithms. Despite the successes of statistical prediction over clinical prediction, there is still widespread resistance to algorithms. We discuss recent attempts to understand that resistance. Current research focuses on when people use algorithmic predictions, how people perceive algorithms, and how algorithms can be made more appealing. We also examine attempts to boost human forecasting accuracy, either by spotting talent, cultivating talent via training, or developing algorithms that aggregate individual forecasts. We hypothesize that hybrid models with both human and algorithmic predictions may encounter less resistance than algorithms alone, especially when the algorithm is “humanized” (with anthropomorphic features) and the human is “algorithmized” (by reducing nose, decreasing bias and increasing signal).

KEYWORDS

AI, algorithms, clinical versus statistical prediction debate, forecasts, predictions

1 | INTRODUCTION

This article examines the role of consumer and firm predictions in the marketplace and beyond. As consumers, we try to predict our preferences for products, services, and experiences on an almost daily basis. Despite the frequency of this process, forecasting is extremely hard (Gal & Simonson, 2021). We make systematic errors, including overpredicting the desire for future variety (Kahneman & Snell, 1992; Simonson, 1990), underpredicting future expenses (Howard et al., 2022; Sussman & Alter, 2012; Ülkümen et al., 2008), overpredicting satiation of high-arousal product designs, such as intense colors or patterns (Buechel & Townsend, 2018), and projecting current desires into future states, such as grocery shopping on an empty stomach (Loewenstein et al., 2003).

Firms make macro and micro forecasts about consumers. Macro forecasts are about aggregate customer behavior, such as market share of a product. Micro forecasts are about individual customers and may involve personalized recommendations, customized images, or tailored promotions. For example, Google News selects personalized stories for customers based on the stories customers previously read and their current locations. Amazon, Netflix, and Spotify likewise recommend products based on customers’ purchase histories. Recommendations shape both the advertising content that customers experience and the products they purchase (Adomavicius et al., 2018; De et al., 2010; Pathak et al., 2010). The accuracy of these predictions is necessary for the success of firms.

This paper reviews both human and algorithmic approaches to micro and macro predictions—and the potential for hybridization. When will people prefer algorithmic predictions or recommendations? What factors make algorithmic predictions more appealing to use? We also discuss advances in subjective-probability forecasting. How much room is there to boost the accuracy of human predictions? How reliably can we distinguish better from worse forecasters? What are the best aggregation algorithms for combining individual forecasts? How can we identify more accurate forecasters ahead of time? And we discuss the broader role of prediction in science. Prediction and decision algorithms are here to stay, and although algorithms have yet to encroach on the decision prerogatives of the highest-status actors in society—top corporate and government officials—their influence will continue to grow. The sooner we recognize and accept their predictive benefits, the more accurate our forecasts will be.

2 | PAST RESEARCH ON CLINICAL VERSUS STATISTICAL PREDICTIONS

We begin with a controversy in psychology that dates back 70 years: the debate between proponents of clinical (intuitive) versus actuarial (statistical) prediction (Meehl, 1954). When people make predictions—selecting among job applicants, making medical diagnoses, or spotting students who will later act violently in middle schools or high schools—are they more accurate if they use their intuitions (clinical predictions) or if they base their predictions on simple models (statistical predictions)? Meehl was one of the first to ask this question, and he established the rules for a fair competition. Both people and methods should have access to the same objective data, and statistical methods should be cross-validated in another sample to avoid overfitting. Within those constraints, Meehl collected and compared the relative accuracies of methods for a wide range of studies. He showed that statistical models were generally just as accurate, and quite often more accurate, than human intuitions.

In light of Meehl's findings, Sawyer (1966) noted that types of data could be separated from types of aggregation methods. Data can be subjective, such as interviews and impressions, or objective, such as grades and standardized test scores. Aggregation methods can be intuitive or statistical. By systematically comparing datasets and aggregation methods, Sawyer found that statistical aggregation was reliably more accurate than human intuition for subjective data, objective data, and combinations of both.

Many more studies compared the relative accuracies of human versus statistical predictions. Statistical models were superior for parole decisions (Carroll et al., 1982), applicant selection in academia (Dawes, 1971; Schofield & Garrard, 1975; Wiggins & Kolen, 1971), bankruptcies (Libby, 1976), cancer survival times (Einhorn, 1972), myocardial infarction (Goldman et al., 1988; Lee et al., 1986), and neuropsychological diagnosis (Leli & Filskov, 1984; Wedding, 1983). Statistical predictions continued to outperform intuitive predictions.

Meta-analyses reached similar conclusions. Grove et al. (2000) surveyed the literature in psychology, medicine, forensics, and finance. Statistical predictions had a small but consistent advantage over human predictions ($d = .12$). Ægisdóttir et al. (2006) gathered human intuitions and statistical predictions of mental health practitioners. Their findings resembled those of Grove et al. (2000); statistical prediction had a reliable edge over human intuition. Humans still have a major role to play. They can provide useful input data and build theoretical models, but algorithms are far better at aggregating the information.

2.1 | Winning doesn't mean acceptance

Rarely do psychological debates produce a clear winner, but this one did. The evidence favoring statistical prediction was strong. Yet, despite the evidence, many people continued to object to statistical models, preferring human judgments in domains such as employee performance (Kuncel et al., 2013), market demand (Sanders & Manrodt, 2003), fraud (Boatsman et al., 1997), crime (Wormith & Goldstone, 1984), consumer preferences (Swearingen & Sinha, 2001), and medicine (Dawes et al., 1989; Eastwood et al., 2012; Grove et al., 2000; Grove & Meehl, 1996). There were many reasons for the opposition, including the (pre-neural-network) belief that models cannot learn (Dawes, 1979) and humans will learn (Highhouse, 2008). In addition to these beliefs, many people had a general lack of understanding about the benefits of statistical models.

3 | CURRENT RESEARCH ON HUMAN PREDICTIONS VERSUS ALGORITHMIC PREDICTIONS

Today, there is still resistance to algorithms, although many things have changed. We talk about algorithms, not simply statistical models, because algorithms perform additional operations besides prediction, such as searching for information, filtering choice sets, giving recommendations, and making decisions. Vastly more data are available, and powerful algorithms pop up in many aspects of daily life (Manyika, 2022).

Much of the current research on human and algorithmic predictions focuses on when people prefer algorithmic predictions. Castelo et al. (2019) surveyed the willingness of people to trust human versus algorithmic predictions for subjective tasks (i.e., intuitive and emotional) and objective tasks (i.e., quantifiable and measurable). They found an overall trust in human predictions over algorithmic predictions, but on some objective tasks, such as predicting the stock market, forecasting the weather, analyzing data, or giving directions, people were willing to trust algorithmic predictions.

Lee et al. (2018) found similar results while studying preferences for human versus algorithmic decision making with social tasks (i.e., hiring employees) and mechanical tasks (i.e., scheduling work). With social tasks, people preferred humans to make decisions. With mechanical tasks, there was no strong preference. People also evaluated decisions based on fairness, trustworthiness, and liking. With social tasks, human decisions were viewed as fairer, more trustworthy, and more likable. With mechanical tasks, human and algorithmic decisions were viewed as equally fair, trustworthy, and likable.

3.1 | The role of predictive accuracy

Do preferences for human versus algorithmic predictions change depending on the relative accuracy of methods? Some studies examine scenarios with little or no information about the accuracy of one or both methods, other studies investigate scenarios in which participants are told the methods are equally accurate, and still other studies explore situations in which participants learn that methods differ in accuracy. We discuss research conducted on each of these cases.

Dietvorst et al. (2015) asked participants to predict the success of MBA applicants given GMAT scores, interview quality, work experience, and other data. Participants saw predictive variables for

15 MBA applicants. They were told that an algorithm to predict MBA performance was "put together by thoughtful analysts." Participants were randomly assigned to groups that saw different predictions. One group saw algorithmic predictions only, and another made their own predictions. A third group saw algorithmic predictions and made their own, and the fourth group saw neither algorithmic predictions nor made their own. After each trial, participants learned how well the applicant had performed.

Participants were then given a chance to earn bonus money based on their predictions of 10 new MBA applicants. They could tie their bonuses to algorithmic predictions or their own predictions. When participants had seen both types of predictions or only the algorithmic predictions, they preferred their own, a result the authors called algorithm aversion. But when participants had not seen algorithmic predictions either because they had made their own or because they saw no predictions, they preferred to use the algorithms.

Logg et al. (2019) investigated whether people preferred human or algorithmic recommendations when asked to estimate unknown quantities or predict events. Participants made initial responses, were given a human or algorithmic recommendation, and were allowed to update their responses. Half of the participants were told the recommendation came from a human, and the other half were told it came from an algorithm. Respondents were more likely to change their guesses toward the algorithmic recommendation than the human one, despite the recommendations being identical. These findings and those of Dievorst et al. suggest that, with little or no information about the accuracy of the algorithm, people prefer algorithmic predictions to those of humans, a result that Logg et al. called algorithm appreciation.

In another study with no information about the accuracy of human and algorithmic recommendations, Kaufmann and Budescu (2020) asked middle and high school teachers to examine student profiles to decide which students should be given tutoring. Teachers could obtain recommendations from algorithms or humans. They requested and followed human guidance more often than algorithmic guidance. Teachers are more familiar with human recommendations, and without information about relative accuracy, the familiarity of human recommendations may have determined their preferences.

Finally, Promberger and Baron (2006) examined the willingness of people to accept human or algorithmic recommendations in a medical scenario when no information about predictive accuracy was provided. Participants assumed they were patients who needed to make decisions about coronary bypass surgery. They received a recommendation from a physician or an algorithm. Participants were more willing to accept the recommendation from the physician than from the algorithm. Again, with no knowledge about relative accuracies of either type of recommendation, participants' preferences may have been based on familiarity.

What happens when participants are given accuracy information about one or both methods? Longoni et al. (2019) conducted studies in which students decided whether to take a stress test. The analysis of their data would be performed by a human or an algorithm. Both groups were told the accuracy rate was 82% to 85%. When physicians and algorithms were equally accurate, participants preferred analyses that were done by physicians. Otherwise, they preferred the most accurate method (see also Pezzo & Beckstead, 2020).

Allen and Choudhury (2022) studied information technology support workers who had varying amounts of experience to resolve customers' computer problems. Workers could rely on their own knowledge or get advice from an algorithm that was 90% accurate. Allen and Choudhury expected workers with more experience would use the algorithm more often due to their keener ability to judge its accuracy. Instead, workers with medium levels of experience used the algorithm more and performed the best. Less experienced workers rejected the algorithm because they could not assess its accuracy, and more experienced workers rejected the algorithm because they believed they knew more. Perhaps not surprisingly, both of these groups performed worse than workers with medium levels of experience.

3.2 | Experts who use algorithms

A few studies have examined peoples' perceptions of experts who did or did not use algorithmic tools. Arkes et al. (2007) measured peoples' perceptions of physicians with access to decision aids. When they relied on their decision aids, physicians were viewed as less capable of making diagnostic decisions. But once again, participants had no information about the relative accuracies of physicians who used or did not use the decision aids.

Physicians' use of decision aids is perceived positively if the decision aid is described as more accurate than the physician. Pezzo et al. (2022) investigated participants' perceptions of physicians who were said to have misdiagnosed a patient whose outcome turned out poorly. The physician was taken to court. In one scenario, participants were told that the physician had consulted a decision aid that was superior to the physician. In another scenario, the physicians didn't use the decision aid, and in a third scenario, the physician consulted it but ignored the advice. Use of the decision aid increased the perceived innocence of the physician and had a protective effect. The physician who used the decision aid was judged more positively than the physician who didn't use it or the physician who consulted it but dismissed the advice.

3.3 | Making algorithms more appealing

Can algorithmic predictions be made more desirable? Several interventions have been tried, including giving people more control over the algorithm, allowing people to have personalized predictions from the algorithm, and providing people with a greater understanding of the algorithm. Other interventions have made algorithms more human-like, conversational, capable of learning, and able to express degrees of uncertainty. We turn to these studies now.

Dietvorst et al. (2018) found that when people could exert some degree of control over algorithmic predictions, they became more acceptable. Dietvorst et al. asked participants to predict the math scores of 20 high school seniors based on nine variables. Participants were told that an imperfect algorithm had been developed (with the average absolute error provided). They were randomly assigned to one of four groups. The first group could only use algorithmic predictions with no changes. The second group could adjust the algorithmic predictions but only by 10 percentiles. The third group could only change half of the algorithmic predictions but by any amount. In the fourth group, participants could adjust the algorithmic predictions as much as they wished.

Next, all participants except those who saw both types of predictions were asked to predict math scores for another 20 high school seniors. They could use their own predictions or the method they had just used. Most participants chose the method that they had experienced. Participants who were allowed to adjust their predictions, even by small amounts, were likelier to use the algorithms than those who could not change their predictions.

Providing users with a better understanding of the algorithm also makes algorithms more appealing. Yeomans et al. (2019) found that people were more accepting of recommender systems that predicted the funniness of jokes when they knew more about how the algorithm worked. Yeomans et al. gave participants either a sparse or a rich explanation of the algorithm. The sparse explanation said, "… we are going to feed your ratings into a computer algorithm which will recommend some other jokes that you might also like." The rich explanation said, "Think of the algorithm as a tool that can poll thousands of people and ask them how much they like different jokes. This way, the algorithm can learn which jokes are the most popular overall and which jokes appeal to people with a certain sense of humor. Using the database ratings, the algorithm will search for new jokes that are similar to the ones you liked, and dissimilar to the ones you did not like." Participants showed a greater preference for the algorithm when they had the rich explanation. In a similar vein, Cadario et al. (2021) showed that resistance to medical AI systems could be reduced by providing people with information about the system. A one-page explanation of the algorithm with graphics increased acceptance of an AI-based skin cancer diagnostic tool over a human provider.

Algorithms can also be made more appealing for subjective tasks when participants are told that algorithms can perform other subjective tasks. Castelo et al. (2019) gave participants a graph of the S&P 500 index over 1 year and asked them to predict its value in a month. Respondents learned that the stock market was driven by emotions, intuitions, and subjective factors. Half of the respondents learned that algorithms could create music and art and predict the popularity of songs. The other half learned that algorithms were incapable of performing these tasks (although in reality, they can). When participants knew that algorithms could perform other human-like tasks, the algorithms became more acceptable.

Informing people that algorithms can give personalized output also encourages their usage. Longoni et al. (2019) asked participants to assume the role of a patient considering surgery for a heart condition. A physician or an algorithm would analyze their data and provide a recommendation. Some participants were told the analysis was personalized and tailored to their unique characteristics. Others were just told that an analysis would be done. When the analysis was described as customized, participants were equally likely to accept the physician's recommendation or the algorithm's recommendation.

Getting advice often involves interactions, especially if the person wanting the advice has follow-up questions. Hildebrand and Bergner (2021) demonstrated that conversational robo-advisors (as opposed to non-conversational robo-advisors) increased perceptions of trust in algorithmic advice of a financial services firm. Conversational robo-advisors used a dialog-based process of inquiry that resembled aspects of human conversations, such as turn-taking. Hildebrand and Bergner (2021) showed that conversational robo-advisors increased consumers' likelihood to follow portfolio recommendations, attributions of benevolence toward the financial services, and enjoyment of the overall experience.

Another study about the benefits of AI systems with human-like features was conducted by Waytz et al. (2014). They investigated driving simulations when participants either drove a normal car, an autonomous vehicle that controlled steering and speed, or a similar autonomous vehicle with anthropomorphic features—name, gender, and voice. Participants who drove the anthropomorphized vehicle were more satisfied than those who drove the other two vehicles. They trusted their vehicle more, felt more relaxed in an accident caused by another car, and blamed their vehicle less for the accident.

In their research, Waytz et al. (2014) used a female voice, a choice that may have increased the acceptability of the autonomous vehicle. Borau et al. (2021) investigated gender preferences for bots in a healthcare context and found that people preferred female bots to male bots because they were perceived as more human. Female bots were also viewed as likelier to consider the unique needs of the individual.

People are likelier to use predictions when they believe those predictions can improve. Berger et al. (2021) asked participants to imagine working as the manager of a call center, responsible for staffing decisions. The call center had taken a new client, and the manager needed to estimate new call volume. Participants were given estimates of call volume from humans or algorithms based on six variables. Some participants saw predictions that improved, while others did not. When giving their final estimates, participants were more likely to use either human or algorithmic predictions when they improved. Algorithmic predictions that improved were more acceptable than human predictions that did not.

Using a different task, Filiz et al. (2021) further explored the effects of algorithmic improvement. Participants predicted a future stock price given prices on 40 previous periods. Participants made several predictions about prices. They were incentivized to be accurate, given clear feedback, and allowed repeated opportunities to improve. Later, they were asked whether they wanted to use algorithmic predictions or make their own. As rounds progressed, participants used the algorithm more often and thereby performed better on the task.

Some people prefer advice expressed with degrees of uncertainty. Tools that calculate confidence and display them in comprehensible ways can increase satisfaction with algorithms. However, this type of information requires at least some knowledge of statistics and probability (Goodyear et al., 2016; Kuncel, 2008; Lodato et al., 2011; Sanders & Courtney, 1985; Whitecotton, 1996). Further understanding of random and systematic errors is also important for an understanding of uncertainty (Kahneman et al., 2021).

4 | IMPROVING HUMAN PREDICTIONS

Since the clinical versus statistical debate, we know much more about how to boost the accuracy of human predictions. IARPA, the research wing of the US intelligence community, conducted a large-scale investigation of human forecasting for high-stakes geopolitical outcomes in a series of forecasting tournaments conducted between 2011 and 2015 (Mellers et al., 2014; Tetlock & Gardner, 2015). IARPA created a level playing field among five competing research groups by challenging them to provide forecasts of the same events over the same period of time. IARPA scored all forecasts for accuracy using the same Brier (1950) scoring rule, a "strictly proper" rule that incentivizes forecasters to report their true beliefs.

Many people, including an author of this paper, were involved in a research group known as The Good Judgment Project. The Good Judgment Project recruited thousands of volunteer forecasters from a wide range of professions and countries. Each year, they were given over 100 wide-ranging geopolitical forecasting questions that remained open for an average of 3 months. Forecasters could update their beliefs about questions at any time prior to the resolution of the question. Questions included, "Will any country officially announce its intention to withdraw from the Eurozone before April 1, 2013?," "Will the number of registered Syrian refugees reported by the United Nations High Commissioner for Refugees exceed 250,000 at any point before April 1, 2013?," "Will the World Health Organization report any confirmed cases of Ebola in a European Union member state before 1 June 2015?," and "On September 15, 2014, will the Arctic Sea ice extent be less than that on September 15, 2013?" Questions spanned such a wide range of topics that no single forecaster could possibly be an expert in all domains.

Initially, researchers in the Good Judgment Project were skeptical about whether human geopolitical forecasting skill even existed. If accuracy was largely attributable to luck, there would be little internal consistency in forecasting accuracy across questions or years. To our surprise, a gauge of internal consistency known as Cronbach's alpha was 0.88, which suggested a considerable degree of skill. In addition, the correlation between accuracy scores of individuals from year to year was as high as .71 (Mellers et al., 2014).

At the beginning of each tournament, the Good Judgment Project gave participants a battery of psychological and political knowledge tests (Mellers, Stone, Atanasov, et al., 2015). Cognitive ability, political knowledge, and open-mindedness were the strongest dispositional correlates of forecasting accuracy. Longer deliberation times, greater frequency of belief updating, smaller steps in updating, and greater discrimination of uncertainty along the probability scale were the strongest behavioral correlates.

The Good Judgment Project won the tournament each year using three interventions: (1) training forecasters in probabilistic reasoning, (2) placing forecasters in teams to work together, and (3) selecting the best forecasters to make predictions in elite teams (Mellers, Stone, Murray, et al., 2015; Tetlock & Gardner, 2015). The probability training module was a 45-min interactive tutorial that, in many respects, encouraged algorithm-like thinking, urging forecasters to look for comparison classes and use base rates. The module also urged them to average estimates when confronted with contradictory clues and advice. Chang et al. (2016) showed that probability training increased forecaster accuracy by 6% to 11% each year for 4 years.

The second intervention of team forecasting (with groups of 10 to 15 members) also improved accuracy relative to individuals who forecasted alone. Despite the statistical argument that independent forecasts will have errors that balance out, the information-pooling benefits of working collaboratively exceeded those of independence. Team members shared information, explained their rationales, and corrected the errors of team members. The presence of other forecasters motivated team members who wished to perform well in the presence of others (Hertel et al., 2000).

Finally, the third intervention of placing the best forecasters in elite teams dramatically increased accuracy. At the end of each tournament, The Good Judgment Project selected the top 2% of forecasters to become *superforecasters*. They were placed in special teams of 10 to 15 members. This social incentive proved to be the most potent single intervention; superforecaster teams were 30% to 40% more accurate than regular teams and even outperformed professional analysts working for the intelligence community (Goldstein et al., 2016).

4.1 | What caused the improvements?

To investigate how the three interventions—probability training, teamwork, and allowing the best forecasters to work together—improved performance accuracy, Satopää et al. (2021) developed a Bayesian BIN model (Bias, Information, Noise) that disentangled three processes underlying forecasts. Improvement can come from reducing systematic bias, decreasing noise (random errors), and amplifying valid signals. The BIN model provided a method for revealing which of these processes were responsible for the accuracy boosts from each of the three interventions used by the Good Judgment Project. Bayesian algorithms at the heart of the BIN model made an interesting discovery: The most reliable path to forecasting improvement was noise reduction.

Prior to working on the BIN model, we had hunches about how training, teaming, and tracking forecasters would improve accuracy. We expected that probability training would reduce cognitive biases, such as base rate neglect, by encouraging individuals to adopt the outside view (Kahneman, 2011). We expected forecasters who worked collaboratively to be less biased and better informed due to the sharing of knowledge. Lastly, we expected superforecasters in elite teams to have less bias, less noise, and more information (Mellers, Stone, Murray, et al., 2015).

The BIN model produced an unexpected set of results. With all three interventions, noise reduction was the most important factor. The main difference between trained and untrained forecasters was that trained forecasters were less noisy. The main difference between forecasters who worked alone and those who worked in teams was that teams were less noisy, with a slight reduction in bias and increase in signal. Superforecasters proved to be the most informed, least noisy, and least biased.

5 | IMPROVING THE ELICITATION OF FORECASTS

Another advance since the clinical versus statistical prediction debate has been the focus on crowd forecasts over individual forecasts (Sunstein, 2006; Surowiecki, 2005). The average of many forecasts, for example, often outperforms the majority of individuals in a crowd by canceling out noise in independent judgments. Simply averaging forecasts does not, of course, guarantee greater accuracy. For example, Simmons et al. (2011) showed that averaging forecasts in a sports betting tournament did not improve crowd wisdom because many individuals had faulty assumptions that led them to predict "favorites" more than "underdogs." But averaging is often an improvement over individuals or experts.

5.1 | Prediction markets

For many years, economists have argued that the fastest way to discover the truth is to rely on self-correcting market mechanisms in which people buy or sell shares of future contracts that pay a fixed amount if an event occurs and nothing otherwise. The price at which a contract trades (the market equilibrium) is, in effect, a collective probability estimate that the event will happen.

To illustrate, imagine a contract that paid $100 if Donald Trump were to be reelected as president in 2024 and $0 otherwise. If that contract were trading at $30, we would say the market predicts that Trump has a 30% chance of reelection. Someone who thinks the probability is higher should buy shares, which will increase the price. If someone else thinks the probability is lower, that person should sell shares, which will decrease the price.

Firms use prediction markets to estimate the likelihood of events, such as whether the company will meet a deadline, reach a level of product quality, or obtain a level of market share (e.g., Cowgill & Zitzewitz, 2015). New product ideas are crowdsourced using consumers and experts in a process known as innovation tournaments. Innovation tournaments allow firms to weed out lower-quality ideas so that only the most promising ones survive (Terwiesch & Ulrich, 2009). Researchers have attempted to fine-tune preference markets in several ways by exploring the effects of positive versus negative feedback on ideas (Camacho et al., 2019) and sharing ideas with others (Hofstetter et al., 2021).

The Good Judgment Project examined prediction markets along with other elicitation methods. Volunteers were randomly assigned to continuous double auction markets (with multiple buyers and sellers). Prediction markets were generally accurate for geopolitical events (Atanasov et al., 2017). Prediction markets did especially well with short-term questions, perhaps because traders did not want to tie up their assets for long periods of time. Long-term questions have opportunity costs.

Prediction markets have an impressive track record for other outcomes, from US elections (Iowa Electronic Markets) to box office revenues for newly released films (Hollywood Stock Exchange, PredictIt, and Predictwise). They have outperformed experts on company sales projections (Plott & Chen, 2002), journalists on Oscar winners (Pennock et al., 2001), and professional macroeconomists on macroeconomic indicators (Gürkaynak & Wolfers, 2021).

A mechanism known as preference markets is popular among firms for evaluating potential products (Dahan et al., 2010). In preference markets and in securities trading of concepts (Dahan et al., 2011), participants trade potential products that consist of bundles of features to measure the strength of the crowd's preferences. Preference markets function like beauty contests in which the winner is the trader who most accurately predicts the preferences of others.

5.2 | Prediction polls

Prediction polling is a popular competitor to prediction markets for gathering crowd wisdom. Prediction polls are not the same as opinion polls. In opinion polls, respondents are often asked about their personal preferences or intentions on one occasion. In prediction polls, respondents make predictions for future events over an extended period of time, as done in prediction markets. Forecasters can update their predictions as often as they wish until a question resolves. The Good Judgment Project tested prediction polls using individuals, teams, and superforecasters (Mellers et al., 2014). Participants made probabilistic forecasts, either independently or in teams, and updated their beliefs as they learned more. When questions closed, forecasters received feedback about their performance using the Brier scoring rule. Prediction polls also proved to be an effective method for eliciting geopolitical forecasts. Later, we will discuss the accuracy of prediction markets relative to prediction polls.

6 | IMPROVING AGGREGATION ALGORITHMS

Given the increased focus on eliciting forecasts from crowds, researchers have explored many algorithms for aggregating individual forecasts (Clemen & Winkler, 1999; Cooke, 1991). A simple aggregation rule is the mean or median (Keuschnigg & Ganser, 2017; Lee &

Lee, 2017; Makridakis & Winkler, 1983). There are many variations on those themes, such as weighted means or averages of quantiles that can yield even sharper forecasts (Lichtendahl et al., 2013). Going beyond a simple average, the Good Judgment Project found that the most successful aggregation rule for prediction polls was a weighted average that favored recent forecasts and forecasters who had either performed well in the past or who had updated their forecasts more frequently on the target question. The weighted average is transformed by a log-odds extremizing function to reflect the diversity of the crowd (Baron et al., 2014; Budescu & Chen, 2015; Chen et al., 2016; Satopää, Baron et al., 2014; Satopää, Jensen, et al., 2014).

This transformation is applied because averages do not describe all of the information in individual forecasts (Baron et al., 2014). There is an intuitive way to think about the transformation. Imagine a variety of groups using forecasting methods to make the most accurate possible predictions of the 2024 US presidential election. Groups could be using prediction markets, prediction polls, opinion surveys, or election models based on variables like the economic growth, favorability ratings, and incumbency status (Abramowitz, 1988; Fair, 1978; Murphy, 1988). Suppose that forecasts made by these groups typically disagree, but in this election, they all fall within a range of 35% to 45% that Trump will win. The fact that groups who usually disagree happen to agree suggests the aggregate may contain more information than it would otherwise.

In these cases, the transformation pushes the weighted aggregate toward the ends of the probability scale. A value of 35% to 45% might be recalibrated to a lower value, such as one between 20% and 30%. Conversely, if diverse groups converged on forecasts ranging from 55% to 65% for a Trump victory, the transformation would push the weighted aggregate to a more extreme value (above 50%), such as one between 70% and 80%. When forecasting methods that are often uncorrelated show agreement, the aggregate should be treated as more informative than if the same forecasting methods that typically agree show agreement. This idea was demonstrated by Wallsten et al. (1997) and later shown to hold under reasonable assumptions by Wallsten and Diederich (2001).

The Good Judgment Project compared the accuracy of aggregated prediction polls to the accuracy of prediction markets (Atanasov et al., 2017; Dana et al., 2019; Mellers & Tetlock, 2019). As mentioned earlier, prediction polls quantify uncertainty using probability judgments, and prediction markets quantify uncertainty using prices that can be converted into probability estimates. Using the Brier scoring rule, we compared the accuracy of both methods and found that the relative winner depended on how probability judgments from prediction polls were aggregated. With simple averages, prediction markets outperformed prediction polls (Atanasov et al., 2017). However, the advantages of prediction markets disappeared when forecasts from prediction polls were aggregated with a transformed weighted average.

The Good Judgment Project conducted an even more demanding test of the relative accuracy of prediction markets and prediction polls (Dana et al., 2019). We used a within-subject design in which participants made a probabilistic judgment before they could buy or sell shares. This design ensured that participants had access to the same information while making exchanges and assigning probabilities—the last trading price and the history of trades in the order book. Once again, prediction markets beat prediction polls when forecasts were combined using a simple mean. But prediction polls outperformed prediction markets when forecasts were aggregated with a transformed weighted average. We further compared a hybrid method using both markets and polls. The hybrid approach proved to be more accurate than the prediction market alone (Dana et al., 2019).

An important reason for the success of crowd wisdom is that individuals in the crowd possess overlapping and unique insights (Chen et al., 2004). The problem of asymmetric information might imply that most individuals have “shallow” information and only a few possess the information needed to make an accurate forecast. To identify more accurate forecasters, some researchers have turned to confidence judgments. However, confidence ratings do not solve the problem because forecasters use different reference groups for evaluating their abilities. Nonetheless, there is a different kind of information that helps solve the problem—predictions about the beliefs of others (Prelec et al., 2017). Such predictions allow researchers to select and use as their forecast the “surprisingly popular answer” or the answer that is chosen by the crowd more frequently than the crowd itself predicts.

The intuition behind this method is that if participants select an answer despite believing they are in the minority, they may possess information that they believe is not widely known. The surprisingly popular answer provides the correct answer in theory across large samples of Bayesian respondents, as well as in practice in a wide range of domains (Prelec et al., 2017). The surprisingly popular answer has been successful at predicting the outcomes of NFL football games (Lee et al., 2018) and purchase intention surveys (Radas & Prelec, 2019). Palley and Soll (2019) have also used forecasters' predictions of others as part of a model for forecasting continuous quantities.

7 | MAKING PREDICTIONS IN SCIENCE

The ability to make accurate predictions is one mark of a good science. Cognitive and behavioral sciences in general, and consumer behavior in particular, have long been concerned with formulating and testing theories of behavior. There is, however, a growing interest in models that make accurate predictions in hold-out samples (Hofman et al., 2021; Yarkoni & Westfall, 2017). Successful predictions of hold-out samples require machine learning and cross-validation, not simply data fitting to an in-sample model.

A related trend is for researchers to conduct competitions among teams who predict data from a set of training data provided to all teams (Erev, Ert, & Roth, 2010; Erev, Ert, Roth, Haruvy, et al., 2010). Erev, Ert, and Roth organized competitions to evaluate models that predicted three types of choices: one-shot decisions under risk, one-shot decisions from experience, and repeated decisions from experience.

Another prediction competition called the "Fragile Families Challenge" required participants to predict six variables for thousands of families based on thousands of variables measured over 15 years (Salganik et al., 2020). Researchers predicted life outcomes in the last wave of the data collection, including child grade point average (GPA), child grit, household evictions, household material hardship, primary caregiver layoff, and primary caregiver participation in job training. Salganik et al. found that even the best models were not much better than a baseline model, and the baseline model was not particularly good. Despite their importance, certain outcomes are extremely difficult to predict, even with access to thousands of variables and powerful prediction tools.

There is also a growing interest in the predictions of experts, especially academics, about the results of laboratory and field experiments (DellaVigna et al., 2019). DellaVigna and Pope asked academics to predict the outcomes of a large-scale online experiment that tested the effects of treatments designed to motivate effort. On average, academics provided forecasts that were more accurate than forecasts made by nonexperts (DellaVigna & Pope, 2018a, 2018b). However, in another large-scale study with treatments designed to increase physical exercise, academics were no better than other groups nor did they outperform the baseline model (Milkman et al., 2021).

8 | LOOKING FORWARD WITH HYBRID PREDICTIONS

Algorithmic predictions are increasingly important to modern life. In healthcare, virtual coaches recommend activities for individuals (Bickmore et al., 2016; Grolleman et al., 2006; Hudlicka, 2013), and AI systems make diagnoses about individual patients, predict health outcomes from an individual's genome, and evaluate the risk of individual patients (Cadario et al., 2021; Davenport & Kalakota, 2019). Judicial and law enforcement sectors also rely on algorithms to predict those who are eligible for bail (Kleinberg et al., 2017) and criminal sentences (Angwin et al., 2022). A missed opportunity in the debate over clinical versus statistical prediction was that of hybridization. Virtually all of the research focused on the relative accuracy of clinical versus statistical predictions without considering whether the combination of predictions could outperform either set of predictions alone. An exception was Blattberg and Hoch (1990) who found that catalog sales and customers' coupon redemption rates were best predicted from the combination of a statistical model and a manager's intuitive judgment than from either method alone. Models could predict trends, and managers were more aware of unique cases. Blattberg and Hoch surmised that models were too consistent and managers were too flexible which made them an ideal combination. If more studies had tested hybrid methods in addition to comparing relative accuracies, we would know more about situations in which humans and models were complementary.

When it comes to making predictions, humans and algorithms have strengths and weaknesses. Humans know what questions to ask and what variables may be useful for accurate predictions. Humans can be flexible about quickly changing conditions. Humans might also be aware of highly unusual but diagnostic cues that are so rare they are not incorporated into models. Yet, humans suffer from errors. They are overconfident and sensitive to irrelevant factors. They get tired, bored, and emotional. Algorithms are less noisy and immune to social or organizational pressures. Algorithms can optimally weigh and aggregate evidence. But too much consistency becomes rigidity.

In the spirit of hybrid models, Steyvers et al. (2022) developed a Bayesian framework that combines human predictions, algorithmic predictions, and confidence measures to determine when hybrid approaches are most likely to succeed. When the correlation between human and algorithmic methods is low, hybrid approaches can be advantageous. They will work best when humans and algorithms are independent yet each has reliable and valid information.

Humans often increase errors when they simply adjust algorithmic predictions (Önkal et al., 2009: Khosrowabadi et al., 2022), but, sometimes, humans and algorithms work better together. For example, Graefe et al. (2014) examined predictions of six US presidential elections from 1992 to 2012. They collected inputs from experts, polls, models, and Iowa Electronic Markets. By averaging similar inputs and predicting elections using the combination of different inputs, they showed that a hybrid model was more accurate than any of the inputs by themselves.

In another study, Phillips et al. (2018) tested the accuracy of experts and super-recognizers against machine learning algorithms at recognizing human faces. By combining the most accurate models with the most accurate facial examiners, predictions were more accurate than those based on combinations of humans or combinations of algorithms.

In the medical domain, Patel et al. (2019) developed a collective intelligence platform, Swarm, that combined the predictions of networked radiologists working together in real time to diagnose pneumonia from chest radiographs. The accuracy of radiologists working together was compared to radiologists working alone and two deep-learning AI models. The greatest accuracy was achieved by combining radiologists working together with the deep-learning predictions of AI.

Lastly, Tschandl et al. (2020) examined the accuracy of human predictions and image-based AI predictions to diagnosis skin cancer. When image-based AI systems were used in conjunction with physicians' diagnoses, accuracy was greater than when AI systems or physicians were used alone.We have an extraordinary opportunity to reduce bias and noise in human predictions. Some people still resist algorithms, especially when their professional identities are threatened, and they will continue to resist algorithms until they gain familiarity and trust. We view hybrid models as a natural step in the transition to a world with widespread use of algorithms. Use of algorithms will mean more accurate predictions. And with more accurate predictions, both consumers and firms will be better positioned to improve decisions about both trivial and high-stakes events.

ORCID

Barbara A. Mellers https://orcid.org/0000-0001-9869-5880

Louise Lu https://orcid.org/0000-0003-2397-1216

REFERENCES

Abramowitz, A. I. (1988). Explaining senate election outcomes. *American Political Science Review*, *82*(2), 385–403. https://doi.org/10.2307/1957392

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, *29*(1), 84–102. https://doi.org/10.1287/isre.2017.0703

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., & Cohen, G. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Allen, R. T., & Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science*, *33*(1), 149–169. https://doi.org/10.1287/ORSC.2021.1554

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In K. Martin (Ed.). *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, *27*(2), 189–202. https://doi.org/10.1177/0272989X06297391

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business and Information Systems Engineering*, *63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bickmore, T. W., Utami, D., Matsuyama, R., & Paasche-Orlow, M. K. (2016). Improving access to online health information with conversational agents: A randomized controlled experiment. *Journal of Medical Internet Research*, *18*(1), e5239. https://doi.org/10.2196/JMIR.5239

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, *36*(8), 887–899. https://doi.org/10.1287/mnsc.36.8.887

Boatsman, J. R., Moeckel, C., & Pei, B. K. W. (1997). The effects of decision consequences on auditors' reliance on decision aids in audit planning. *Organizational Behavior and Human Decision Processes*, *71*(2), 211–247. https://doi.org/10.1006/obhd.1997.2720

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology and Marketing*, *38*(7), 1052–1068. https://doi.org/10.1002/mar.21480

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. https://doi.org/10.1287/mnsc.2014.1909

Buechel, E. C., & Townsend, C. (2018). Buying beauty for the long run: (Mis)predicting liking of product aesthetics. *Journal of Consumer Research*, *45*(2), 275–297. https://doi.org/10.1093/jcr/ucy002

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Camacho, N., Nam, H., Kannan, P. K., & Stremersch, S. (2019). Tournaments to crowdsource innovation: The role of moderator feedback and participation intensity. *Journal of Marketing*, *83*(2), 138–157. https://doi.org/10.1177/0022242918809673

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review*, *17*(1), 199–228. https://doi.org/10.2307/3053536

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

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.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, *13*(2), 128–152. https://doi.org/10.1287/deca.2016.0329

Chen, K. Y., Fine, L. R., & Huberman, B. A. (2004). Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, *50*(7), 983–994. https://doi.org/10.1287/mnsc.1040.0247

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), 187–203. https://doi.org/10.1111/j.1539-6924.1999.tb00399.x

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, Ford, and Firm X. *Review of Economic Studies*, *82*(4), 1309–1341. https://doi.org/10.1093/restud/rdv014

Dahan, E., Kim, A. J., Lo, A. W., Poggio, T., & Chan, N. (2011). Securities trading of concepts (STOC). *Journal of Marketing Research*, *48*(3), 497–517. https://doi.org/10.1509/jmkr.48.3.497

Dahan, E., Soukhoroukova, A., & Spann, M. (2010). New product development 2.0: Preference markets—How scalable securities markets identify winning product concepts and attributes. *Journal of Product Innovation Management*, *27*(7), 937–954. https://doi.org/10.1111/j.1540-5885.2010.00763.x

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? The surprising informational value of "just asking". *Judgment and Decision Making*, *14*(2), 135–147.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, *26*(2), 180–188. https://doi.org/10.1037/h0030868

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–682. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

De, P., Hu, Y., & Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science*, *56*(11), 1930–1945. https://doi.org/10.1287/mnsc.1100.1233

DellaVigna, S., & Pope, D. (2018a). Predicting experimental results: Who knows what? *Journal of Political Economy*, *126*(6), 2410–2456. https://doi.org/10.1086/699976

DellaVigna, S., & Pope, D. (2018b). What motivates effort? Evidence and expert forecasts. *Review of Economic Studies*, *85*(2), 1029–1069. https://doi.org/10.1093/restud/rdx033

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science: Systematic collection of predictions of research findings can provide many benefits. *Science*, *366*(6464), 428–429. https://doi.org/10.1126/science.aaz1704

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making*, *25*(5), 458–468. https://doi.org/10.1002/bdm.741

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, *7*(1), 86–106. https://doi.org/10.1016/0030-5073(72)90009-8

Erev, I., Ert, E., & Roth, A. E. (2010). A choice prediction competition for market entry games: An introduction. *Games*, *1*(2), 117–136. https://doi.org/10.3390/g1020117

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., Stewart, T., West, R., & Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47. https://doi.org/10.1002/bdm.683

Fair, R. C. (1978). The effect of economic events on votes for president. *The Review of Economics and Statistics*, *60*(2), 159–173. https://doi.org/10.2307/1924969

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, *31*, 100524. https://doi.org/10.1016/j.jbef.2021.100524

Gal, D., & Simonson, I. (2021). Predicting consumers' choices in the age of the internet, AI, and almost perfect tracking: Some things change, the key challenges do not. *Consumer Psychology Review*, *4*, 135–152.

Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., Acampora, D., Stasiulewicz, C., Walshon, J., Terranova, G., Gottlieb, L., Kobernick, M., Goldstein-Wayne, B., Copen, D., Daley, K., Brandt, A. A., Jones, D., Mellors, J., & Jakubowski, R. (1988). A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *New England Journal of Medicine*, *318*(13), 797–803. https://doi.org/10.1056/nejm198803313181301

Goldstein, S., Hartman, R., Comstock, E., & Baumgarten, T. S. (2016). Assessing the accuracy of geopolitical forecasts from the US Intelligence Community's prediction market. Retrieved from https://goodjudgment.com/wp-content/uploads/2020/11/Goldstein-et-al-GJP-vs-ICPM.pdf

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience*, *10*, 542. https://doi.org/10.3389/fnhum.2016.00542

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, *30*(1), 43–54. https://doi.org/10.1016/j.ijforecast.2013.02.005

Grolleman, J., van Dijk, B., Nijholt, A., & van Emst, A. (2006). Break the habit! Designing an e-therapy intervention using a virtual coach in aid of smoking cessation. Paper presented at the PERSUASIVE 2006: Persuasive Technology, Berlin, Heidelberg

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gürkaynak, R. S., & Wolfers, J. (2021). Macroeconomic derivatives: An initial analysis of market-based macro forecasts, uncertainty and risk. In J. Galí & K. D. West (Eds.). *NBER international seminar on macroeconomics* (pp. 11–64). The University of Chicago Press.

Hertel, G., Kerr, N. L., & Messé, L. A. (2000). Motivation gains in performance groups: Paradigmatic and theoretical developments on the Köhler effect. *Journal of Personality and Social Psychology*, *79*(4), 580–601. https://doi.org/10.1037/0022-3514.79.4.580

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*, 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Hildebrand, C., & Bergner, A. (2021). Conversational robo advisors as surrogates of trust: Onboarding experience, firm perception, and consumer financial decision making. *Journal of the Academy of Marketing Science*, *49*(4), 659–676. https://doi.org/10.1007/s11747-020-00753-z

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Hofstetter, R., Dahl, D. W., Aryobsei, S., & Herrmann, A. (2021). Constraining ideas: How seeing ideas of others harms creativity in open innovation. *Journal of Marketing Research*, *58*(1), 95–114. https://doi.org/10.1177/0022243720964429

Howard, R. C. C., Hardisty, D. J., Sussman, A. B., & Lukas, M. F. (2022). Understanding and neutralizing the expense prediction bias: The role of accessibility, typicality, and skewness. *Journal of Marketing Research*, *59*(2), 435–452. https://doi.org/10.1177/00222437211068025

Hudlicka, E. (2013). Virtual training and coaching of health behavior: Example from mindfulness meditation training. *Patient Education and Counseling*, *92*(2), 160–166. https://doi.org/10.1016/j.pec.2013.05.007

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Kahneman, D., & Snell, J. (1992). Predicting a changing taste: Do people know what they will like? *Journal of Behavioral Decision Making*, *5*(3), 187–200. https://doi.org/10.1002/bdm.3960050304

Kaufmann, E., & Budescu, D. V. (2020). Do teachers consider advice? On the acceptance of computerized expert models. *Journal of Educational Measurement*, *57*(2), 311–342. https://doi.org/10.1111/jedm.12251

Keuschnigg, M., & Ganser, C. (2017). Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Management Science*, *63*(3), 818–828. https://doi.org/10.1287/mnsc.2015.2364

Khosrowabadi, N., Hoberg, K., & Imdahl, C. (2022). Evaluating human behaviour in response to AI recommendations for judgemental forecasting. *European Journal of Operational Research*, *303*(3), 1151–1167. https://doi.org/10.1016/j.ejor.2022.03.017

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.3386/w23180

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology*, *1*(3), 343–346. https://doi.org/10.1111/j.1754-9434.2008.00059.x

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, R. M., Behar, V. S., Floyd, W. L., Morris, J. J., Waugh, R. A., & Whalen, R. E. (1986). Predicting outcome in coronary disease statistical models versus expert clinicians. *The American Journal of Medicine*, *80*(4), 553–560. https://doi.org/10.1016/0002-9343(86)90807-7

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*(4), 322–333.

Lee, M. D., & Lee, M. N. (2017). The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making*, *12*(4), 328–343.

Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology*, *40*(6), 1435–1441. https://doi.org/10.1002/1097-4679(198411)40:6<1435::AID-JCLP2270400629>3.0.CO;2-0

Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance*, *16*(1), 1–12. https://doi.org/10.1016/0030-5073(76)90002-7

Lichtendahl, K. C., Grushka-Cockayne, Y., & Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, *59*(7), 1594–1611. https://doi.org/10.1287/mnsc.1120.1667

Lodato, M. A., Highhouse, S., & Brooks, M. E. (2011). Predicting professional preferences for intuition-based hiring. *Journal of Managerial Psychology*, *26*(5), 352–365. https://doi.org/10.1108/02683941111138985

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics*, *118*(4), 1209–1248. https://doi.org/10.1162/003355303322552784

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(7), 987–996. https://doi.org/10.1287/mnsc.29.9.987

Manyika, J. (2022). Getting AI right: Introductory notes on AI & society. *Daedalus*, *151*(2), 5–27. https://doi.org/10.1162/DAED_e_01897

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Emlen Metz, S., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., & Tetlock, P. (2019). From discipline-centered rivalries to solution-centered science: Producing better probability estimates for policy-makers. *American Psychologist*, *74*, 290–300.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J. J., … Duckworth, A. L. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, *600*(7889), 478–483. https://doi.org/10.1038/s41586-021-04128-4

Murphy, S. (1988). A comparison of the selection of bargaining representatives in the United States and Canada: Linden Lumber, Gissel, and the right to challenge majority status. *Comparative Labor Law Journal*, *10*, 65.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Palley, A. B., & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291, mnsc.2018.3047–2309. https://doi.org/10.1287/mnsc.2018.3047

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarappallil, J., Mariano, A. J., Riley, G., Seekins, J., Shen, L., Zucker, E., & Lungren, M. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, *2*(1), 111. https://doi.org/10.1038/s41746-019-0189-7

Pathak, B., Garfinkel, R., Gopal, R., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, *27*(2), 159–188. https://doi.org/10.2753/MIS0742-1222270205

Pennock, D. M., Lawrence, S., Giles, C. L., & Nielsen, F. Å. (2001). The real power of artificial markets. *Science*, *291*(5506), 987–988. https://doi.org/10.1126/science.291.5506.987

Pezzo, M. V., & Beckstead, J. W. (2020). Algorithm aversion is too often presented as though it were non-compensatory: A reply to Longoni et al. (2020). *Judgment and Decision Making*, *15*(3), 449–451.

Pezzo, M. V., Nash, B. E. D., Vieux, P., & Foster-Grammer, H. W. (2022). Effect of having, but not consulting, a computerized diagnostic aid. *Medical Decision Making*, *42*(1), 94–104. https://doi.org/10.1177/0272989X211011160

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J. C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115

Plott, C. R., & Chen, K. Y. (2002). Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem (1131). Retrieved from Pasadena, CA: https://resolver.caltech.edu/CaltechAUTHORS:20140317-135547085

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/nature21054

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Radas, S., & Prelec, D. (2019). Whose data can we trust: How meta-predictions can be used to uncover credible respondents in survey data. *PLoS ONE*, *14*(12), e0225432. https://doi.org/10.1371/journal.pone.0225432

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Sanders, G. L., & Courtney, J. F. (1985). A field study of organizational factors influencing DSS success. *MIS Quarterly*, *9*(1), 77–93. https://doi.org/10.2307/249275

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega*, *31*(6), 511–522. https://doi.org/10.1016/j.omega.2003.08.007

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. https://doi.org/10.1016/j.ijforecast.2013.09.009

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *Annals of Applied Statistics*, *8*(2), 1256–1280. https://doi.org/10.1214/14-AOAS739

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *Management Science*, *67*(12), 7599–7618. https://doi.org/10.1287/mnsc.2020.3882

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*(3), 178–200. https://doi.org/10.1037/h0023624

Schofield, W., & Garrard, J. (1975). Longitudinal study of medical students selected for admission to medical school by actuarial and committee methods. *Medical Education*, *9*(2), 86–90. https://doi.org/10.1111/j.1365-2923.1975.tb01900.x

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, *38*(1), 1–15. https://doi.org/10.1086/658070

Simonson, I. (1990). The effect of purchase quantity and timing on variety-seeking behavior. *Journal of Marketing Research*, *27*(2), 150–162. https://doi.org/10.1177/002224379002700203

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), e2111547119. https://doi.org/10.1073/pnas.2111547119

Sunstein, C. R. (2006). *Infotopia: How many minds produce knowledge*. Oxford University Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Sussman, A. B., & Alter, A. L. (2012). The exception is the rule: Underestimating and overspending on exceptional expenses. *Journal of Consumer Research*, *39*(4), 800–814. https://doi.org/10.1086/665833

Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. Paper presented at the Proceedings of ACM SIGIR Workshop on Recommender Systems, Berkeley, CA.

Terwiesch, C., & Ulrich, K. T. (2009). *Innovation tournaments: Creating and selecting exceptional opportunities*. Harvard Business Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Random House.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Ülkümen, G., Thomas, M., & Morwitz, V. G. (2008). Will I spend more in 12 months or a year? The effect of ease of estimation and confidence on budget estimates. *Journal of Consumer Research*, *35*(2), 245–256. https://doi.org/10.1086/587627

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*(3), 243–268. https://doi.org/10.1002/(sici)1099-0771(199709)10:3<243::aid-bdm268>3.0.co;2-m

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *41*(1), 1–18. https://doi.org/10.1016/S0165-4896(00)00053-6

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wedding, D. (1983). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology*, *5*, 49–55.

Whitecotton, S. M. (1996). The effects of experience and a decision aid on the slope, scatter, and bias of earnings forecasts. *Organizational Behavior and Human Decision Processes*, *66*(1), 111–121. https://doi.org/10.1006/obhd.1996.0042

Wiggins, N., & Kolen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology*, *19*(1), 100–106. https://doi.org/10.1037/h0031147

Wormith, J., & Goldstone, D. (1984). The clinical and statistical prediction of recidivism. *Clinical Justice and Behavior*, *11*, 3–34. https://doi.org/10.1177/0093854884011001001

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

How to cite this article: Mellers, B. A., Lu, L., & McCoy, J. P. (2023). Predicting the future with humans and AI. *Consumer Psychology Review*, *6*(1), 109–120. https://doi.org/10.1002/arcp.1089