Combining Human and Machine Intelligence for Making Predictions

by

Yiftach Nagar

B.Sc. Industrial Engineering
Tel-Aviv University, 1997

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTERS OF SCIENCE IN MANAGEMENT RESEARCH

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2012 [2013]

Signature of Author: _____
MIT Sloan School of Management
May 1st, 2012

Certified by: _____
Thomas W. Malone
Patrick J. McGovern Professor of Management, MIT Sloan School of Management; Director, MIT Center for Collective Intelligence
Thesis Supervisor

Accepted by: _____
Ezra Zuckerman Sivan
Nanyang Technological University Professor of Technological Innovation, Entrepreneurship, and Strategic Management
Chair, MIT Sloan PhD Program

Combining Human and Machine Intelligence for Making Predictions

by

Yiftach Nagar

Submitted to the MIT Sloan School of Management
on April 15, 2013, in partial fulfillment of the requirements for the degree of
Master of Science in Management

Abstract

An extensive literature in psychology, economics, statistics, operations research and management science has dealt with comparing forecasts based on human-expert judgment vs. (statistical) models in a variety of scenarios, mostly finding advantage of the latter, yet acknowledging the necessity of the former. Although computers can use vast amounts of data to make predictions that are often more accurate than those by human experts, humans are still more adept at processing unstructured information and at recognizing unusual circumstances and their consequences. Can we combine predictions from humans and machines to get predictions that are better than either could do alone? We used prediction markets to combine predictions from groups of people and artificial intelligence agents. We found that the combined predictions were both more accurate and more robust in comparison to those made by groups of only people, or only machines. This combined approach may be especially useful in situations where patterns are difficult to discern, where data are difficult to codify, or where sudden changes occur unexpectedly.

Thesis Supervisor: Thomas W. Malone
Title: Patrick J. McGovern Professor of Management, MIT Sloan School of Management; Director, MIT Center for Collective Intelligence

Acknowledgments

This work would not have been possible without the work, advice, ideas and intelligence – individual and collective – of many people, whose contributions I acknowledge and am grateful for. I thank DuWayne J. Peterson, Jr. for a generous fellowship that supported my graduate research, and to MIT Lincoln Laboratory and the U.S. Army Research Laboratory's Army Research Office (ARO) for funding this project. I am grateful to John Willett for his patience, dedication and help with statistical analyses and to Chris Hibbert for software development, and education about prediction markets. This project originated, and has benefited greatly from discussions with Sandy Pentland, Tomaso Poggio, Drazen Prelec, and Josh Tenenbaum. Jason Carver, Wendy Chang, Jeremy Lai have developed much of the software used in this thesis, and together with Rebecca Weiss have greatly helped with designing and running the experiments. For their wise comments, I thank John Carroll, Gary Condon, Robin Hanson, Haym Hirsh, Ben Landon, Retsef Levi, Cynthia Rudin, and Paulina Varshavskaya. Thanks also go to our undergraduate research assistants – Jonathan Chapman, Catherine Huang, Natasha Nath, Carry Ritter, Kenzan Tanabe and Roger Wong – for their help in running experiments, and to Richard Hill and Robin Pringle for administrative support. Sharon Cayley, the Director of Sloan's PhD program, and Maria Brennan, Assistant Director at the International Students Office have, in many ways and occasions, helped resolve various questions and issues, since before I got to MIT, and made it easier for me to focus on my work. I'm thankful to them as well.

Finally, I am indebted to my extended family, for continuously supporting my studies even when times were not easy; and of course, to my advisor, Thomas Malone, for inspiring, supporting, and constantly encouraging my work.

About the Author

Yiftach Nagar is a doctoral candidate at the MIT Sloan School of Management, and a member of the MIT Center for Collective Intelligence and the MIT Intelligence Initiative. He explores collective forms of intelligence in sociotechnical settings – from face-to-face groups, to distributed collaborations of minds and machines. His research combines social and cognitive psychology, organization studies and computer-supported-cooperative-work (CSCW) and has been featured in top academic venues, and in the media.

Before returning to academia, Mr. Nagar's career spanned R&D, project management, and product management roles in several hi-tech organizations. He has led projects and products which were successfully deployed by some of the world's leading telecom companies in Europe, Asia and the US. Mr. Nagar holds a B.Sc in Industrial Engineering from Tel-Aviv University.

Contents

Introduction

The creation of accurate and reliable predictions[1] has been the subject of extensive research in many fields and disciplines. Theoretical advancements and supporting empirical findings from multiple domains suggest several conclusions.

First, substantial evidence from multiple domains supports the claim that models usually yield better (and almost never worse) predictions than do individual human experts (e.g. Dawes, Faust, & Meehl, 1989; Dawes & Kagan, 1988; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Whereas models (or machines) are better at information processing and are consistent (Einhorn, 1972), humans suffer cognitive and other biases that make them bad judges of probabilities (c.f. Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1973; Lichtenstein, Baruch, & Phillips, 1982; Rabin, 1996). In addition, *"Such factors as fatigue, recent experience, or seemingly minor changes in the ordering of information or in the conceptualization of the case or task can produce random fluctuations in judgment"* (Dawes et al., 1989), and it is therefore not surprising that models of judges often outperform the judges themselves (Armstrong, 2001b; Goldberg, 1970; Stewart, 2001). When working in groups, humans often exhibit effects such as *Groupthink* (Janis, 1972; Janis & Mann, 1977) and *group polarization* (see Chapter 6 in Brown, 1986) that negatively affect their judgment. Nevertheless, the role of humans is still recognized as valuable in real-life situations, for at least two good reasons: humans are still better at retrieval and acquisition of many types of information – especially unstructured types of information (Einhorn, 1972; Kleinmuntz, 1990) and this advantage is not soon to disappear. In addition, humans' common-sense is required to identify and respond to "broken-leg" situations (Meehl, 1954) in which the rules normally characterizing the phenomenon of interest do not hold. Therefore, combining the human and machine/model predictions may help in overcoming and mitigating human and model respective flaws and yield better predictions (Blattberg & Hoch, 1990; Bunn & Wright, 1991; Einhorn, 1972).

---

[1] In this paper, we use the terms predictions/forecasts (and predictor/forecaster) interchangeably.

Second, a vast body of theoretical and empirical research suggests that combining forecasts from multiple independent, uncorrelated forecasters that have relevant knowledge and information leads to increased forecast accuracy[2]. This unanimous result holds whether the forecasts are judgmental or statistical, econometric or extrapolation (Armstrong, 2001a; Clemen, 1989). Combining forecasts is recommended not only for improving accuracy, but also because it may be difficult or impossible to identify a single forecasting method that is the best (Makridakis & Winkler, 1983) and *"it is less risky in practice to combine forecasts than to select an individual forecasting method"* (Hibon & Evgeniou, 2005).

Weaving together the lessons learned from three threads of inquiry – the combination of predictions from different models, the combination of predictions from people, and the comparison of human predictions to model predictions – it is natural to hypothesize that there may be ways to combine predictions from multiple human experts *and* models that will emphasize their relative advantages and mitigate their respective flaws. Indeed, it is surprising that this path has hardly been explored. Previous work by Blattberg & Hoch, Bunn & Wright, and Einhorn, emphasizes the complementary nature of man and model, but does not stress the potential of improving predictions by combining predictions from multiple humans and multiple models. Interacting and/or combining human and model predictions may be especially valuable in scenarios where patterns such as time-series are either very difficult to identify, or where they do not exist at all. The literature explains why mathematical/statistical models such as those that are widely discussed in the forecasting literature, may outperform humans where relatively stable and predictable patterns, such as seasonal sales trends, or weather change, exist. However, many phenomena of interest do not behave in such smooth manner. For example, predicting the behavior of groups of people such as business competitors, or – to use a different example – insurgent groups, may be an important task for an organization.

---

[2] (For discussion of the philosophical and the mathematical principles underlying the logic of combining forecasts, see Armstrong, 2001a; Larrick & Soll, 2006; Makridakis, 1989; Sunstein, 2005, pp. 972-974; Winkler, R. L., 1989).

How can these be predicted? Resulting from the interaction of extrinsic causal forces, purposeful intrinsic goal-oriented action, and local context, the patterns and rules governing such behaviors (to the extent they exist) are complex, and may prove difficult, or even practically impossible to detect or to model. Recent advancements in artificial-intelligence enable artificial-neural-network agents to relatively successfully identify patterns even in complex scenarios (e.g. Jain, Duin, & Mao, 2000; Mannes et al., 2008; Pao, 1989). But while artificial neural networks offer advantages over "static" statistical models, machines are still limited in access to unstructured information, and while the implementation can be adaptive to new data, humans can still probably be better judges of whether unorthodox patterns are due to a real change in the environment, or just plain noise.

We conjecture therefore that when it comes to predicting events under complex situations, where rules that may be fuzzy or difficult to discern, and where some data may not be easily codifiable, combining predictions made by humans with those made by artificial-intelligence agents is a strategy that can prove robust and, in the long-run, outperform both relying solely on human experts (working alone or in groups, but without the aid of the model), or relying solely on artificial-intelligence (or statistical models) with no human intervention.

To test this hypothesis, we performed a study in which humans and artificial intelligent agents made predictions of the moves of teams playing football. We used prediction markets[3] to combine predictions from human subjects and artificial-intelligence agents, and compared 3 conditions: markets with human participants, markets of artificial-intelligence agents, and markets where humans and artificial-intelligence agents participate simultaneously.

---

[3] Also known as information markets, decision markets, electronic markets, virtual markets, idea futures, event futures and idea markets (Tziralis & Tatsiopoulos, 2007; Wolfers & Zitzewitz, 2004).

Study Description

Our goal in this study was to compare the quality of predictions made by 3 different types of predictors: groups of humans, groups of artificial-intelligence agents, and 'hybrid' groups of humans and agents. We used prediction markets to combine the predictions made by humans and artificial-neural-network agents, of the moves of football teams to during a football game. This enabled us to emulate a realistic situation where humans and agents would have access to different information (specifically, humans had access to video information that is difficult or costly to codify for the agents, and in addition, their familiarity with the game may lead them to devise more sophisticated heuristics, which may be right or wrong). We hypothesized that 'hybrid' markets of humans and computers would do better than both markets of computer-agents with no humans and markets of humans with no computer (we discuss what 'better' means in more detail before presenting the results).

Lab Experiments

We conducted 20 laboratory sessions, lasting about 3 hours each, in which human subjects participated in prediction markets, with and without computer agents, as well as 10 more computer-only sessions, totaling 30 experimental sessions. In each of the 20 lab sessions we had a group of 15 to 19 participants (median group size was 18; mean 17.55; mode 19) who participated in the entire session (though they could choose not to 'play' in the markets), totaling 351 subjects overall. For every experimental session we recruited a new group of people, from the general public, mainly through web advertisements. We encouraged the participation of football fans by stressing the fun part and by clearly stating that knowledge of football could help make higher profits, however domain expertise was not a mandatory requirement. Compensation to participants included a base payment and an additional performance-based bonus that was proportional to the ending balance in each participant's account and could reach up to 75% of the base pay.

The AI agents were artificial neural-net agents developed by graduate students from our computer science department, using the *JOONE* open-source package[4]. The information they had for each play included 3 parameters: the down number, the number of yards to first down, and the previous play's move (i.e. whether it was a pass). In addition, the agents considered the market price and would trade only if they were confident about their prediction. The agents were trained on a dataset of plays from a previous game.

After initial explanation[5] and training rounds, each experimental session included 20 plays. For each play, a short video excerpt from a football game was shown to all participants. The video was automatically stopped just before the team possessing the ball was about to make a move. At that stage, a new online prediction market was opened and the group of participants (be it human participants only, or human participants along with AI agents[6]) started trading contracts of RUN and PASS[7]. We ran the experiments using custom-tailored version of the *ZOCALO* open-source prediction markets platform[8], and employed its automated market maker to simplify trading and ensure liquidity in the markets. The market was then closed after 3.5 minutes[9], and the video continued, revealing what had actually happened, and stopping before the next play. The same set of 20 plays (taken from one game) was shown in all the sessions. The AI agents participated either in the first half (first 10 plays) or the second half (last 10 plays) of the experiment (according to a random

---

[4] Available at http://sourceforge.net/projects/joone/

[5] Subjects were given an elaborate verbal explanation on the goal of the experiment, and on trading in the prediction market. In addition, subjects were prompted to read a short manual the day before coming to the lab, doing which – as they were truthfully told, would raise their chance to succeed in the markets and make a higher bonus. We regularly checked by vote of hands how many of them actually read the manual and the overwhelming majority did. The manual was also available on subjects' screens, though they rarely, if ever, referred to it during the sessions.

[6] We ran 10 neural-net agents. They used the same code and same training dataset, but their logic of trading was also based on the market price, and they were started in a staggered manner so that they encountered different market conditions.

[7] RUN and PASS were two exhaustive options in this case (we eliminated other moves from the video).

[8] Available at http://zocalo.sourceforge.net/

[9] We decided on 3.5 minutes after several pilot sessions in which we monitored participation and noted when trading had stopped. We alerted participants 30 seconds prior to closing each market.

draw previously performed). Human participants were told that AI agents would trade in some of the markets but were not told in which, and could not generally tell. Thus in each lab session we collected data from 10 'human only' markets and 10 'hybrid' (humans and agents) markets. In addition we ran 10 "computer-only" experimental sessions with no human participants, where the agents traded in all 20 markets, predicting the same plays. Those were conducted in a similar manner, other than the relevant technical adjustments (e.g. agents read the play parameters through an API as in the lab experiments but we did not need to project the video, etc.). We thus got a total of 600 observations (10 observations of each of our 3 conditions for each play).

In our analysis, we took the market closing price as representing the collective group estimation of the probability of the football team to either RUN or PASS the ball. While the exact degree of accuracy to which the market closing price genuinely represents the mean belief of market participants may not be fully known (Manski, 2006), it is widely acknowledged de-facto as the best estimator of it (see Wolfers & Zitzewitz, 2004, 2006).

Results

Assessing the outcome: What makes a better predictor

Prediction quality is a multidimensional concept that aims to capture the degree of correspondence between predictions and observations. There are many measures by which predictions can be assessed, but no single measure is sufficient for judging and comparing forecast quality (Jolliffe & Stephenson, 2003). Thus, assessment of prediction quality is a matter of analyzing and understanding trade-offs. To compare the three groups of predictors, we therefore look at three criteria common in the forecasting literature: ***Accuracy, Reliability (a.k.a Calibration)*** *and* ***Discrimination*** which, combined, help understand those trade-offs. We augment our analysis with a comparison of accuracy vs. variability, using the Sharpe ratio (Sharpe, 1966, 1994), commonly used in economics to compare reward-vs.-risk performance, and then also present an analysis based on the ***Receiver-Operating-Characteristic (ROC)*** approach (Swets, 1988; Swets & Pickett, 1982;

Zweig & Campbell, 1993) that has been established and widely accepted in many domains as a method of assessing and comparing predictors who make predictions about binary events. The ROC analysis is in itself a trade-off analysis, which sheds more light on our findings.

Accuracy

Accuracy is a measure or function of the average distance/error between forecasts and observations. A common way to assess the accuracy of predictions and to compare the skill of the people or methods that created them is to use a scoring rule. TABLE 1 summarizes the evaluations of accuracy for the human-only markets, agent-only markets, and hybrid markets, over the experimental play set, according to three popular scoring rules: the Mean Absolute Error (MAE), the Mean Square Error (MSE[10]) – also known as the Brier Score (Brier, 1950), and the Log Scoring Rule (LSR, introduced by Good, 1952).

TABLE 1 – ACCURACY OF PREDICTION MARKETS

	Scoring Rule		
	Mean Absolute Error	Mean Squared Error	LSR
Humans-only Markets	0.415	0.197	0.250
Agents-only Markets	0.349	0.172	0.225
Hybrid Markets	0.350	0.150	0.205

The lines to the right of the MAE and the LSR scores indicate where the differences between the scores of the different predictors were found statistically significant[11] ($p<0.05$). Under all

[10] Some authors offer that comparing forecasts/forecasters accuracy based on the Mean Square Forecasting Error (Brier score) has some limitations – (cf. Clements & Hendry, 1993; Ferro, 2007; Jewson, 2004), and have suggested other measures such as the likelihood function (Clements & Hendry, 1993; Jewson, 2004), and many other (Diebold & Mariano, 1995). Another point that is all-too-often ignored is that since the mean square error is not normally distributed, it also has the drawback of not yielding to ordinary parametric tests of statistical significance without violating assumptions underlying the statistical analysis.

[11] To compare the conditions we built a mixed model to account for nesting, and used SAS's PROC MIXED (Littell, Milliken, Stroup, & Wolfinger, 2006) with the first-order autoregressive AR(1) error-covariance-matrix structure (ibid., pp. 175-176). The squared errors are not normally distributed,

of these scoring rules, a score that is closer to zero is better, and under all of them a perfect predictor who assigns a probability estimation of 100% to actual events and a probability of zero to all other potential options will score zero.

Albeit popular, this comparison should be interpreted with care that is too often lacking in analyses. Scoring rules (to be exact: *proper* scoring rules[12]) are useful in eliciting honest probability estimations. But, as Winkler (1969) cleverly points out, using them to evaluate and rank predictors ex post may be misleading, as it confounds the measurement of accuracy with the cost function of errors. Different scoring rules punish small and large errors to different extents, and can yield contradicting results when used to rank predictors. Indeed in our table, the Hybrid markets are more accurate, on average, than the Agent-only markets according to the MSE and the LSR, but not according to the MAE where they tie. It is up to the decision maker therefore to select the rule to be used for evaluation, and this would be done according to the nature of the setting and the corresponding cost functions. For example, in weather forecasting, small errors are tolerable on a daily basis (say, ±1 degree in temperature predictions), but big errors (predicting a very hot day which turns out to be very cold, or failing to predict a tornado) are not. In a production setting, on the other hand, it may be OK to throw away a unit due to a large prediction error on rare occasions, but precision is very important on a regular basis. While there may be some ambiguity in selecting a scoring rule when the cost of errors is unknown, in our case it appears that the number of large errors matters more than the average accuracy (e.g. it is likely that a prediction of 90% and prediction of 95% for a PASS attempt by the offense team would both translate to the same decision by the defense team) and hence, the MSE and the LSR seem more appropriate than the MAE.

---

which hinders a parametric statistical comparison of the MSE scores. Distributions of the absolute errors and of the log-predictions are quasi-normal.

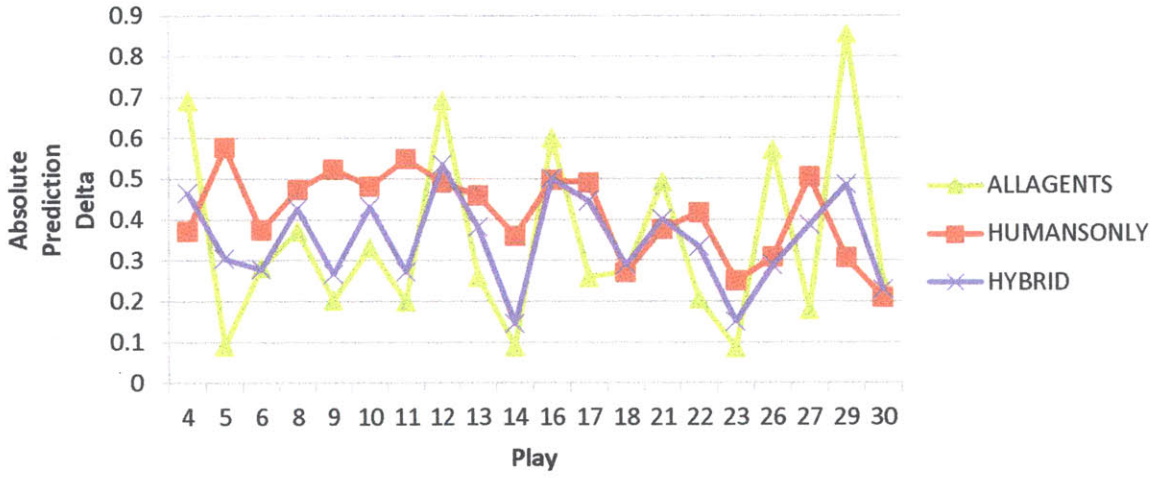
[12] A scoring rule is proper if the forecaster maximizes the expected score for an observation drawn from the distribution F if he or she issues the probabilistic forecast F, rather than $G \neq F$. In prediction problems, proper scoring rules encourage the forecaster to make careful assessments and to be honest (Gneiting & Raftery, 2007)

Taken together, these results suggest that the hybrid markets were the most accurate – beating both human-only and agent-only markets under the MSE and LSR, and yielding a tie with the agents under the MAE. We also note that although the agents were very simple, on average the agent-only markets were more accurate, than the human-only markets, as one could expect based on previous evidence. We later turn to use the ROC method to make a comparison that is agnostic to the cost of errors, but first we explore how well our predictors predicted each play and consider a few other criteria.

A deeper look at the play level

A deeper look at the play level provides better understanding of the behavior of the predictors and reveals several interesting patterns. FIGURE 1 depicts the mean absolute prediction error (average of 10 observations from 10 markets) of each condition, per play.

FIGURE 1 – MEAN PREDICTION ERRORS OF HUMAN, AGENTS AND HYBRID MARKETS[13]



We note a strong interaction between condition and play. As could be expected, humans and agents predicted differently on different plays. While *on average* agents were more

[13] (The play numbers are actual numbers of plays from the game, by order. The plays that are not included in the graph were not included in the experiment as they were not clearly defined as RUN/PASS). There is a total of 20 plays.

accurate than humans (i.e. had smaller errors), in several cases they made severe errors while humans predicted correctly (notably: plays 4, 12, 16, 29). But why? Informal interviews with subjects suggested that they incorporated more information into their decision-making than did the agents. Notably, they gleaned from the video the formation of the offensive and the defensive teams. For example: before both play 4 and play 29, the offense team formed a "Shotgun" formation[14], with a running-back standing next to the quarterback, which to football savvy fans implies a higher probability for a pass attempt. In both those plays, the 'human-only' markets clearly indicated a pass (70% and 77% on average) whereas the 'all-agents' markets indicated a RUN (69% and 85.5% on average, corresponding to 31%, 14.5% predictions for PASS). A few subjects also reported that commentary by anchors was helpful, and several others mentioned that the body language of players was revealing.

Beyond mean errors: considering prediction-error variability

Measures of accuracy alone do not provide sufficient information to convey the complexity of the data, as they are essentially comparisons of single numbers representing entire distributions. Two predictors can yield the same mean F (Prediction Error), where F is some scoring rule, and yet offer very different predictions and risk profiles. Therefore, it is important to consider the variability of prediction errors of the different predictors being compared. After assigning economic values to the predictions using scoring rules, the ex post Sharpe ratio (Sharpe, 1966, 1994), originally developed to compare *reward-to-risk* performance of mutual funds, enables us to consider accuracy against variability of prediction errors, making the comparison more informative.

Too keep with the familiar logic of the Sharpe ratio that assumes a higher positive financial return is better, we adjust our scoring rules such that the adjusted MAE score (AMAE) equals 1-MAE and the adjusted MSE score (AMSE) equals 1-MSE. The adjusted Log score is

---

[14] (Mallory & Nehlen, 2006 ch.7-8)

$\log_{10}$(Absolute Prediction Error)-1. We calculated the Sharpe ratio according to equations 3-6 in Sharpe (1994, p. 50). As a simple and straightforward benchmark, we use an "ignorant" predictor who bets 50% PASS all the time (and whose error variance is therefore zero). The corresponding AMAE, AMSE and ALSR for the benchmark predictor are therefore 0.5, 0.75 and 0.699, correspondingly. The results are summarized in TABLE 2.

TABLE 2 - EX POST SHARPE RATIO FOR PREDICTION MARKETS, UNDER 3 SCORING RULES

	Scoring Rule		
	AMAE (Benchmark = 0.5)	AMSE (Benchmark = 0.75)	ALSR (Benchmark = 0.699)
Humans-only Markets	0.54	0.41	0.41
Agents-only Markets	0.67	0.39	0.37
Hybrid Markets	0.91	0.74	0.72

Clearly, the hybrid markets yield the highest Sharpe ratio and outperform both the human-only and agent-only markets. This result holds under three different scoring rules. According to the Sharpe ratio index criterion, therefore, the Hybrid markets are more robust, offering a better trade-off between prediction accuracy and variability.

Calibration and Discrimination

Reliability (Murphy & Winkler, 1977), (also: Calibration, e.g. Lichtenstein et al., 1982), refers to the degree of correspondence between forecast probabilities and actual (observed) relative event frequencies. For a predictor to be perfectly calibrated, assessed probability should equal percentage correct where repetitive assessments are being used (ibid.).

Calibration diagrams, built by binning predicted probabilities into 10% bins, are commonly used to portray observed event frequencies against predicted probabilities (e.g. see Murphy & Winkler, 1977). In the left panel of Figure 2 we depict the calibration diagram for our 3 conditions. The dotted straight diagonal line stretching from (0,0) to (100,100) represents the ideal reference of a hypothetical perfectly-calibrated predictor. Evidently, both the human and hybrid markets were reasonably calibrated, while the agents were not. We complement this with the plot on the right panel of Figure 2, which depicts the distribution of predictions

made by the 3 conditions. Humans' risk-aversion is evident, as 75% of their predictions are between 30-70% (50% of the predictions are between 40-60%). For binary events like the ones in our case, whose estimated base rate is near 50%, predictions in the range near 50% are not very informative, and their value for a decision-maker is questionable, as they do not provide much discrimination. For the agents, about 70% of the predictions were in the ranges of 0-30% or 70-100%.

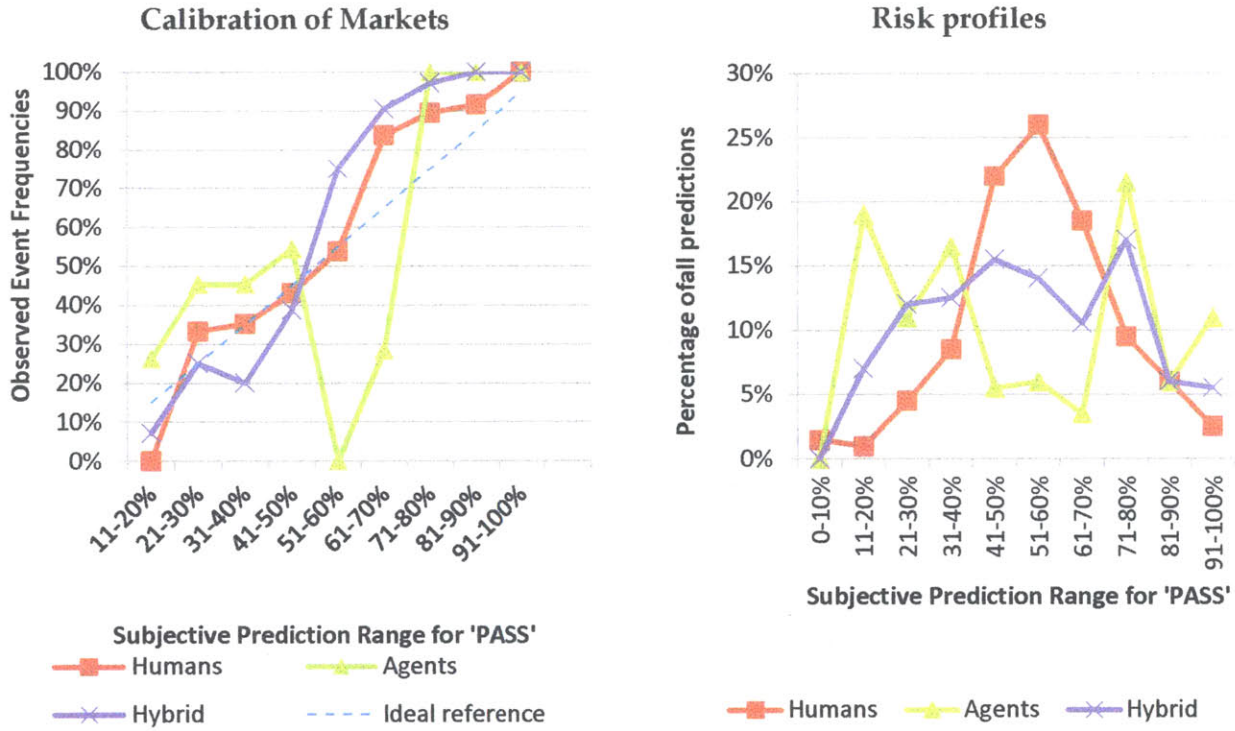
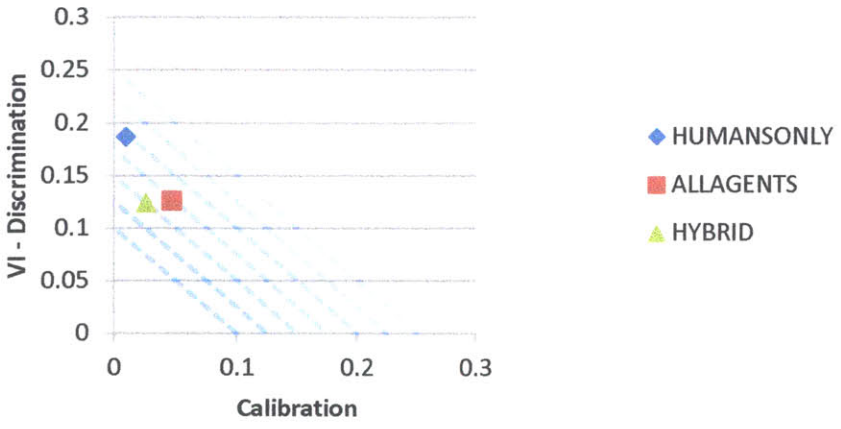


Figure 2

Discrimination (a.k.a Resolution) taps forecasters' ability to do better than a simple predict-the-base-rate strategy. Observers get perfect discrimination scores when they infallibly assign probabilities of 1.0 to things that happen and probabilities of zero to things that do not (Tetlock, 2005, pp. 47-48, 274). It is important to note that calibration skill and discrimination skill are two separate skills. For example, a predictor that always predicts the base-rate of the event will score high on calibration but low on discrimination (for such predictor, the calibration plot will only include a single point, on the diagonal reference

line). It has been offered that the MSE can be decomposed as VI+CI-DI where VI is the variability index representing the uncertainty of the phenomena, CI is the calibration index of the forecasts and DI is the discrimination index of the forecasts (Murphy, 1973; Murphy & Winkler, 1987; see also Tetlock, 2005, pp. 274-275). While the MSE may have drawbacks as a criterion by which to judge the quality of predictions, this decomposition seems nevertheless useful in orienting our understanding of the trade-off between calibration and discrimination of our predictors. Given that the variability of the events in our case is identical for the 3 conditions we want to compare (since they made predictions about the same events) we can draw a plot of (Variability – Discrimination) vs. Calibration for each predictor. For a given variability, we can also draw "efficient front" isopleths of MSE. We present such a plot depicting the performance of our 3 conditions in Figure 3. VI in this study was 0.24. In this plot, the more calibrated a predictor is, the more to the left it would appear (CI closer to zero is better). The more discriminating a predictor is, the lower it would appear. It is evident in this plot that the Hybrid markets were about as discriminating as the agent markets, but more calibrated. It is also clear that compared to human markets, the hybrid markets were slightly less calibrated, but more discriminating. Overall, the hybrid markets are on a more efficient front compared to both agents markets and human markets – as reflected in the MSE scores.

Figure 3 - Variability - Discrimination vs. Calibration (with MSE Isolines. VI=0.24)



While the agents were more accurate than the humans *on average*, their predictions were less calibrated, and they made more severe errors. For any practical matter, they were *utterly* wrong about 4 out of 20 plays (with errors ranging 60-85%; and wrong to a lesser degree on one other play), potentially rendering them untrustworthy for a decision maker (though, of course that depends on the cost of the errors to the decision maker). Humans, on the other hand, had only 3 plays where their prediction (average of 10 markets) was in the wrong direction – but in 2 out of those, their average error was less than 0.55 (and in the third, less than 0.58), conveying their uncertainty to the decision maker by a prediction that was very close to the base rate. Then again, they were also very hesitant (non-discriminating) in most other cases, even when predicting the correct outcome, raising doubt about their value as predictors. The 'hybrid' combination of humans and agents proved to be useful in mitigating both those problems. In terms of accuracy or discrimination, it did not fall far from the agents (in fact, according to the MSE and the LSR criteria, the hybrid markets were more accurate than the agents). In addition, it provides better calibration than the agents, and better discrimination than that of the humans. Importantly, the hybrid groups were on average wrong only about a single play (12), yet their prediction for that play (53.5% Run/46.5% Pass) clearly indicates their lack of confidence in this case to the decision maker.

ROC Analysis

Our comparisons of accuracy, and of the Sharpe ratio, both rely on attaching values to prediction errors using scoring rules. While we used common rules, they may not represent the actual economic value of predictions (or corresponding errors), and in reality, it is not always possible to determine those values. A way of comparing the predictions which does not rely on their unknown economic value can provide additional support for our conclusions.

The Receiver-Operating-Characteristic (also: Relative-Operating-Characteristic; ROC) is an established methodology for evaluating and comparing the performance of diagnostic and prediction systems that has been widely used in many different domains including signal

detection, radiology, weather forecasting, psychology, information retrieval etc. (Swets, 1973, 1988; Swets & Pickett, 1982; Zweig & Campbell, 1993). An ROC curve is obtained by plotting the hit rate (i.e. correctly identified events) versus the false alarm rate (incorrect event predictions) over a range of different thresholds that are used to convert probabilistic forecasts of binary events into deterministic binary forecasts (Jolliffe & Stephenson, 2003, p. 211). For example, a decision maker may set a decision threshold, T. In that case, every prediction that an event is going to occur with a probability estimation $P \geq T$ is considered a positive prediction leading to action, whereas any prediction that the probability of the event to occur is less than T is considered a negative prediction leading to inaction (or, to different action). Under that scenario, small differences in probability estimation may be of low importance. Thus for example, if we set the decision threshold at 50%, we get the result depicted in Table 3.

Table 3 – Results of binary classification of predictions (using a 50% decision threshold)

	Humans	Agents	Hybrid
Number of Cases	200	200	200
Times Correct	139	142	166
Accuracy	69.50%	71.00%	83.00%
Percent of PASS events correctly predicted	76.70%	65.80%	80.00%
Percent of RUN events correctly predicted	58.80%	78.80%	87.50%
Cases of PASS predicted to be a RUN	28	41	24
Cases of RUN predicted to be a PASS	33	17	10

From this table, it appears that the hybrid markets outperformed both other types of markets. But how would this result change if we changed the decision threshold? The ROC, plotted for a range of different thresholds[15], offers a more credible view of the entire spectrum of accuracy of the different predictors (Zweig & Campbell, 1993, pp. 563-564), and

[15] For more details on constructing ROC curves, see Swets (1988, pp. 1286, 1287 and see also endnote 4 on pp. 1291-1292) and others (e.g. Harvey, Hammond, Lusk, & Mross, 1992, pp. 864, 865; Stephenson, 2003)

serves to highlight the tradeoff between sensitivity and specificity of each predictor. The ROC curves[16] of our conditions are presented in Figure 4.

ROC Curves for predictions of football plays by Human-only, Agent-only and Hybrid Prediction Markets (20 plays, 10 observations of each play by each condition)

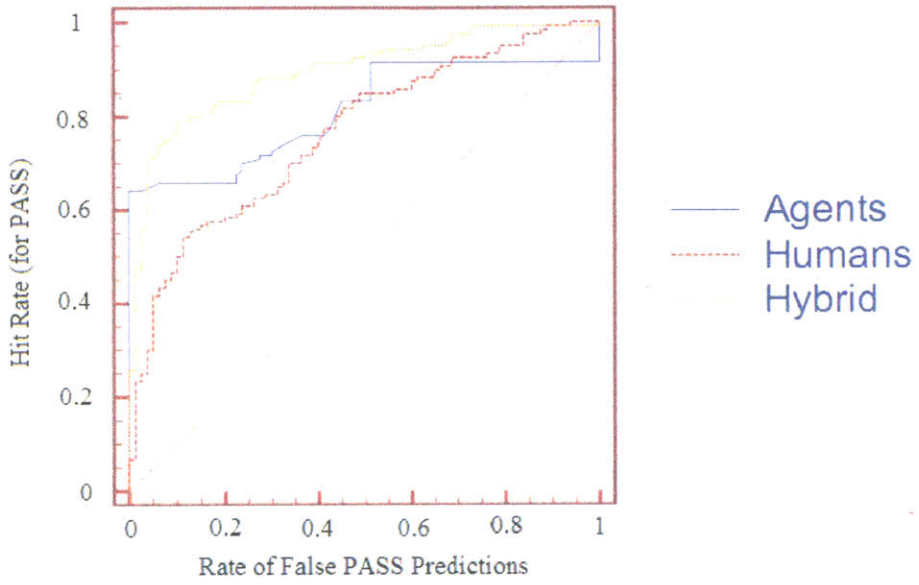


Figure 4

The most common and preferred way to compare the accuracy of predictors using ROC curves is to calculate and compare the area under the curve (usually denoted Az or simply A). The better a predictor is at discriminating events, the closer its curve would be to the (0,1) point. An ideal predictor would be characterized by a curve that goes through that

[16] The construction of ROC curves assumes a reference category. In this case we took "PASS" to be the event of interest and thus, a hit is the correct prediction of a PASS, and a false positive is the prediction of a PASS that turns out to be a RUN. Had we used RUN as the event of interest we would have gotten transformed curves, but they would be equivalent, and the area under the curves would be the same as in the plot we display.

point and the area under its curve would be 1. A non-discriminating predictor would be characterized by a curve laid on the diagonal and an the area A=0.5 (Hanley & McNeil, 1982; Swets, 1988). Calculation of the area A are usually done using software (Swets, 1988, p. 1287). We used MedCalc software17 to calculate the area under the three curves and the results are summarized in Table 4.

	Area under ROC Curve	SE[18], [19]
Humans	0.763	0.0334
Agents	0.813	0.0305
Hybrid	0.895	0.0224

Table 4

This result suggests that the hybrid prediction markets may provide a better trade-off between sensitivity and specificity when compared to either human-only or agent-only prediction markets. In that, it echoes our previous analyses.

Discussion

We compared predictions created by combining individual predictions from multiple humans and artificial-neural-net agents, to those created by collectives of either humans, or agents. We used prediction markets to aggregate the predictions of individuals (humans and/or agents) in all three conditions. We used several different measures and criteria to assess and compare the quality of the predictions, including accuracy (measured using 3 common scoring rules), Sharpe ratios, calibration, discrimination and receiver-operating characteristic plots.

---

[17] (Schoonjans, Zalata, Depuydt, & Comhaire, 1995). MedCalc is available from http://www.medcalc.be/

[18] Standard errors were calculated using the method offered by DeLong, DeLong, & Clarke-Pearson (1988). However they may be inaccurate as we used repeated measurements.

[19] Currently, there is no widely accepted way to test the statistical significance of differences of areas under the ROC curve for repeated measurements of the same events.

The combination of humans and agents proved to be more accurate than either humans or agents according to 2 scoring rules (MSE and LSR). This result holds regardless of the combination mechanism. Under the MAE scoring rule, the accuracy of the hybrid prediction markets was indistinguishable from that of the agent markets, but the agents-only markets were more accurate on average than markets of humans and agents.

The combination of humans and agents provided predictions that were more calibrated than those of the agents, more discriminating than those of the humans and overall providing a better tradeoff of calibration and discrimination compared to the humans or the agents. It also provided the best Sharpe ratios, i.e. the best tradeoff of accuracy and variability of prediction errors. Similar results showing the best tradeoff are also reflected in the ROC analysis, which does not rely on any assumptions about the cost of errors. Overall, therefore, the combination of human and agent predictions in our setting proved more robust, and arguably, superior to either the agents only predictions or the humans only predictions in our setting.

What do these results imply about combining predictions of humans and models or agents in general? Granted, this study, by design, has many limitations that constrain our ability to generalize its conclusions. We use a limited set of events with binary outcomes, from a single domain. Our implementation of the neural-net agents was simplistic, and, admittedly, their trading in the prediction markets was naïve and void of any sophistication. In addition, since cash balance of subjects was carried from one market to the next, success early-on could significantly increase the endowment of the player, giving him/her more choice in strategy, and more influence of the group outcome. Similarly, early failures in predicting could significantly diminish a player's endowment, posing a limit on that player's ability to play, and to influence the group's prediction. And of course, 'regular' limitations of generalizing from lab studies apply. But, as mentioned above, our goal in this study was not to prove or claim in a definitive way that one method is superior. Rather, it was to test a proof of concept of the existence of scenarios where combining predictions from humans and artificial-intelligence agents, in various ways, can outperform those of

either group alone. To that end, our results have supported the hypothesis and provide an existence proof of a situation in which the hypothesis holds true.

This thesis thus contributes to the growing body of knowledge about predictions in the following ways. First, we offer explicitly that *multiple* models *and multiple* human predictions be mechanically combined. Previous work (Blattberg & Hoch, 1990; Bunn & Wright, 1991; Einhorn, 1972; Kleinmuntz, 1990) offering to combine human and model predictions focused more on the differences between humans and models, devoting attention to analyzing their respective strength and weakness points. However, even though some of these works refer to the work on combining forecasts that has developed in parallel, none of them is explicit about the option of combining *multiple* humans with *multiple* models. For example, Blattberg & Hoch *(ibid.)* write: *"Most research has focused on combinations of multiple models or multiple experts, but not model and expert"* (sic!). Thus, although the idea of combining predictions from multiple humans and multiple models can implicitly be deduced from previous literature, to the best of our knowledge it has not been previously explicated, nor has it been empirically tried, and our study therefore should be but a first example (although, admittedly, our agents were too correlated, and our study would have benefitted from the incorporation of additional types of agents).

Second, we propose that using artificial intelligence (e.g. artificial-neural-nets) in such combinations may be beneficial. Artificial-neural-nets offer some advantages over "traditional" statistical models, such as the ability to dynamically adapt the model as new data becomes available (Tam & Kiang, 1992). Empirical examples elsewhere (ibid.) demonstrate the power of artificial-intelligence in inferring rules from large datasets and supporting the case for use of artificial-intelligence to make better predictions than those of traditional methods.

Finally, we offer, and empirically demonstrate, that prediction markets could be an interesting way to mechanically combine predictions from humans and models, providing what we believe to be the first attempt at using them for this purpose. Previous literature

(e.g. Armstrong, 2001a; Bates & Granger, 1969; Blattberg & Hoch, 1990) offers that the simple average should be taken as the default mechanism of combining predictions to increase accuracy, unless another way is found to be better suited in some case. However, these studies did not consider prediction markets. We are not sure whether prediction markets would prove better than a simple average in any case (one recent study from Goel, Reeves, Watts, & Pennock, 2010 suggests that prediction markets may only offer minute improvements over other methods), and it seems reasonable to assume that they are usually more costly to implement. However, prediction markets may be appealing in some settings for reasons other than improvement of the quality of predictions themselves. First, they incentivize participation, of the mindful kind. Second, by increasing attentive participation and by tying compensation to performance while giving participants a sense of both fun and challenge, they serve to increase both extrinsic and intrinsic motivation. They also induce a sense of participation which supports the legitimacy and acceptance of the predictions made. Markets can also be open for people to run their own 'pet' agents, thus potentially incorporating an open innovation pattern into the forecasting process, which may in the long term improve it.

Additional work is required to identify and compare other ways of combining human and machine predictions, and understand their respective advantages and disadvantages in different contexts. For example, in line with recommendations to use computers to aggregate predictions (Einhorn, 1972; Sawyer, 1966), one way of combining human and agent predictions which seems promising is using Adaboost (Freund & Schapire, 1996). Attention should also be given to additional modes of eliciting and expressing human predictions, such as confidence intervals, odds or odds-ratios, as these may have an 'unsuspected role' in forming the collective prediction (Genest & Zidek, 1986). Future work should also examine our approach in more complex domains, and with more sophisticated, domain-specific agents.

Conclusion

We believe that combining predictions from humans and agents may be particularly beneficial in complex scenarios, such as the prediction of actions of human groups, where the rules governing the predicted phenomena are difficult to discern or formulate. In such domains, machine learning can be useful in building sophisticated and adaptive models (for recent examples, see Bohorquez, Gourley, Dixon, Spagat, & Johnson, 2009; Mannes et al., 2008), whereas humans' tacit knowledge, ability to acquire unstructured information, and intuition can help in both information retrieval, and preventing catastrophic prediction errors. Beyond the realm of making predictions, exploring new ways of connecting the knowledge, skill and intelligence residing people's minds with the power of artificial-intelligence may also prove beneficial in other types of tasks performed by individuals, groups and organizations. We hope this initial work will encourage others to further investigate this promising direction.

References

Armstrong, J. S. (2001a). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*: Kluwer Academic Publishers.

Armstrong, J. S. (2001b). JUDGMENTAL BOOTSTRAPPING: INFERRING EXPERTS' RULES FOR FORECASTING. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *OR, 20*(4), 451-468.

Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerial Intuition: 50% Model+ 50% Manager. *Management Science, 36*(8), 887-899.

Bohorquez, J. C., Gourley, S., Dixon, A. R., Spagat, M., & Johnson, N. F. (2009). Common ecology quantifies human insurgency. *Nature, 462*(7275), 911-914.

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

Brown, R. (1986). *Social psychology* (2nd ed.). New York, NY: Free Press.

Bunn, D., & Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: Issues and analysis. *Management Science, 37*(5), 501-518.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Clements, M. P., & Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting, 12*(8).

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Dawes, R. M., & Kagan, J. (1988). *Rational choice in an uncertain world*: Harcourt Brace Jovanovich San Diego.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837-845.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 253-263.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*(1), 86-106.

Ferro, C. A. T. (2007). Comparing Probabilistic Forecasting Systems with the Brier Score. *Weather and Forecasting, 22*(5), 1076-1088.

Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm.*

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science, 1*(1), 114-135.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359-378.

Goel, S., Reeves, D. M., Watts, D. J., & Pennock, D. M. (2010). *Prediction Without Markets.* Paper presented at the 11th ACM conference on Electronic commerce, Cambridge, MA.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*(6), 422-432.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological), 14*(1), 107-114.

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.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Harvey, L., Hammond, K. R., Lusk, C. M., & Mross, E. F. (1992). The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review, 120*(5), 863-883.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting, 21*(1), 15-24.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 4-37.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*: Houghton Mifflin Boston.

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*: The Free Press New York.

Jewson, S. (2004). The problem with the Brier score. *Arxiv preprint physics/0401046*.

Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast verification: a practitioner's guide in atmospheric science*: Wiley.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review, 80*(4), 237-251.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: toward an integrative approach. *Psychological Bulletin, 107*(3), 296-310.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*(1), 111.

Lichtenstein, S., Baruch, F., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*: Cambridge University Press.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (2006). *SAS for mixed models* (2nd ed.): SAS Publishing.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting, 5*(4), 601-603.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, 987-996.

Mallory, B., & Nehlen, D. (2006). *Football offenses & plays*: Human Kinetics Publishers.

Mannes, A., Michael, M., Pate, A., Sliva, A., Subrahmanian, V. S., & Wilkenfeld, J. (2008). Stochastic Opponent Modeling Agents: A Case Study with Hezbollah. In H. Liu, J. J. Salerno & M. J. Young (Eds.), *Social Computing, Behavioral Modeling, and Prediction* (pp. 37-45).

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters, 91*(3), 425-429.

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

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*(4), 595–600.

Murphy, A. H., & Winkler, R. (1987). A general framework for forecast verification. *Monthly Weather Review, 115*(7), 1330-1338.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 41-47.

Pao, Y. (1989). *Adaptive pattern recognition and neural networks*.

Rabin, M. (1996). Psychology and Economics. *Journal of Economic Literature, 36*(1), 11-46.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*(3), 178-200.

Schoonjans, F., Zalata, A., Depuydt, C. E., & Comhaire, F. H. (1995). MedCalc: a new computer program for medical statistics. *Computer Methods and Programs in Biomedicine, 48*(3), 257-262.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of business, 39*(1), 119-138.

Sharpe, W. F. (1994). The sharpe ratio. *Journal of portfolio management* (Fall), 49-58.

Stephenson, D. B. (2003). Glossary of Forecast Verification Terms. In I. T. Jolliffe & D. B. Stephenson (Eds.), *Forecast verification: a practitioner's guide in atmospheric science*: Wiley.

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81-106): Kluwer Academic Publishers.

Sunstein, C. R. (2005). Group Judgments: Statistical Means, Deliberation, and Information Markets. *New York University Law Review, 80*, 962.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182*(4116), 990-1000.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*: Academic Press.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science, 38*(7), 926-947.

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* : Princeton University Press.

Tziralis, G., & Tatsiopoulos, I. (2007). Prediction Markets: An Extended Literature Review. *Journal of Prediction Markets, 1*(1), 75-91.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association, 64*(327), 1073-1078.

Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting, 5*(4), 605-609.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *The Journal of Economic Perspectives, 18*(2), 107-126.

Wolfers, J., & Zitzewitz, E. (2006). Interpreting Prediction Market Prices as Probabilities. *CEPR Discussion Paper No. 5676*.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561-577.