An Explanation-Based approach to Combining Forecasts

Gilat Levy and Ronny Razin[1]

Abstract: We analyze the implications of an explanation-based approach to combining forecasts. This approach, first studied by Pennington (1981), assumes that decision makers interpret multiple sources of information by adopting an explanation which connects and helps them interpret their observed forecasts. We model explanations as joint information structures which are consistent with a set of forecasts the individual observes. In line with the legal notion of balance of probabilities and the philosophical notion of plausible argumentation we assume that the explanation that is adopted is the one that maximizes the likelihood of the observed forecasts. We show that this procedure leads to a simple and dynamically consistent mechanism for aggregating forecasts. The procedure implies that some forecasts are ignored. In particular the individual adopts explanations under which some predictions follow, in a statistical sense, from other more extreme forecasts. Therefore, the procedure provides a rationalization for why extreme forecasts might get more weight when aggregating multiple sources of information.

1 Introduction

When confronted with multiple sources of information, or forecasts, we often have a better understanding of each source separately than we do of how the sources relate to one another. This is apparent in many situations, even when experts make predictions.[2] In this paper we use the explanation-based approach, advocated by Pennington and Hastie (1988), to model how decision makers combine forecasts in such complex environments.

In an influential set of papers based on experiments with jurors' decisions in court cases, Pennginton and Hastie suggest a novel theory of how individuals combine multiple forecasts. In the experiments, jurors were exposed to evidence and testimonies from real court cases. Pennington and Hastie (1988) document a process by which jurors first choose a summary representations of the evidence, which includes a narrative of the causality of significant events and relationships between different pieces of information. As they show, this mental

[1]This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No SEC-C413.

[2]In the finance literature this has been long recognised. Jiang and Tian (2016) point to several problems in estimating correlation, including the lack of enough market data, instabilities in the correlation process and the increasingly interconnected market patterns. The US financial crisis inquiry (FCIC) report from 2011 cites the acknowledgment of the rating agency Moody's that "In the absence of meaningful default data, it is impossible to develop empirical default correlation measures based on actual observations of defaults."

model facilitates evidence comprehension, directs inference and allows to make a decision as well as determine the confidence level about the success of the decision.

Pennington and Hastie (1988) propose an explanation-based theory of decision making. The theory suggests the following procedure of decision making: (i) An individual is faced with the raw evidence pertaining to their decision problem. (ii) The individual entertains different "explanations", i.e., narratives that provide assumptions about causality and relation between the different pieces of information. (iii) The individual chooses one of the explanations based on some criterion. (iv) The chosen explanation is used to interpret the raw evidence to form a prediction about the unknowns behind the decision at hand.

In this paper we provide a model of the explanation-based procedure in the context of a decision maker who is combining forecasts. We consider an environment in which the decision maker observes multiple forecasts, taken to be posteriors about some state of the world. The decision maker then adopts one narrative that explains the forecasts: this is a joint information structure that has generated the forecasts. This explanation allows her to create a posterior about the state of the world which can then facilitate taking a decision.

Pennington and Hastie (1988) suggest several criteria for choosing explanations, such as coherence and goodness of fit. In this paper we focus on the criterion of *balance of probabilities,* an intuitive criterion used in different forms of argumentation. For example, in legal affairs, this is the standard of proof in civil cases, demanding that the case that is the more probable should succeed. While most legal systems agree that the actual truth may never be known, it is prescribed that the court weighs up the evidence and decides which version is most probably true. In line with this practice, in the model we assume that the decision maker chooses the explanation that yields the highest likelihood of observing the evidence. We then analyze the behavioral implications of such a *most plausible explanation* (MPE) procedure for combining forecasts.

We first show that the MPE prediction takes a simple form in which extreme forecasts get prominence while others are ignored. The intuition for why some forecasts are ignored in the final prediction stems from the possibility to construct more plausible theories of the data whereby these forecasts are statistical derivatives of other forecasts. Doing that can increase the likelihood of these theories in explaining the data.

Specifically, it is extreme forecasts that play an important role, while weak forecasts will play no role in the prediction. To see this, note that the more extreme is a forecast (measured by the distance of the forecast to the prior) the lower is the likelihood of any "explanation" of this observation. For example, assume that all forecasts are based on a common prior and some information structure that generates different posteriors. To explain an observation which is in line with the prior, one can find an explanation that happens with probability

one; the forecast is based on someone who knew the prior and didn't receive any signal about the state. Therefore, given that she knew the prior, the probability she will end up with the prior as her posterior is one. On the other hand, to explain a forecast that is different from the prior and is degenerate on one particular state of the world, the highest likelihood you might get is the probability of that state under the prior.

But an explanation has to include all of the observed forecasts, and so the most extreme forecast provide the upper bound on the likelihood of any explanation. As we show in our main result, the decision maker can achieve this upper bound. The decision maker constructs an explanation which first rationalises all the extreme forecasts. The explanation then builds an information structure in which all other forecasts are statistical derivatives of the extreme forecasts. Therefore, the explanation of the decision maker includes a high level of correlation between forecasts which implies that "weaker" forecasts are ignored in the final prediction.

The above characterization of the MPE leads to several important behavioral implications. First, we show that an individual that uses the MPE procedure dynamically changes her prediction in a directional way. In particular, as new forecasts come to light, the individual will respond to them only if they are more extreme than the current set of extreme forecasts she has observed in the past. In a dynamic framework, this also implies *stagnation*; even when new forecasts arrive over time, the prediction of the decision maker may remain the same. This is true even for repeated forecasts; they have no effect on the prediction.

We study the dynamic properties of the MPE procedure. First, the procedure is simple to compute and involves low memory constraints. In particular, the discussion above implies that the decision maker has to remember only the set of the current extreme forecasts. As we show in the paper, the number of such forecasts at any time is bounded by the number of states in the state space.

We show that the MPE procedure is time consistent. In particular, we ask whether the decision maker can maintain a coherent explanation across time or does she have to entertain a completely new explanation when she confronts a new forecast. We show that the MPE explanation evolves in a way that new forecasts are embedded into the previous explanations. Finally, it satisfies path independence. At any period, the explanation and hence prediction are the same no matter what is the order in which forecasts arrived prior to that period.

We consider the implications of our results to social learning, when each forecast is observed by a different decision maker. In this case a decision maker at time t only observes the MPE predictions of those before her and not their forecasts. We show that if all the history of predictions is observed, then this is equivalent to a single decision maker observing all forecasts. However, with limited observability of previous predictions, path dependency may arise. Moreover, stagnation in predictions is less likely to arise in the limited observability

case.

Finally we discuss how the MPE procedure relates to biases in group decision making explored in the literature. In the context of group decision making, when each individual provides a forecast, the group MPE prediction will be biased towards extreme ones.[3] We shed light on the puzzling Talmudic law that requires that if judges (in the Sanhedrin court) are unanimous in conviction, the defendant should be set free. As in Glatt (2013), who uses a maximum likelihood approach, MPE implies that unanimity among many judges is most likely is a result of strong correlation between the judges, and therefore demands caution.

There is by now a growing literature in economics on how individuals and organizations combine forecasts in a non-Bayesian manner, specifically when individuals have a good understanding of each information source separately but are not sure of how the sources relate to one another. In Levy and Razin (2017) we consider decision makers who perceive all possible consistent explanations with a bounded levels of correlation. We introduce the notion of "correlation capacity" which implies a restricted pointwise mutual information, and show that the set of predictions is completely characterised by the Naïve-Bayes heuristic, which completely neglects correlation, and the degree of this correlation capacity. To derive behavioural implications we consider preferences over the ambiguous set of explanations. In contrast, our analysis here allows for all degrees of correlation and indeed decision makers who use MPE predictions consider models with full correlation. While not Bayesian, the decision maker considers all joint information structures that are consistent with Bayesian updating for the forecasts she observes, and chooses the one that is most plausible given the forecasts.

At what seems to be the other extreme, which arises when there is no correlation capacity, decision makers have correlation neglect, as for example in DeMarzo et al (2003), Glaeser and Sunstein (2009), Levy and Razin (2015a, 2015b) and Golub and Jackson (2012).[4] This literature either uses the Naïve-Bayes approach so that individuals believe that sources are (conditionally) independent, or the DeGroot heuristic (when applied to normal distributions). Ellis and Piccione (2017) consider an axiomatic approach to decision makers who ignore correlation. The literature shows that polarisation or extreme beliefs are likely to arise when individuals neglect correlation. As we show, MPE predictions which arise from decision makers who consider large degrees of correlation can also lead to polarisation and

[3] See Schkade et al (2000) for experiments that show how group decisions are biased towards extreme views. Ambrus et al (2015) on the other hand show in an experiments that it is moderate members that are more persuasive.

[4] See also Ortoleva and Snowberg (2015), Enke and Zimmerman (2013), Kallir and Sonsino (2009) and Eyster and Weiszacker (2011), who provide experimental and empirical evidence about correlation neglect.

prominence to extreme forecasts.[5]

The MPE prediction uses maximum likelihood, a key notion in statistics, econometrics and machine learning since Fisher (1912). Our analysis in this paper differs from maximum likelihood procedures used by econometricians or in machine learning;[6] in these environments, the econometricians knows (or learns) the family of joint information structures which generate the data. Thus, by the law of large numbers, estimators of correlation parameters will be efficient, as if the decision maker is also Bayesian. We analyze such an example in Section 5. However, in the main model, we focus on a dynamic model in which every period the decision maker encounters another forecast, and she considers all possible explanations without restricting to any functional form. Thus at period t, she considers t forecasts and a joint information structure generating t forecasts which is analogous to one observation in the data set of the econometrician. Thus such a procedure is not likely to lead to the true information structure. Note that while typically maximum likelihood explanations are considered to be over-fitting data, in our case we look for explanations that are consistent with rationality of forecasters, so that such over-fitting will not typically arise.

In decision theory, Gilboa and Schmeidler (2003) identify axioms that can rationalize maximum likelihood procedures while Gilboa and Schmeidler (2010) highlight the trade-off that can arise with a very complex theory providing an extreme likelihood of 1, and a simple explanation. We show that for our framework, the maximum likelihood criterion of choosing an explanation has a simple characterisation which also induces dynamic consistency and low memory requirements. Ortoleva (2012) provides axioms that rationalise how individuals change their prior when they encounter zero probability events, where he shows that such individuals can use maximum likelihood in order to switch between priors.

Our analysis also relates to the literature on opinion pooling (Dietrich and List 2016, 2017) and combining probability distributions (Genest and Zidek 1989). In particular, this literature has taken an axiomatic approach to the aggregation of profiles of probability distributions into one probability distribution. We can show that the version of independence considered in this literature is violated by the MPE procedure.

2 The Model

A decision maker is forming a prediction about a state of the world $\omega \in \Omega$. For expositional purposes in the main body of the paper we assume a binary state of the world so that

[5] Sobel (2014) shows that polarisation can also arise in a purely Bayesian model.

[6] See Mullainathan and Waggoner (2017) for a survey of the use of maximum likelihood in machine learning.

$\Omega = \{0, 1\}$. All our results extend to the general case, see Section 6.[7] The decision maker has a prior $p = \Pr(\omega = 1) \in (0, 1)$.

At every period $t \in \{1, 2, ..., T\}$, the decision maker observes a forecast. A forecast t is a (full support) probability distribution over Ω. Let then q^t denote the probability that the state is 1 according to forecast t. The decision maker knows that for any t the forecast q^t is the posterior derived using Bayes rule from the prior p, when a forecaster observed some informative signal about the state. The distribution of the signal that led to forecast q^t is unknown to the decision maker.

Let $Q^t = \{q^1, q^2...q^t\}$ denote the history of forecasts up to period $t \in \{1, 2, ..., T\}$, for $T \leq \infty$. At any period t, the agent's observed history is Q^t. We will discuss later on other possibilities for example when the decision maker forgets forecasts but remembers her predictions (which, as we show, makes no difference), or when the decision maker does not even remember all her predictions.

At every t, the decision maker will combine these forecasts and the prior into a prediction about the state. A prediction about the state is a probability distribution η over Ω. To form a prediction, the decision maker will consider *consistent explanations*, which are Bayesian models that are consistent with Q^t and the prior.

Note that the decision maker only observes forecasts here; our approach can be generalized to the case in which the decision maker observes the forecasts and the marginal distributions generating each forecast (which is equivalent to observing also the signal of each information source providing the forecasts).

We now formally define what an explanation is. An explanation is an information structure, that is, a tuple $\mathcal{I} = (S, \Omega, p, q(\mathbf{s}, \omega))$, such that $S = \times_{j=1}^{t} S^j$ is a set of signals where S^j is finite and for $\mathbf{s} \in S$, $q(\mathbf{s}, \omega)$ is a joint probability distribution over signals and states. Given $q(\mathbf{s}, \omega)$, we have $q(\mathbf{s}|\omega)$, $q(\omega|\mathbf{s})$ and $q(\mathbf{s})$ that are defined accordingly. Moreover, $q(\mathbf{s}, \omega)$ also implies the marginal distributions on signal s^j, $q^j(s^j|\omega)$.

An *explanation* of the data implies that prediction q^j in Q^t was generated by a Bayesian forecaster who was exposed only to the information contained by the marginal distribution of signals in S^j, $q^j(s^j|\omega)$.

A *consistent explanation* of Q^t will be a rationalisation of the observations in Q^t according to an information structure. Formally,

Definition 1 *A consistent explanation of Q^t is a couple, $(\mathcal{I}, \mathbf{s}^*)$ where $\mathcal{I}$ is a joint information structure and $\mathbf{s}^* = (s^{*,j})_{j=1}^{t} \in S$ such that for all $j \in \{1, 2, .., t\}$, $q^j = \Pr_{\mathcal{I}}(\omega = 1|s^{*,j}) = \frac{pq^j(1|s^{*,j})}{pq^j(1|s^{*,j}) + (1-p)q^j(0|s^{*,j})}$.*

[7] All proofs in the Appendix are for the general case.

In other words, the agent perceives some information structure $\mathcal{I}$, and a particular realisation of signals $\mathbf{s}^*$, such that given the prior, the marginal of the joint information structure $q^j(s^j|\omega)$ derived from the joint distribution $q(\mathbf{s}|\omega)$, and the relevant element of $\mathbf{s}^*$, $s^{*,j}$, all the forecasts q^j can be rationalised by Bayes rule, assuming that each forecast q^j was based only on the signals generated by $q^j(s^j|\omega)$. Let the set of consistent models of Q^t be given by C^{Q^t}.

In this paper we assume that the decision maker uses the maximum likelihood criterion to select which consistent explanation to adopt to set her prediction.[8] To formalise this, we first define the likelihood function of observing Q^t given the consistent model $(\mathcal{I}, \mathbf{s}^*)$:

$$L(Q^t|(\mathcal{I}, \mathbf{s}^*)) = \sum_{\omega_i \in \Omega} p(\omega_i) q(\mathbf{s}^*|\omega_i).$$

We are now ready to formalise the behavioural assumption about how the decision maker forms a prediction given an observation of Q^t :

Assumption A1 (Most Plausible Explanation-based prediction): At any period t, the decision maker forms a prediction about Ω according to the following procedure: (i) She chooses a consistent explanation, $(\mathcal{I}, \mathbf{s}^*)$, in $\arg\max_{(\mathcal{I}', \mathbf{s}'^*) \in C^{Q^t}} L(Q^t|(\mathcal{I}', \mathbf{s}'^*))$. (ii) Her prediction is given by $\eta_{Q^t}^{MPE}(\omega) = q_{|(\mathcal{I}, \mathbf{s}^*)}(\omega|\mathbf{s}^*)$.

3 The MPE prediction and ignored forecasts

We now characterize the MPE prediction for the observed history Q^t for $t \in \{1, ..., T\}$. We then use the properties of the characterisation to discuss the features of the simple dynamic algorithm that allows us to generate predictions over time, as well as the positive properties of the observed sequence of predictions.

3.1 The MPE prediction

Consider the t forecasts. Let $q^{t,\max} = \arg\max_{j \in \{1,2,..,t\}} q^j$ and let $q^{t,\min} = \arg\min_{j \in \{1,2,..,t\}} q^j$.

Proposition 1: For any Q^t, the MPE prediction of the decision maker is well-defined,

[8] Pennington and Hastie (1988) discuss other criteria relating to coverage, coherence and goodness of fit. In Levy and Razin (2017) we consider all explanations and use preferences over ambiguity to derive behavioural implications.

unique and satisfies:

$$\eta_{Q^t}^{MPE}(1) = \begin{cases} q^{t,\max} \text{ if } q^{t,\min} \geqslant p \\ q^{t,\min} \text{ if } q^{t,\max} \leqslant p \\ \frac{q^{t,\min} q^{t,\max}}{q^{t,\min} q^{t,\max} + \frac{p}{(1-p)}(1-q^{t,\min})(1-q^{t,\max})} \end{cases}$$

Proof of Proposition 1: First note that for any information structure, for any $\omega \in \Omega$, $q(\mathbf{s}^*|\omega) \leq \min_{j=1,\ldots,t} q^j(s^{*,j}|\omega)$, to satisfy the definition of a joint distribution function. Second note that by the consistency of the explanation (Definition 1), for any j we must have:

$$\frac{q^j}{1-q^j} = \frac{p}{1-p}\frac{q^j(s^{*,j}|1)}{q^j(s^{*,j}|0)}$$

This implies that by setting $q^j(s^{*,j}|1)$ at some level, we pin down $q^j(s^{*,j}|0)$. Using this fact and the inequality above, we can write the upper bound for the likelihood as:

$$\begin{aligned} \sum_{\omega_i \in \Omega} p_i q(\mathbf{s}^*|\omega_i) &\leq p \min_j\{q^j(s^{*,j}|1)\} + (1-p)\min_j\{q^j(s^{*,j}|0)\} \\ &= p \min_j\{q^j(s^{*,j}|1)\} + (1-p)\min_j\{\frac{1-q^j}{q^j}\frac{p}{1-p}q^j(s^{*,j}|1)\} \\ &= p[\min_j\{q^j(s^{*,j}|1)\} + \min_j\{\frac{1-q^j}{q^j}q^j(s^{*,j}|1)\}] \end{aligned}$$

We will now maximise the right hand side and show that we can achieve a likelihood equal to the maximal upper bound. Note that the problem is separable across the forecasts, and that increasing $q^j(s^{*,j}|1)$, for each forecast j, increases all the values for forecasts.

Note that as $q^j(s^j|0) = \frac{1-q^j}{q^j}\frac{p}{1-p}q^j(s^{*,j}|1) \leq 1$, this implies an upper bound on $q^j(s^{*,j}|1)$. Therefore we can set $q^j(s^{*,j}|1)$ at the upper bound,

$$\begin{aligned} q^j(s^{*,j}|1) &= \min\{1, \frac{1-p}{p}\frac{q^j}{1-q^j}\} \equiv \gamma^j \Rightarrow \\ q^j(s^{*,j}|0) &= \min\{\frac{1-q^j}{q^j}\frac{p}{1-p}, 1\} \equiv \eta^j \end{aligned}$$

Given that we do this for each j, going back to the original problem we can set:

$$\begin{aligned} q^{MPE}(\mathbf{s}^*|1) &= \min_j \gamma^j, \\ q^j(\mathbf{s}^*|0) &= \min_j \eta^j. \end{aligned}$$

And as a result,

$$\frac{\eta^{MPE}(1)}{\eta^{MPE}(0)} = \frac{p}{1-p}\frac{q^{MPE}(\mathbf{s}^*|1)}{q^{MPE}(\mathbf{s}^*|0)} = \frac{p}{1-p}\frac{\min_j \gamma^j}{\min_j \eta^j}.$$

This implies the expression in the Proposition.■

Intuitively, the most extreme forecast has the lowest ex ante likelihood; if the forecast is very different from the prior and for example is degenerate on one particular state of the world, its likelihood is at most the probability of that state under the prior. As the maximum likelihood for the joint set of forecast is bounded by the likelihood of each individual one, the likelihood of the most extreme forecast provide the upper bound on the likelihood of any explanation. The proof shows that the decision maker can achieve this upper bound by rationalising all the extreme forecasts and then creating an information structure in which all other forecasts are statistical derivatives of the extreme forecasts. Therefore, the explanation of the decision maker includes a high level of correlation between forecasts which implies that all forecasts other than $q^{t,\max}$ and $q^{t,\min}$ are always ignored.[9] This is also true in the general, non-binary case, and will induce important dynamic implications that we will explore below.

Before doing so, let us illustrate the result further by constructing the MPE itself, i.e., the joint information structure which has the highest likelihood of generating the forecasts. Let us consider two forecasts, so that $t = 2$. With out loss of generality we consider information structures with two signals, s^* and s^{-*}. Assume that $q^2 \geq q^1$. Assume wlog that $\frac{1-p}{p}\frac{q^2}{1-q^2} > 1$. When also $\frac{1-p}{p}\frac{q^1}{1-q^1} > 1$, then the unique MPE explanation is:

$\omega = 0$	s_1^*	s_1^{-*}	$\omega = 1$	s_1^*	s_1^{-*}
s_2^*	$\frac{1-q^2}{q^2}\frac{p}{1-p}$	0	s_2^*	1	0
s_2^{-*}	$\frac{1-q^1}{q^1}\frac{p}{1-p} - \frac{1-q^2}{q^2}\frac{p}{1-p}$	$1 - \frac{1-q^1}{q^1}\frac{p}{1-p}$	s_2^{-*}	0	0

.

In the table above, each cell in the matrix represents the joint probability over receiving the two signals by the two forecasters.

Note that forecast 1's signal realisation s_1^* always follows when forecast 2's signal s_2^* is generated. As a result, the signal realisation of forecast 2, s_2^*, is a sufficient statistic for the signal realisation of forecast 1, s_1^*, and therefore forecast 1 is ignored. The information structure exemplifies why moderate forecasts are sometimes ignored: they are considered to be a weaker signal, correlated with the strong and extreme forecast.

On the other hand, when $\frac{1-p}{p}\frac{q^2}{1-q^2} > 1 > \frac{1-p}{p}\frac{q^1}{1-q^1}$, then the information structure that achieves maximum likelihood is

[9] The result here is that the MPE prediction either identifies with one of the extreme forecasts or is a compromise. When we extend to more than two state we can also have MPE beliefs that are more extreme than the extreme forecasts (and so outside the convex hull of the forecasts). Still, it is only the extreme forecasts that matter. Also, the compromise in the simple binary case treats the two extremum forecasts as if they were independent; this will not generalize for the case of non-binary realisations of the state. See Section 6.

$$\begin{array}{llll|llll} \omega=0 & s^* & & s^{-*} & \omega=1 & s^* & & s^{-*} \\ s^* & \frac{1-q^2}{q^2}\frac{p}{1-p} & & 0 & s^* & \frac{1-p}{p}\frac{q^1}{1-q^1} & & 1-\frac{1-p}{p}\frac{q^1}{1-q^1} \\ s^{-*} & 1-\frac{1-q^1}{q^1}\frac{p}{1-p} & & 0 & s^{-*} & 0 & & 0 \end{array} .$$

Now the signal of forecast 2 is not a sufficient statistic for the one of forecast 1 and forecast 1 is indeed taken into consideration. Both information structures satisfy full correlation in the sense that the matrices above have zero entries.

Remark 1 *(A sufficient statistic forecast is not necessarily relevant):* Note that when a forecast is ignored, then we can generate an information structure in which the other forecast is a sufficient statistic for it. However, the opposite is not true. For two predictions, we can generate an information structure in which one is a sufficient statistic of the other, but this does not imply that the MPE prediction will ignore the latter. For example, take the information structure:

$$\begin{array}{llll|llll} \omega=0 & s_1^* & & s_1^{-*} & \omega=1 & s_1^* & & s_1^{-*} \\ s_2^* & \frac{1-q^2}{q^2}\frac{p}{1-p} & & \frac{1-q^1}{q^1}\frac{p}{1-p}-\frac{1-q^2}{q^2}\frac{p}{1-p} & s_2^* & 1 & & 0 \\ s_2^{-*} & 0 & & 1-\frac{1-q^1}{q^1}\frac{p}{1-p} & s_2^{-*} & 0 & & 0 \end{array} .$$

Now the $*$ signal of forecaster 1 is a sufficient statistic for the $*$ signal of forecaster 2. However this is not the MPE.

The key feature of the MPE prediction is that it demands a high degree of correlation between the signals generating Q^t. This can be seen from the fact that some forecasts in Q^t may actually be ignored; the important or decisive set of forecasts will be the two most extreme beliefs. This, together with the fact that this belief is based on a Bayesian rationalisation implies that it is based on a joint information structure which involves a high degree of correlation between the signals that have generated Q^t. If, in contrast, these signals were assumed to be conditionally independent, the likelihood ratios would have been responsive to all beliefs in Q^t. Thus, the MPE decision maker "looks for" correlations.

3.2 Dynamics of MPE predictions: stagnation and simplicity

We first present two behavioural implications of the dynamic MPE predictions: path independence and stagnation. We also discuss the dynamic properties of the process itself; we show that the MPE is in itself consistent through time, and that the decision maker needs only a bounded memory to reach optimal MPE predictions, as well as use a simple explanation. All the results generalize to the case of non-binary state of the world.

3.2.1 Path independence and stagnation

An immediate derivative of the above results is the invariance of the procedure to the order of forecasts:

Corollary 1: *(Path independence) At period t, the MPE prediction is the same for any permutation of Q^t.*

Another observation relates to *stagnation*. Assume just for illustration equal priors, i.e., $p = \frac{1}{2}$. Imagine the decision maker observing $q^1 > \frac{1}{2}$ and then $q^2 > \frac{1}{2}$, with $q^1 > q^2$. The decision maker will not change her mind in period 2, and $\eta^{1,MPE} = \eta^{2,MPE} = q^1$. Similarly, if forecasts are repeated, such repetition does not affect the MPE prediction. For example, completely replicating the set of forecasts will result in the same prediction. Stagnation will arise in this case as well. Formally, we say that stagnation arises in period t when $\eta^t = \eta^{t-1}$.

Corollary 2: *(Stagnation and no effect for repetition): (i) For any history Q^t, $\eta^{t,MPE}(Q^t) = \eta^{t+1,MPE}(Q^t, \hat{q})$ for any $\hat{q} \in [\min\{p, q^{t,\min}\}, \max\{q^{t,\max}, p\}]$. (ii) Consider the history $Q \subseteq Q^t$ and replicate it k times so that $Q^{t+k|Q|} = (Q^t, Q, Q, ..., Q)$. Then $\eta^{MPE}_{Q^t} = \eta^{MPE}_{Q^{t+k|Q|}}$.*

Another way to describe the particular dynamics of stagnation is noting that MPE leads to *directional updating*. This can be seen again in the example with two forecasts: When $q^1 > \frac{1}{2}$ then the decision maker strongly responds to $q^2 > q^1$, as it is more extreme and points in the same direction away from the prior as initial belief, and does not respond at all to weaker beliefs in that direction, when $q^2 \in [\frac{1}{2}, q^1]$. When confronted with forecasts pointing in the other direction from the prior, $q^2 < \frac{1}{2}$, the decision maker does respond to this, but only partially.

Therefore, although we have not assumed any auxiliary reason to favour one direction to another, the MPE prediction produces such a bias. Weaker forecasts (pointing to the same direction away from the prior) can be assumed to follow directly from the signals that have produced the stronger forecast, which allows the decision maker to build consistent explanations with a high correlation and high likelihood.

3.2.2 Simplicity of the dynamics of the MPE

We now highlight "efficient" features of the dynamic MPE prediction: the dynamic process is time-consistent so one can "extend" the current explanation to accommodate a new forecast, and moreover the process demands only a low memory capacity. Finally, we show that it is sufficient to consider information structures with two signals for each forecaster. Together these results illustrate the simplicity of the process.

Let us consider time-consistency first: We show that when the decision maker observes a new forecast, she can accommodate it within the (adjusted) previous explanation rather than consider a whole new explanation.

The following definition will be useful to formulate time-consistency across explanations.

Definition 2: *For any consistent explanation* $(\mathcal{I}, s^*)$ *of* Q^{t+1}*, let the marginal of* $(\mathcal{I}, s^*)$ *on* Q^t*,* $(\mathcal{I}_{|Q^t}, s^*_{|Q^t})$*, be given by* $\mathcal{I}_{|Q^t} = (S_{|Q^t}, \Omega, p(.), q_{|Q^t}(s,\omega))$ *where* $S_{|Q^t} = S = \times_{j \text{ such that } q^j \in Q^t} S^j$ *and* $q_{|Q^t}(s,\omega)$ *is the projection of* $q(s,\omega)$ *on the predictions in* Q^t*, and* $s^*_{|Q^t} = (s^{*,j})_{j \in Q^t}$*.*

Proposition 2: *For any* Q^t *and additional forecast* q*, there exists a consistent explanation* $(I, s^*) \in \arg\max_{(\mathcal{I}', \mathbf{s}'^*) \in C^{Q^{t+1}}} L(Q^{t+1}|(I', s'^*))$ *such that* $(I_{|Q^t}, s^*_{|Q^t}) \in \arg\max_{(\mathcal{I}', \mathbf{s}'^*) \in C^{Q^t}} L(Q^t|(I', s'^*))$.

The dynamic MPE prediction also has an attractive simplicity characteristic: As at each point in time only the extreme forecasts matters, it is enough for the decision maker to remember across periods at most the two most extreme forecasts she had received. More generally, whenever a forecast is ignored at one point, it can be ignored forever and can be permanently forgotten. This is a good property of the procedure in particular when the data gets large:

Corollary 3 *To generate the same sequence of MPE predictions as a decision maker who always knows all the forecasts, a decision maker needs to carry over to the next period at most* 2 *forecasts.*

Finally, another feature of the MPE prediction is that the explanations involved -namely the joint information structures- are themselves simple. In the previous subsection we have generated the MPE information structure with only two signals. This is in fact a general sufficient feature which is technically helpful; while the set of consistent explanations the decision maker considers is very large, without loss of generality we can focus only on information structures with $\#|S^j| = 2$ for any $j = 1, ..., m$:

Lemma 1 *For any* $(\mathcal{I}', \mathbf{s}'^*) \in C^{Q^t}$ *there exists another consistent explanation* $(\mathcal{I}, \mathbf{s}^*) \in C^{Q^t}$ *such that* $S = \{s^*, s^{-*}\}^m$ and for which (i) $\hat{q}(\omega_i|\mathbf{s}^*, \mathcal{I}) = \hat{q}(\omega_i|\mathbf{s}^*, \mathcal{I}') = \eta^{ML}_{Q^t}(\omega)$ and (ii) $L(Q^t|(\mathcal{I}, \mathbf{s}^*)) = L(Q^t|(\mathcal{I}', \mathbf{s}'^*))$.

4 Source amnesia and social learning

We have so far assumed that the decision maker remembers forecasts; she either remembers all forecasts, or as we saw, the two most extreme forecasts (the "decisive" set of forecasts). We now consider an extension to a decision maker that does not remember forecasts but

predictions. This is important to consider as predictions which possibly turn into actions are typically easier to remember as these have consequences. It is therefore reasonable to assume that the decision maker remembers her predictions but not how she reached them. This feature is sometimes referred to as *source amnesia*.

We consider first the case in which she remembers all her predictions, i.e., *full observability*, and then the case of *limited observability*, where she remembers only her last prediction. We then compare between the two.

4.1 Full observability: equivalence result

In our previous model, at time t, the decision maker observes $Q^t = (q^1, q^2, ..., q^t)$. Assume an equal prior for simplicity. We now assume instead that at time t, a decision maker observes $\Upsilon^t = (\eta^1, \eta^2, ..., \eta^{t-1}, q^t)$, that is, previous η^{MPE} predictions, and a new forecast.

When all history of play is observable, q^1 is the forecast and action at Period 1, $\eta^2 = \eta^{MPE}(q^2, q^1)$ is the observed action at Period 2, $\eta^3 = \eta^{MPE}(q^3, \eta^2, q^1)$ is the observed action and belief at Period 3, and so on.

Suppose that $q^1 > \frac{1}{2}$. Let us consider Period 3, where the decision maker observes q^3 and η^2, q^1. She knows that for whichever q^2 she imagines, it has to be that $\eta^2 = \eta^{MPE}(q^2, q^1)$. By Proposition 1, if $\eta^2 \geq q^1$, then $\eta^2 = q^2$. If $\eta^2 = q^1$, then $q^1 > q^2 > \frac{1}{2}$ and it is not important for her to know the exact value of q^2, so one can assume $q^2 = q^1$. If $\eta^2 < q^1$, then $q^2 = \frac{\eta^2(1-q^1)}{q^1+\eta^2(1-2q^1)}$. Thus, at Period 3 (and hence at all subsequent periods) she can extract $\min_{1,2}\{\frac{q^1}{1-q^1}, \frac{q^2}{1-q^2}\}$ from her previous behaviour. It follows then that at Period t she can understand $\min_i\{\frac{q^i}{1-q^i}\}$ for $i < t$ for any i. We therefore have:

Proposition 3: *When the decision maker remembers only her previous predictions (and all of them) but not forecasts, at any period t the MPE prediction is the same as in the model in which she observes Q^t.*

In the previous Section we have shown that if the decision maker remembers forecasts, then it is sufficient to remember only the most extreme ones. The result here also exemplifies the simple memory requirements needed by the MPE procedure. Proposition 3 states that it is also equivalent to remember all predictions (in the order they have been derived). However, when one remembers predictions and not forecasts, it is not sufficient to have a limited observability or memory, as we now illustrate.

4.2 Limited observability: path dependence and less stagnation

We now consider a decision maker who has limited observability of predictions. For simplicity, let us assume that she remembers her latest prediction, η^{t-1}, and learns the forecast q^t, when making her prediction at time t. Imagine an individual that receives two pieces of information, q^1 and q^2, and delivers a combined forecast η^2. At this point, the individual forgets *how* she reached η^2, that is, the exact q^1 and q^2, only remembers η^2 as her current beliefs and only carries this through to the next period until she needs to act again. Thus in period 3 she knows q^3 and η^2.[10]

We saw above that when decision makers observe the full sequence of forecasts, the sequence of moves did not affect the prediction. However, this is not the case when decision makers only observes the previous prediction. To consider an example, let $q > \frac{1}{2}$, and $\hat{q} < \frac{1}{2}$. If $q^1 = q, q^2 = \hat{q}$, $q^3 = q$, then $\eta^2 = \frac{\frac{\hat{q}}{1-\hat{q}}}{\frac{\hat{q}}{1-\hat{q}}+\frac{1-q}{q}}$. Suppose that $\hat{q} > 1-q$, and thus $\eta^2 \in (\frac{1}{2}, q)$. We then have at period 3 the combined forecast of η^2 and q which is $\eta^3 = \eta^{MPE}(\eta^2, q) = q$. However, if $q^1 = q, q^2 = q$, $q^3 = \hat{q}$, then $\eta^2 = q$ and $\eta^3 = \eta^{MPE}(\hat{q}, q) = \frac{\frac{\hat{q}}{1-\hat{q}}}{\frac{\hat{q}}{1-\hat{q}}+\frac{1-q}{q}} < q$. Thus, history matters and specifically, a lower belief $\hat{q}$ will have a greater effect the later it is in the sequence.

The pattern of the example above is more general: at any period t, beliefs are more likely to make a difference when arriving at the end rather than earlier in the sequence:

Corollary 4: *Consider a sequence of T beliefs composed of $t > 1$ forecasts $q > \frac{1}{2}$ and $T - t > 1$ forecasts q' satisfying $1 - q < q' < \frac{1}{2}$. The final prediction will be q if $(q^{T-1}, q^T) = (q, q)$ and will be lower than q if $q^T = q'$.*

We have derived stagnation in the basic model in which the decision maker observes forecasts. From Proposition 3 we know that this will arise also in the case in which she observes the set of predictions, as these cases are equivalent. As we now show, the instances of stagnation are lower in the case of limited memory of forecasts. This is intuitive as limited previous predictions cannot carry with them the intensity with which they were derived and thus the decision maker will underestimate the strength of the last predictions.

Proposition 4: *(ii) For any sequence of forecasts $\{q^t\}_{t=1}^{\infty}$, for any period t, if under the full observability of predictions model there is no stagnation in period t then there is no stagnation in the limited observability model. (iii) There exists sequences of forecasts $\{q^t\}_{t=1}^{\infty}$ and periods t such that there is stagnation under the full observability model and no stagnation under the limited observability model.*

[10] The decision maker may or may not realize that η^2 was based on two forecasts; we show in the Appendix (Lemma 2) that this is not important. That is, the new prediction will be the same if the decision maker mistakenly believes that $\eta^2 = q^2$, or if she believes that it was derived based on q^1 and q^2.

4.3 Social learning

Social learning is typically modelled as an environment in which an individual at period t observes the sequence of the actions or predictions of her predecessors, as well as her own forecast.[11] Thus as above, she will observe a segment of $\Upsilon^t = \{\eta^1, \eta^2, \eta^3 ... \eta^{t-1}, q^t\}$.

The key difference between a social learning environment and a single decision maker is that in the former we have to assume what each individual thinks about how others combine forecasts.

Assume then that all players use MPE predictions and that all players know that others do so as well. Formally, we define this recursively. Player 2 needs to rationalize q^1, q^2 with an MPE as in Definition 1 and Assumption 1. Player 3 needs to rationalize q^1, q^2, q^3 with an MPE subject to Player 2 rationalizing q^1, q^2 with an MPE satisfying $\eta^{MPE}(q^2, q^1) = \eta^2$. Thus, a player k needs to rationalize some $q^1, q^2, q^3, ..., q^{k-1}, q^k$, by having an MPE with a set of signals $s^1, ..., s^k$, so that $q^i = \hat{q}(\omega|s^i)$ for each player $i \leq k$, and for each player $i < k$, there exists an MPE so that $\eta^i = \eta^{MPE}(q^1, q^2, ..q^i)$.

Given the above, all the results above hold. When all previous predictions are observed, it is as if forecasts are shared as well. When predictions are not fully observed, then path dependence will arise, with latter forecasts being more important, as Corollary 4 indicates.

Within the literature on social learning, our results then show that even in a model with continuous actions and beliefs, stagnation might arise, so that individuals may stick to the same action even when they receive more and more different forecasts. Still, stagnation may be temporary and the prediction or action may change at a latter point. This differs from the results in the current literature (see for example Eyster and Rabin 2010, Smith and Sorensen 2001).

5 A homogenous pool model: non monotonicity of decision rules

In our analysis above we have assumed a general set of possible explanations. In some applications there are some restrictions imposed on the set of possible explanations. One reason to put restrictions on the set of possible explanations is tractability. Poll aggregation analysts for example use specific families of correlations structures (copulas) to model different scenarios of correlation between different polls in different constituencies. Or it can be

[11] In the standard social learning literature an individual observes a signal and knows her marginal probability distribution generating the signal conditional on the state of the world. As long as she does not know the correlation between the signals, our model in which individuals only observe Bayesian forecasts is qualitatively equivalent.

the case that the decision maker has some prior information about the environment which identifies a specific family of information structures.

In this Section we provide a simple example of a homogenous pool model to illustrate how an explanation-based approach might work when the set of possible explanations is constrained. We show that some of the results above are a product of the generality in the set of explanations. In particular, as we illustrate below, an explanation-based decision maker might not ignore forecasts as often as above when the set of explanations is more restricted, and thus may change her mind with repetition of forecasts. It also illustrates the difference between our approach above and the standard use of maximum likelihood in econometrics. We then discuss the implications of this example to some non-monotonicity properties of group decision rules.

Let us focus on a simple, homogenous pool, family of possible explanations. In particular, assume a binary state of the world with a uniform prior and assume that all information structures in this family have the following structure. There are $T+1$ binary, conditionally independent signals, $(s^*, s^1, ..., s^T)$, that each agree with the state of the world with probability $q > 0.5$. Each forecast q^t is determined by either observing the realisation of s^*, with probability $\alpha \in [0,1]$, or by observing the realisation of s^t, with probability $1-\alpha$. This family of explanations is characterized by the parameter α which represents the degree of correlation across forecasts. When $\alpha = 1$ we have full correlation across forecasts and when $\alpha = 0$ we have conditionally independent forecasts.

If this family of information structure is indeed what is behind the observed forecasts, then the observed forecasts could either be q or $1-q$. Therefore, given that the order of forecasts should not make a difference, each Q^t can be fully characterized by the number k of forecasts q. The following Proposition characterizes the maximum likelihood explanation-based prediction as a function of k for large n.

Proposition 5: *Let $k = \gamma T$. (i) When $\gamma = q$, there is a large enough T such that the MPE explanation is $\alpha = 0$ and the MPE prediction is 1. (ii) When $\gamma = 1$, then for any T, the MPE explanation is $\alpha = 1$ and the MPE prediction is q.*

The proof follows the law of large numbers. Intuitively, when all forecasts are the same, then the most likely explanation is that all are fully correlated. On the other hand when the set of forecasts is large enough, and the share of q forecasts is exactly as predicted by the (conditional) independent case, then this is the most likely one.[12]

[12]Note also that an MPE who does not restrict the set of explanations to the homogenous pool model would behave in the same way when $\gamma = 1$, but would have a low confidence with a prediction of $\frac{1}{2}$ for all interior γ.

Note that this example is similar to how econometricians use maximum likelihood estimators. When the set of explanations is restricted to a specific family, and this family is the correct one, then a large enough set of observations yields the asymptotically correct estimator of α and as a result of the state. As a result, decision making based on such a procedure will be equivalent in the limit to a standard Bayesian procedure.

The result above allows us to shed light on a peculiar form of judicial making. The Talmudic Sanhedrin court law is an interesting example of a decision rule that is not monotonic in the share of the votes. Specifically, it requires that if judges are unanimous in conviction, the defendant should be set free, while if only a majority convict, this majority verdict pertains.[13] While our main analysis focused on a single decision maker, we can interpret the set of forecasts as a set of information held by committee members or juries. If they have the same preferences and share these forecasts, then a committee with T members is equivalent to a single decision maker who faces T forecasts. The result in Proposition 5 shows how non-monotonicity arises. A majority -but not unanimity- of votes to convict will send the defendant to jail, while she will be spared with a unanimity of votes to convict. Consensus leads to low confidence and cautiousness compared to the majority case.

6 More than two states of the world

We conclude the paper by generalizing the results of Section 3 to more than two states of the world. The differences from the binary environment are not substantial, but some of them are interesting to note. We start with some extension of the notation.

Consider then a decision maker who is forming a prediction about a state of the world $\omega \in \Omega = \{\omega_1, ..., \omega_n\}$. She has a prior p, $p = (p_1, ..., p_n)$, a probability distribution over over Ω. A forecast t is a (full support) probability distribution, $q^t = (q_1^t, ..., q_n^t)$, over Ω. A *consistent explanation* of Q^t is a couple, (I, s^*) where I is a joint information structure and $s^* \in S$ such that $\forall \omega_i \in \Omega$ for all $j \in \{1, 2, .., t\}$, $q_i^j = \Pr_{\mathcal{I}}(\omega_i | s^{*,j}) = \frac{p_i q^j(\omega_i | s^{*,j})}{\sum_{l=1}^n p_l q^j(\omega_l | s^{*,j})}$.

For any forecast $j = 1, ..., t$, let $i^j \in \arg\max_{i \in \{1,...,n\}} \frac{q_i^j}{p_i}$. In what follows we assume that for any $j = 1, ..., t$, i^j is unique.

Proposition 1*: Given a set Q^t, the MPE prediction of the decision maker is well-defined,

[13] Glatt (2013) offers a maximum likelihood rationalisation of this rule; unanimity among many judges most likely is a result of strong correlation between the judges, and therefore demands caution. Gunn et al (2016) discuss this interpretation also in other legal scenarios.

unique and satisfies, for any $k, k' \in \{1, ..., n\}$:

$$\frac{\eta^{MPE}_{Q^t}(\omega_k)}{\eta^{MPE}_{Q^t}(\omega_{k'})} = \frac{\min_{j=1,...,t}\{\frac{p_{ij}}{q^j_{ij}} q^j_k\}}{\min_{j=1,...,t}\{\frac{p_{ij}}{q^j_{ij}} q^j_{k'}\}}$$

Note that the likelihood ratio between two states arising from the MPE prediction does not satisfy independence of irrelevant states. As this ratio depends on "special" states i^j for some forecasts j, the likelihood ratio between two states k and k' will depend on these states i^j. This of course differs from the binary case for which there are only two realisation and thus the issue of independence of the MPE maximum likelihood ratio plays no role.[14]

Another difference is that an ignored forecast will not be necessarily "dominated" by one other forecast, but potentially by a set of several other forecasts. In the binary model for example with equal priors, a forecast q^2 was dominated by q^1 when $q^2 \in (0.5, q^1)$. When we have more than two states this relation will be more involved.

In what follows for any history of forecasts Q^t, let $Q^t_{-t'}$ be the same history of forecasts excluding period $t' \leq t$.

Definition 3: A forecast t' is *ignored in* Q^t if $\eta^{MPE}_{Q^t} = \eta^{MPE}_{Q^t_{-t'}}$.

We show below that the notion of ignoring a forecast implies a transitive incomplete relation on forecasts. This result is sufficient in order to extend the proofs of all results in the subsection describing the dynamic properties of the MPE.

Let (Q^t, q^{t+1}) be the history Q^t and the additional forecast for period $t+1$, q^{t+1}.

Proposition 6: *(i) (no path dependence): The MPE prediction is the same for any history with the same set of forecasts as in Q^t even if they arrive in different order; (ii) (Weak monotonicity): If q^{t+1} is ignored in (Q^t, q^{t+1}), it is ignored in any (Q^s, q^{t+1}) where $Q^t \subset Q^s$. (ii) (Transitivity) If q^{t+1} is ignored in (Q^t, q^{t+1}), and $q^{s+1} \in Q^t$ is ignored in (Q^s, q^{s+1}) then q^{t+1} is ignored in $\{Q^t_{-s} \cup Q^s, q^{t+1}\}$. (iii) For any history Q^t, there is a decisive set $Q' \subset Q^t$ such that no $q' \in Q'$ is ignored in Q^t, all $q \in Q^t \backslash Q'$ are ignored in Q', and $\#|Q'| \leq n$ implying that when $t > n$ at least $t - n$ forecasts are ignored.*

A final difference from the binary case relates to whether predictions end up in the convex hull of forecasts, or whether additional polarisation arises. In fact, in the general model, some situations we are assured to be outside of the convex hull:

[14] For the same reason we have the violation of event independence, see Dietrich and List (2017).

Corollary 5: *Suppose there exists an* i^* such that *for any* j, $i^j = i^*$ *and that the set of forecasts* j *such that* $j \in \arg\min_{t=1,...,T}\{\frac{p_{i^*}}{q^t_{i^*}} q^t_i\}$ *for some* i *is not a singleton. Then polarisation arises so that* $\eta^{MPE}(\omega_{i^*}) > \max_j q^j(\omega_{i^*})$.

This case arises for example when all forecasts have the same mode but differ in how they view other states. This implies that the MPE prediction on the other states becomes weaker due to the disagreement between the forecasts, while the MPE belief becomes stronger on the mode of the forecasts.

7 Conclusion

We have analyzed a decision maker who combines forecasts by finding the most likely explanation that have yielded these forecasts, and using this explanation to form a prediction. We show how such explanations will be based on a high degree of correlation. As a result, when no meaningful constraints are imposed, many forecasts that are ignored, while extreme forecasts are prominent. When the decision maker knows the particular family of information structures that generates the forecasts, then we have identified a specific functional form that rationalizes why decision rules are not necessarily monotonic in forecasts and why unanimity can lead to low confidence.

8 Appendix

We first present the proof of Proposition 1 for the non-binary case, as well as of Proposition 6, which is important for the other proofs that follow. Henceforth for the remainder of the Appendix, we denote by a superscript * a result that is worded more generally for the non-binary case and provide its proof which will therefore prove the result without the * as well.

Proof of Proposition 1*:

Step 1: For any information structure, for any $\omega \in \Omega$, $q(\mathbf{s}^*|\omega) \leq \min_{j=1,...,m} q^j(s^{*,j}|\omega)$.

Proof: This is by the definition of a joint distribution function.

Step 2: By consistency of the model (Definition 1), we must have for any $\omega_k, \omega_{k'} \in \Omega$:

$$\frac{q^j_k}{q^j_{k'}} = \frac{p_k}{p_{k'}} \frac{q^j(s^{*,j}|\omega_k)}{q^j(s^{*,j}|\omega_{k'})}$$

This implies that by setting $q^j(s^{*,j}|\omega_1)$ at some level, we pin down all values $q^j(s^{*,j}|\omega_i)$, $i = 1, ..., n$.

Using the inequality and equality above, we can write the upper bound for the likelihood as:

$$\begin{aligned}\sum_{\omega_i\in\Omega} p_i q(\mathbf{s}^*|\omega_i) \quad \leq \quad & p_1 \min_j\{q^j(s^{*,j}|\omega_1)\} \\ & +p_2 \min_j\{q^j(s^{*,j}|\omega_2)\} \\ & ... + p_n \min_j\{q^j(s^{*,j}|\omega_n)\} = \\ & p_1[\sum_{i=1}^{n} \min_j\{\frac{q_i^j}{q_1^j} q^j(s^{*,j}|\omega_1)\}]\end{aligned}$$

We will now provide a solution to the above problem by maximising the right hand side and showing that we can achieve a likelihood equal to the maximal upper bound. Note that the problem is separable across the forecasts, and that increasing $q^j(s^{*,j}|\omega_1)$, for each forecast j, increases all the values for forecasts.

Note that for any $i \in \{1, ..., n\}$, $q^j(s^j|\omega_i) = \frac{q_i^j}{q_1^j}\frac{p_1}{p_i} q^j(s^j|\omega_1) \leq 1$ which implies an upper bound on $q^j(s^{*,j}|\omega_1)$. Therefore we can set $q^j(s^{*,j}|\omega_1)$ at the upper bound,

$$q^j(s^{*,j}|\omega_1) = \frac{p_{i^j}}{p_1}\frac{q_1^j}{q_{i^j}^j}$$

Note that $\frac{p_{i^j}}{p_1}\frac{q_1^j}{q_{i^j}^j} \leq 1$ by the definition of i^j.

Thus going back to the original problem, we can set

$$q^{MPE}(\mathbf{s}^*|\omega_i) = \min_j\{\frac{p_{i^j}}{p_i}\frac{q_i^j}{q_{i^j}^j}\}$$

And as a result,

$$\begin{aligned}\eta^{MPE}(\omega_i) \quad &= \quad \frac{p_i q^{MPE}(\mathbf{s}^*|\omega_i)}{\sum_{i=1}^n p_i q^{MPE}(\mathbf{s}^*|\omega_j)} \\ &= \quad \frac{\min_j\{\frac{p_{i^j}}{q_{i^j}^j} q_i^j\}}{\sum_{k=1}^n \min_j\{\frac{p_{i^j}}{q_{i^j}^j} q_k^j\}}.\end{aligned}$$

This implies the expression in the Proposition. ■

Proof of Proposition 6: By Proposition 1*, we know that the solution depends for any state ω_i on $\min_{j=1,...,m}\{\frac{p_{i^j}}{q_{i^j}^j} q_i^j\}$. (i) Let q^1 *be ignored in* $Q = (q^1, Q')$. This implies that for any i, $\min_{j\in Q'}\{\frac{p_{i^j}}{q_{i^j}^j} q_i^j\} \leq \frac{p_{i^1}}{q_{i^1}^1} q_i^1$. This will remain so if more forecasts are added. (ii) Let q^1 be ignored in Q' as above and q^2 be ignored in Q''. By the fact that q^2 is ignored in Q'', and by (i), we know that $\min_{j\in\{q,Q'\setminus q^2\cup Q''\}}\{\frac{p_{i^j}}{q_{i^j}^j} q_i^j\} \leq \min_{j\in\{q,Q'\}}\{\frac{p_{i^j}}{q_{i^j}^j} q_i^j\} \leq \frac{p_{i^1}}{q_{i^1}^1} q_i^1$ where the latter inequality follows from the fact that q^1 is ignored in Q'. Thus we get the result. (iii)

As for any state we choose the forecast that solves $\min_{j=1,...,t}\{\frac{p_{ij}}{q_{ij}^j}q_i^j\}$, there will be at most "decisive" n forecasts. Moreover, by the finiteness of the problem there must exist at least one forecast which is not ignored and thus such a set always exists. Thus at least $t-n$ forecasts will be ignored.■

Proofs of Corollaries 1 and 3: These are included in the proof of Proposition 6.

Corollary 2* *To generate the same sequence of predictions as a decision maker who always knows all the forecasts, a decision makers needs to carry over each period at most n* forecasts.

Proof of Corollary 2*: By Proposition 5, we know that at period n, once the decision maker has n forecasts, the set of predictions that are not ignored in Q^n is at most of size n. Denote this set by $\hat{Q}^n$. By Proposition 5, all forecasts in $Q^n \backslash \hat{Q}^n$ will be ignored in every future period. Thus in period $n+1$ the decision maker needs to consider $\{\hat{Q}^n, q^{n+1}\}$. Again from part (iii) we know that the set of forecasts that are not ignored in $\{\hat{Q}^n, q^{n+1}\}$ will be of size of at most n and that ignored forecasts in $\{\hat{Q}^n, q^{n+1}\}$ will be ignored in every future period and thus the decision maker does not need to remember them. A simple inductive argument completes the proof.■

Proof of Proposition 2*: Suppose that $(\mathcal{I}, \mathbf{s}^*) \in \arg\max_{(\mathcal{I}',\mathbf{s}'^*)\in C^{Q^{t+1}}} L(Q^{t+1}|(\mathcal{I}', \mathbf{s}'^*))$. Let $\hat{q}^{Q^{t+1}}(\mathbf{s}, \omega)$ be the distribution of signals in a consistent model of Q^{t+1} in $\arg\max_{(\mathcal{I}',\mathbf{s}'^*)\in C^{Q^{t+1}}} L(Q^{t+1}$ and let $\hat{q}^{Q}(\mathbf{s}, \omega)$ be the distribution of signals in a consistent model of Q^t in $\arg\max_{(\mathcal{I}',\mathbf{s}'^*)\in C^{Q^t}} L(Q^t|(\mathcal{I}', s$

Following the proof of Proposition 1* we have either (i) $\hat{q}^{Q^{t+1}}(\mathbf{s}^*|\omega_i) = \hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i) \Leftrightarrow \min_{j\in Q^{t+1}}\{\frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}\} = \min_{j\in Q^t}\{\frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}\}$, or (ii) $\hat{q}^{Q^{t+1}}(\mathbf{s}^*|\omega_i) < \hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i) \Leftrightarrow \min_{l\in Q^{t+1}}\{\frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}\} < \min_{j\in Q^t}\{\frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}\}$.

Now we can choose a new consistent model of Q^{t+1} with distribution $\tilde{q}^{Q^{t+1}}(s^*|\omega_i)$ as follows. First, following the proof of Proposition 1*, all the marginal probabilities for receiving $s^{j,*}$ are set at $\tilde{q}^{j,Q^{t+1}}(s^{j,*}|\omega_i) = \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}$.

In case (i): For states ω_i in this case we have $\tilde{q}^{j,Q^{t+1}}(s^{j,*}|\omega_i) = \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j} \geqslant \min_{j'\in Q^{t+1}}\{\frac{p_{ij'}}{p_i}\frac{q_i^{j'}}{q_{ij'}^j}\}$ so we construct $\tilde{q}^{Q^{t+1}}$ so that:

Let $\alpha \in [0,1]$ satisfy $\hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i) + (1-\hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i))\alpha = \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}$. Note that such an α exists as $\hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i) = \min_{j'\in Q^{t+1}}\{\frac{p_{ij'}}{p_i}\frac{q_i^{j'}}{q_{ij'}^j}\} \leq \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j} = \tilde{q}^{j,Q^{t+1}}(s^{j,*}|\omega_i)$. Now set $\tilde{q}^{Q^{t+1}}(\mathbf{s}^*_{|Q^t}|\omega_i) = \hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i)$, $\tilde{q}^{Q^{t+1}}(s^{j,*}|\mathbf{s}^*_{|Q^t}) = 1$ and $\forall \mathbf{s} \in S_{|Q^t}$, $\mathbf{s} \neq \mathbf{s}^*_{|Q^t}$ $\tilde{q}^{Q^{t+1}}(s^{j,*}|\mathbf{s}) = \alpha$. For these values we have $\tilde{q}^{Q^{t+1}}(s^{j,*}|\omega_i) = \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}$ by the definition of α and $\tilde{q}^{Q^{t+1}}(s^{j,*}, \mathbf{s}^*_{|Q^t}|\omega_i) + \tilde{q}^{Q^{t+1}}(s^{j,-*}, \mathbf{s}^*_{|Q^t}|\omega_i) = \hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i)$.

In case (ii): For states ω_i in this case we have $\tilde{q}^{j,Q^{t+1}}(s^{j,*}|\omega_i) = \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j} < \min_{j' \in Q^{t+1}}\{\frac{p_{ij'}}{p_i}\frac{q_i^{j'}}{q_{ij'}^j}\}$ so we construct $\tilde{q}^{Q^{t+1}}$ so that $\tilde{q}^{Q^{t+1}}(s^{j,*}|\omega_i) = \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j}$, $\tilde{q}^{Q^{t+1}}(\mathbf{s}^*_{|Q^t}|s^{l*}) = 1$ and $\tilde{q}^{Q^{t+1}}(\mathbf{s}^*_{|Q^t}|s^{j,-*}) = \beta$, where $\beta \in [0,1]$ satisfies $\tilde{q}^{Q^{t+1}}(s^{j,*}|\omega_i) + (1 - \tilde{q}^{Q^{t+1}}(s^{j,*}|\omega_i))\beta = \hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i)$. Note that such an β exists as $\hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i) = \min_{j' \in Q^{t+1}}\{\frac{p_{ij'}}{p_i}\frac{q_i^{j'}}{q_{ij'}^j}\} > \frac{p_{ij}}{p_i}\frac{q_i^j}{q_{ij}^j} = \tilde{q}^{j,Q^{t+1}}(s^{j,*}|\omega_i)$. For these values we have, by the definition of β, $\tilde{q}^{Q^{t+1}}(s^{j,*}, \mathbf{s}^*_{|Q^t}|\omega_i) + \tilde{q}^{Q^{t+1}}(s^{j,-*}, \mathbf{s}^*_{|Q^t}|\omega_i) = \hat{q}^{Q^t}(\mathbf{s}^*_{|Q^t}|\omega_i)$.■

Proof of Lemma 1: Construct the new model by maintaining the same distribution over signals as in $(\mathcal{I}', \mathbf{s}'^*)$, but re-labeling the signal names. In particular, label all the individual elements in $\mathbf{s}'^*$ as s^*. In addition, for any $j = 1, ..., t$, bundle all other signal values by one value s^{-*}. Note that this relabeling does not affect the consistency of the model, the prediction nor the likelihood as they all just depend on $\mathbf{s}'^*$.■

Lemma 2 *Suppose a sequence of forecasts* $\{q^t\}_{t=1}^{\infty}$. *The sequence of predictions* $\{\eta^{t,MPE}(q^t, \eta^{t-1,MPE})$ *arising for the naïve and sophisticated individuals in the limited memory model is the same.*

Proof of Lemma 2: The highest likelihood ratio for any $\eta^{t-1,ML}$ will be achieved when $q^1 = q^2 = ... = q^{t-1} = \eta^{t-1,ML}$, which is the same as the naive individual would use.■

Proof of Corollary 4: Note that the lowest that $\eta^{T-2,MPE}$ can be is q'. Thus $\eta^{T-1,MPE}(^{T-2,MPE}, q)$ must be above $\frac{1}{2}$ and thus $\eta^{T,MPE}(\eta^{T-1,MPE}, q)$ will be q. If $q^T = q'$ then even if $\eta^{T-1,MPE} = q$, which is the highest possible, then $\eta^{T,MPE}(\eta^{T-1,MPE}, q')$ will be lower than q.■

Proof of Proposition 4: (i) is a corollary of Proposition 1 and shown in Example 1. To see (ii) notice that if in some period t the action in period t is different to that in period $t+1$ under the full observability model, this must mean that q^{t+1} has to be further than the prior than any other forecast q^s with $s < t+1$. But by Proposition 1 this implies that q^{t+1} has to be further than the prior than any other prediction under the limited observability model for periods s with $s < t+1$. Therefore, the prediction under the limited observability model in period $t+1$ must be different to that of period t. (iii) Consider the following example: Let $q^1 > \frac{1}{2}$, and $q^2 < \frac{1}{2}$ and $q^2 < q^3 < \frac{1}{2}$. For this sequence, under the full observability model the prediction in period 3 will be the same as that in period 2. But assume that $\frac{\frac{q^1}{1-q^1}}{\frac{\hat{q}}{1-q^1} + \frac{1-q^2}{q^2}} > \frac{1}{2}$ then under the limited observability model the actions in the three periods are all different.■

Proof of Proposition 5: Suppose that the state is 1. Fix α. By the law of large numbers we know that almost surely the fraction of observed forecasts equal to q will be either $q(1-\alpha) + \alpha$ with probability q or $q(1-\alpha)$ with probability $1-q$. Now consider the

state 0. By the law of large numbers we know that almost surely the fraction of observed forecasts equal to q will be either $(1-q)(1-\alpha)+\alpha$ with probability $1-q$ or $(1-q)(1-\alpha)$ with probability q.

Suppose that a fraction $\gamma = q$ of forecasts q is observed. By the law of large numbers this could have arisen when the state was 1 and $\alpha = 0$ or when the state was 0 and the signal s^* was realised to signal that the sate was 1, with probability $(1-q)$. Among these two possibilities, the first has a likelihood proportional to $\frac{1}{2}$ while the second to $\frac{(1-q)}{2} < \frac{1}{2}$. Therefore, in this case the explanation is $\alpha = 0$ and given this explanation, the prediction is that the state is 1 with probability one.

Now suppose that a fraction $\gamma = 1$ of forecasts p is observed. In this case, for any n, this could arise as $\alpha = 1$ and the state is either 1 or 0 but the signal s^* was realised to signal that the sate was 1. The likelihood of the observation given this observation is $\frac{1}{2}$. But the likelihood of this observation under any other explanation is strictly smaller than $\frac{1}{2}$. To see this, fix $\alpha < 1$. The likelihood of observing n forecasts p is then given by,

$\Pr(n \text{ forecasts } q) =$
$0.5q \sum_{m=0}^{n} \binom{n}{m} \alpha^m ((1-\alpha)q)^{n-m}$
$+0.5(1-q)((1-\alpha)q)^n$
$+0.5(1-q) \sum_{m=0}^{n} \binom{n}{m} \alpha^m ((1-\alpha)(1-q))^{n-m}$
$+0.5q((1-\alpha)(1-q))^n =$
$0.5q(\sum_{m=0}^{n} \binom{n}{m} \alpha^m ((1-\alpha)q)^{n-m} + ((1-\alpha)(1-q))^n)$
$+0.5(1-q)(\sum_{m=0}^{n} \binom{n}{m} \alpha^m ((1-\alpha)(1-q))^{n-m} + (1-\alpha)q)^n$
$< 0.5q + 0.5(1-q) = 0.5.$
The inequality follows as
$\sum_{m=0}^{n} \binom{n}{m} \alpha^m ((1-\alpha)q)^{n-m} + ((1-\alpha)(1-q))^n =$
$\Pr(n \text{ forecasts } q | s^* = \omega = 1) + \Pr(n \text{ forecasts } 1-q | s^* = \omega = 1) < 1$
and $\sum_{m=0}^{n} \binom{n}{m} \alpha^m ((1-\alpha)(1-q))^{n-m} + (1-\alpha)q)^n =$
$\Pr(n \text{ forecasts } q | s^* \neq \omega = 0) + \Pr(n \text{ forecasts } 1-q | s^* \neq \omega = 1) < 1.$■

Proof of Corollary 5: Let $j^* = \arg\max_j q^j(\omega_{i^*})$. Now note that $\eta^{MPE}(\omega_{i^*}) = \frac{p_{i^*}}{\sum_{k=1}^{n} p_{i^*} \min_j \{\frac{q_k^j}{q_{i^*}^j}\}} =$

$\frac{1}{\sum_{k=1}^{n} \min_j \{\frac{q_k^j}{q_{i^*}^j}\}} > \frac{1}{\sum_{k=1}^{n} \frac{q_k^{j^*}}{q_{i^*}^{j^*}}} = q_k^{j^*}.$■

References

[1] Ambrus, A., Greiner, B. and P. Pathak (2015), How individual preferences get aggregated in groups – An experimental study, Journal of Public Economics, 129, 1-13.

[2] De Marzo PM, Vayanos D, Zwiebel J. (2003). Persuasion bias, social influence and unidimensional opinions. Q. J. Econ. 118:909–68.

[3] Dietrich F. and C. List (2017), Probabilistic Opinion Pooling Generalized, Part One: General Agendas, Part Two: The Premise-based Approach, Social Choice and Welfare

[4] Dietrich F. and C. List (2016), Probabilistic opinion pooling [an introductory review], Oxford Handbook of Probability and Philosophy, 2016.

[5] Ellis, A. and M. Piccione (2017), Correlation Misperception in Choice, American Economic Review 107(4):1264-92.

[6] Enke, B. and F. Zimmerman (2013). Correlation Neglect in Belief Formation. mimeo.

[7] Eyster, E. and M. Rabin (2010). Naïve Herding in Rich-Information Settings. American Economic Journal: Microeconomics, 2(4): 221-43.

[8] Eyster, E. and G. Weizsäcker (2011). Correlation Neglect in Financial Decision-Making. Discussion Papers of DIW Berlin 1104.

[9] Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41 155-160.

[10] Genest C. and J.V. Zidek (1986), Combining Probability Distributions: A Critique and an Annotated Bibliography, Statistical Science, Vol. 1, No. 1 (Feb., 1986), pp. 114-135. Institute of Mathematical Statistics.

[11] Gilboa, I. and D, Schmeidler (2003), Inductive Inference: An Axiomatic Approach, Econometrica, Vol. 71, No. 1. pp. 1-26.

[12] Gilboa, I. and D, Schmeidler (2010), Simplicity and likelihood: An axiomatic approach, Journal of Economic Theory, Elsevier, vol. 145(5), pages 1757-1775

[13] Glatt, E. (2013), The Unanimous Verdict According to the Talmud: Ancient Law Providing Insight into Modern Legal Theory, Pace International Law Review, Vol 3, No. 10.

[14] Glaeser, E. and C. R. Sunstein (2009), Extremism and social learning, Journal of Legal Analysis, Volume 1, Number 1.

[15] Golub, B. and M. Jackson (2012), "How Homophily Affects the Speed of Learning and Best-Response Dynamics", Quarterly Journal of Economics, pp. 1287–1338.

[16] Gunn, L.J., F.s Chapeau-Blondeau, M.D. McDonnell, Bruce R. Davis, Andrew Allison, Derek Abbott, (2016), Too good to be true: when overwhelming evidence fails to convince, Proceedings of the Royal Society A. 472 20150748; DOI: 10.1098/rspa.2015.0748. Published 23 March 2016.

[17] Jiang, J. and W. Tian (2016). Correlation Uncertainty, Heterogeneous Beliefs and Asset Prices. mimeo, University of North Carolina.

[18] Kallir, I. and Sonsino, D. (2009). The Perception of Correlation in Investment Decisions. Southern Economic Journal 75 (4): 1045-66.

[19] Levy, G. and R. Razin (2017), Combining Forecasts: Why Decision Makers Neglect Correlation, mimeo, LSE.

[20] Levy, G. and R. Razin (2015a), Does polarization of opinions lead to polarization of platforms? the case of correlation neglect, with Ronny Razin, Quarterly Journal of Political Science, Vol. 10: No. 3, pp 321-355.

[21] Levy, G. and R. Razin (2015b), Correlation Neglect, Voting Behaviour and Information Aggregation, with Ronny Razin, American Economic Review.

[22] Mullainathan S. and R. C. Waggoner (2017), Machine Learning: An Applied Econometric Approach, Journal of Economic Perspectives, Volume 31, Number 2, Spring, Pages 87–106.

[23] Ortoleva, P. (2012), Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News, American Economic Review 2012, 102(6): 2410–2436.

[24] Ortoleva, P. and E. Snowberg. (2015). Overconfidence in political economy. American Economic Review, 105: 504-535.

[25] Pennington, N. (1981), Causal reasoning and decision making: The case of juror decisions. Unpublished doctoral dissertation, Harvard University.

[26] Pennington, N. and R. Hastie (1988), Explanation-based decision making: The effects of memory structure on judgment, Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 521-533.

[27] Schkade, D., C.R. Sunstein and D. Kahneman (2000), Deliberating about Dollars: The Severity Shift, Columbia Law Review, Vol. 100, No. 4, pp. 1139-1175.

[28] Smith, L. and Sorensen, P. 2000. Pathological outcomes of observational learning. Econometrica, 68, 371–98.