Synthese (2022) 200:315
https://doi.org/10.1007/s11229-022-03809-5

ORIGINAL RESEARCH

Prediction with expert advice applied to the problem of prediction with expert advice

Daniel A. Herrmann[1]

Received: 28 September 2021 / Accepted: 5 July 2022 / Published online: 24 July 2022

Abstract
We often need to have beliefs about things on which we are not experts. Luckily, we often have access to expert judgements on such topics. But how should we form our beliefs on the basis of expert opinion when experts conflict in their judgments? This is the core of the novice/2-expert problem in social epistemology. A closely related question is important in the context of policy making: how should a policy maker use expert judgments when making policy in domains in which she is not herself an expert? This question is more complex, given the messy and strategic nature of politics. In this paper we argue that the prediction with expert advice (PWEA) framework from machine learning provides helpful tools for addressing these problems. We outline conditions under which we should expert PWEA to be helpful and those under which we should not expect these methods to perform well.

Keywords Social epistemology · Machine learning · Expertise · Probability · Prediction · Policy

1 Introduction

Goldman (2001) poses the *novice/2-expert problem*: how should a novice (non-expert) make judgments about the reliability of two rival experts when they disagree?[1] This is a difficult problem because the novice cannot deploy any domain knowledge in order to evaluate the different expert advice. That is, the very thing about which the novice is trying to learn seems to be needed in order to learn.

[1] A related question social epistemologists have investigated is how to assign belief to a group based on the beliefs of the group members (for example, List & Pettit, 2011).

✉ Daniel A. Herrmann
daherrma@uci.edu

[1] Department of Logic and Philosophy of Science, University of California, Irvine, Irvine, CA, USA

Far from being an idle philosophical problem, this problem comes up in the important context of policy making.[2] Indeed, policy makers are often in situations in which they are not themselves experts about certain domains—climate science, technology, epidemiology—and yet they need to make critical decisions in these domains. In these cases a policy maker might like to aggregate the opinions of experts so that she can make good predictions about the world. Thus, we see that policy makers must solve a version of the novice/2-expert problem: to whom should they listen when making policy?

In addition to the core of the novice/2-expert problem (the novice lacks the relevant expertise) there are a number of other epistemic perils that a policy maker faces. For example, there may be bad actors in the community who are willing to deceive the policy makers in order to influence policy choice, or there may be psychological and political forces that corrupt the epistemic hygiene of the experts.[3]

Given that a real policy maker will want a solution that works in a complex and perilous political world, the policy maker's strategy for aggregating expert judgments must satisfy some *strategic* as well as epistemic desiderata. While there are many aspects of an aggregation method worth considering, there are at least three minimal conditions it must satisfy:

1. The method must be feasible to deploy.
2. The method should make good predictions, given the experts available.[4]
3. The method should be palatable to different interest groups in the policy maker's community.[5]

One strategy that Goldman discusses[6] to solve the novice/2-expert problem is that the novice could use the past track records of the experts to help solve this question. Goldman notes a number of issues with this approach, and observes that this strategy "[o]f course...provides no algorithm by which novices can resolve all their two-expert problems" (2001, p. 108). In this paper we follow up on this strategy and discuss a

[2] Goldman himself points out that this problem has practical importance. Indeed, he writes, "...some issues in epistemology are both theoretically interesting and practically quite pressing. That holds of the problem to be discussed here: how lay-persons should evaluate the testimony of experts and decide which of two or more rival experts is most credible" (2001, p. 85).

[3] For example, cognitive biases, belief polarization, and the accidental spread of misinformation make this task more difficult.

[4] This will of course be sharpened once the formal framework is on the table.

[5] This will also be sharpened, but an intuitive description is the following. Consider a society in which there has been radical belief polarization between two groups, the A's and the B's. The method should be such that both groups, if they have confidence in their own experts' predictions, should welcome the use of the method by the policy maker. Thus this is a sort of "unbiased" condition. In particular, the kind of groups for which this must hold are groups that would have some ability to challenge this system of aggregation and make it unworkable. Thus if there were a third group, the C's, that had no ability to disrupt the efforts of the policy maker, it would be less important for this method to appease the C's. Of course, the policy maker might want to appease some less powerful groups for moral reasons, but that only strengthens this as a desideratum.

[6] See Sect. 6 of Goldman (2001).

class of algorithms that have precise characterizations and theoretical guarantees from a subfield of machine learning, *prediction with expert advice* (PWEA).[7]

In Sect. 2 we discuss the novice/2-expert problem. In Sect. 3 we introduce the prediction with expert advice framework. In Sect. 4 we consider how these algorithms provide a kind of solution to the novice/2-expert problem, and also their limitations. We then take the lessons gleaned from the analysis of prediction with expert advice algorithms in the context of the novice/2-expert problem and apply it to the practical case of policy making. To do so, in Sect. 5 we discuss the connection between prediction and policy making, and the application of PWEA to policy contexts. In Sects. 6 and 7 we argue that, under certain conditions, PWEA provides us with a powerful set of tools for helping to inform policy decisions. In Sect. 8 we consider how these methods hold up in an adversarial political landscape. We then offer some brief concluding remarks.

2 The novice/2-expert problem

Goldman states the novice/2-expert problem as follows:

> The novice/2-expert problem is whether a layperson can *justifiably* choose one putative expert as more credible or trustworthy than the other with respect to the question at hand, and what might be the epistemic basis for such a choice? (2001, p. 92)

Goldman considers five different sources of evidence that a novice might use in order to choose which expert to trust in the novice/2-experts problem. In this paper we consider the last of the sources he considers: the experts' past track records. Goldman himself states that track records “may provide the novice's best source of evidence for making credibility choices" (2001, p. 106).

In order to follow up on Goldman's suggestion, we first review an important difficulty that Goldman identifies with this approach. This difficulty is that the novice might not be in a position to evaluate the past success of the experts. For example, imagine a novice who is watching two mathematicians writing out the answers to a challenging math problem that she cannot herself solve. How will she know which, if either, is correct? If she cannot determine which answer is correct, this track record of answers does not help her at all.

Goldman offers a distinction that helps us understand when such an approach will be feasible and when it will not. Following Goldman, we call a statement *esoteric* if the truth-value of the statement is *inaccessible* to the novice, and we call a statement *exoteric* if the truth-value of the statement is *accessible* (2001, p. 94).

[7] For an excellent introduction to the field see Cesa-Bianchi & Lugosi (2006). Philosophers have already used the prediction with expert advice framework to offer new solutions to Hume's problem of induction (Schurz, 2008, 2009, 2019; see Arnold, 2010 and Sterkenburg, 2019 for a discussion of limitations of the approach) and to clarify the foundations and assumptions of Solomonoff's theory of inductive inference (Solmonoff, 1964 and Li & Vitányi, 2008; see Sterkenburg, 2018, chapter 6 for the analysis with prediction with expert advice).

With this distinction in hand, we can diagnose the problem in our example. The answers to the math questions are esoteric for the novice; she cannot evaluate their truth value. However, suppose there were a reliable computer that would check the answers for her. Then, even though she could not compute the answer herself, if she had access to the computer the statements would become exoteric for her.

We will keep this distinction in mind when we evaluate the application of the PWEA algorithms to the novice/2-expert problem. We will see that an important strength, and limitation, of the framework is that it forces us to consider only exoteric statements.

3 Prediction with expert advice

In this section we introduce the core of the prediction with expert advice framework. In the following section we will consider the extent to which it address the novice/2-expert problem.

3.1 Prediction games

Imagine that you are playing a game with your friend *Environment*. You each have a deck of cards in front of you. Every round you first make a *prediction* by placing a card face up on the table in front of you and Environment. After you make your prediction, Environment chooses an *outcome* from her deck of cards. If your card matches the outcome then you don't suffer any penalty; otherwise you lose a point. Luckily for you, you have assembled a crew to help you in this game. You have a group of friends who are also making predictions, and you get to see their predictions before you have to make your own. So, each round, in addition to trying to figure out how Environment is generating outcomes so that you can avoid loss, you are also thinking about which of your friends make good predictions so that you can copy them.

Such a game is similar to a *prediction game*—the core formal framework of prediction with expert advice.[8] Formally, a prediction game is a repeated game between a forecaster N and the environment (we use "N" to draw the parallel between the forecaster in PWEA and the novice in the novice/2-expert problem). The forecaster is trying to predict an unknown sequence of events $y_1, y_2, \ldots$ where $\mathcal{Y}$ is an outcome space and $y_t \in \mathcal{Y}, \forall t = 1, 2, 3, \ldots$. The outcome space is analogous to Environment's deck of cards above, where each card is a possible outcome.

The forecaster's predictions $\hat{p}_1, \hat{p}_2, \ldots$ are elements of a decision space $\mathcal{D}$. The decision space is analogous to your deck of cards above, where each card you can play is a prediction. However, there is an important difference. In the example above, you could only play a single card. In the prediction with expert advice framework, the decision space $\mathcal{D}$ needs to be a convex subset of a vector space. What this means intuitively is that the predictions can be *mixtures* of the possible outcomes. For example, if in the game above you were allowed to give a probability distribution of the cards as your prediction, then your decision space would be the set of all probability

[8] Cesa-Bianchi and Lugosi describe the prediction protocol as a repeated game (2006, p. 7). Schurz (2008) also describes PWEA using the idea of a prediction game.

distributions over cards, which is in fact a convex set. One thing this tells us is that, in general, $\mathcal{D}$ need not be identical to $\mathcal{Y}$—that is, the set of possible predictions does not need to be equal to the set of possible outcomes.[9]

There are two other key components in a prediction game. The first is a set of experts $\mathcal{E} = \{E_i\}$, which also predict the sequence.[10] This is analogous to the crew you assembled in the card game. The second is a nonnegative loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}^+$. This is analogous to the loss of points you were trying to avoid in the card game. The main idea is that the worse your prediction is, the larger the loss you suffer.

The prediction game is played at each round $t = 1, 2, 3, \ldots$. At each round t the environment chooses an outcome $y_t \in \mathcal{Y}$ and chooses the expert advice $\{f_{E,t} \in \mathcal{D} | E \in \mathcal{E}\}$. The forecaster is able to view the expert advice and then make a prediction $\hat{p}_t$ given the advice. Finally, the outcome y_t is revealed to the forecaster and all the experts, and the forecaster and each expert incur a loss according to the loss function ℓ. Each predictor's loss will be based on that predictor's prediction, and the actual outcome.

Notice that we describe the *environment* as also choosing the expert's predictions. This is a little counter-intuitive, but is meant to capture the idea that we can think of everything that is not under the forecaster's control as part of the environment. Schurz and Thorn suggest we consider the outcomes the *natural* part of the environment, and the predictions of the other players the *social* part of the environment (2016, p. 36).

One important feature to note here is that we don't assume that the experts are *agential* in any way—that is, they need not be rational players also trying to minimize their loss. So when we say that each predictor receives a loss, this is more of a helpful tool for the forecaster to keep track of how well each predictor is doing than anything that *needs* to influence the future predictions of an expert (of course, it could).

For a simple example of a prediction game consider the context in which $\mathcal{Y} = \{0, 1\}$, and $\mathcal{D} = [0, 1]$. That is, the forecaster is attempting to predict a binary sequence, and the predictions she issues are probabilities that the next element of a sequence is a 1. The loss function ℓ might be the absolute loss function $\ell(d, y) = |d - y|$. Our set of experts might be $\mathcal{E} = \{E_i\}_{i \in 11}$, where $f_{E_i,t} = i/10$ for all t. That is, the ith expert always predicts that the next element of the sequence will be a 1 with probability $i/10$. Each round the forecaster is able to look at how well each expert has done in the past, look at their current prediction for this round, and then make a probabilistic prediction of her own.

3.2 Forecaster's goal

When we evaluate different possible ways to combine expert advice in the PWEA framework, we evaluate them with respect to how well they achieve a particular goal. This goal is for the forecaster to make predictions of the sequence that are *optimal*[11] with respect to her set of experts, $\mathcal{E}$. The intuition is that we want to our forecaster to

[9] This is important for the later part of our paper because, as we will discuss in Sect. 5, policy makers will need to be able to make probabilistic predictions.

[10] In general we assume $\mathcal{E}$ is finite.

[11] We have not yet said what "optimal" means in this context; this will be made precise soon.

do about as well as the best expert in $\mathcal{E}$ (or better!), *no matter which sequence obtains*. In order to make this precise we introduce the following notions.

For $E \in \mathcal{E} \cup \{N\}$ let $L_{E,T} := \sum_{t=1}^{T} \ell(f_{E,t}, y_t)$ be the cumulative loss of predictor E.[12] That is, $L_{E,T}$ tells you how much loss predictor E has accumulated by time T. The lower $L_{E,T}$ the better E has been doing. If we think back to the card game above, this is analogous to the total cumulative loss of points.

Recall that we want our forecaster to do well relative to the best expert to which she has access. In order to capture this ideally formally, we define the *cumulative regret* for each expert $E \in \mathcal{E}$ to be

$$R_{E,T} := L_{N,T} - L_{E,T}$$

We think of this as the regret that the forecaster N has about not listening to expert E. The goal of the forecaster is to choose its predictions at each round so that the regret is *small* compared to the *best* expert.

There are many different ways to precisify the condition that we want this regret to be small. One ambitious formulation of this requirement is that as $n \to \infty$ we have

$$\frac{1}{T}\left(L_{N,T} - \min_{E \in \mathcal{E}} L_{E,T}\right) \to 0.$$

Informally, this says that as the forecaster plays the game longer her per-round regret of having not listened to the best expert gets closer and closer to 0, meaning she is approximating the best expert well. As we will discuss below, there are very simple prediction algorithms that satisfy this condition. Furthermore, their regret actually decreases in T quite quickly, which is essential for our purposes in this paper.

3.3 Forecasting methods and results

A well-studied and general type of forecaster is the weighted average forecaster. The idea is to average the predictions of all the experts, according to how well each expert has done in the past. We keep track of how well each expert has done in the past with a *weight*. The larger the weight of an expert, the more of a contribution their prediction will make to the average. Formally, at time t the prediction of the weighted average forecaster is

$$\hat{p}_t = \frac{\sum_{E \in \mathcal{E}} w_{E,t-1} f_{E,t}}{\sum_{E \in \mathcal{E}} w_{E,t-1}}$$

where $w_{E,t-1}$ is the weight assigned to an expert E at time t and $f_{E,t}$ is the prediction that expert E makes at time t.

[12] Note that if $E = N$, then $f_{E,t} = \hat{p}_t$.

The variations of this forecasting method consist in different ways of assigning weights to experts, usually as some function of their past success, where more successful experts get more weight.[13]

There are many different variants of the weighted average forecaster, which are appropriate in different contexts. In order to give the reader a flavour for the theory and some of the main results we introduce here an important forecaster in the prediction with expert advice framework—the *exponentially weighted average forecaster*.

The exponentially weighted average forecaster is defined by

$$\hat{p}_t = \frac{\sum_{E \in \mathcal{E}} e^{-\eta L_{E,t-1}} f_{E,t}}{\sum_{E \in \mathcal{E}} e^{-\eta L_{E,t-1}}}$$

where η is a positive parameter that controls how much the learner decreases the weight of an expert that suffers a high loss. Note that the prediction of the exponentially weighted average forecaster does not depend on the past predictions of experts, but only on their performance. This makes it fairly straightforward to compute. If we allow η to be a function of time η_t, then with a judicious choice of η_t it is possible to prove that the per-round regret of the exponentially weighted average forecaster converges to 0 very quickly.[14]

The exponentially weighted average forecaster is only one method from the PWEA literature. There are many different methods that perform well relative to the best expert in their pool. Cesa-Bianchi and Lugosi (2006) summarize the general result from the field nicely:

> [I]t is possible to construct algorithms for online forecasting that predict an arbitrary sequence of outcomes almost as well as the best of M experts. Namely, the per-round cumulative loss of the predictor is at most as large as that of the best expert plus a term proportional to $\sqrt{\ln M/T}$ for any bounded loss function, where T is the number of rounds in the prediction game. (p. 99)[15]

Thus we see that these methods do well compared to the best expert in the pool, and they do well quickly. Furthermore, what is striking is that we achieve these bounds by considering the *worst* possible sequence of observations for the forecaster. Obviously we do not expect an adversarial world, and thus for actual prediction tasks we would expect the regret to be smaller than the worse-case scenario.

4 PWEA applied to novice/2-expert problem

Let us consider the extent to which PWEA addresses the novice/2-expert problem. They key lesson from PWEA is that, when the novice uses the expert advice to inform

[13] Note that we are guaranteed that $\hat{p}_t \in \mathcal{D}$ since $\mathcal{D}$ is convex and $\hat{p}_t$ is a convex combination of elements of $\mathcal{D}$.

[14] The rate is of order $1/\sqrt{T}$, for any sequence of outcomes. This holds for any loss function ℓ that is convex in its first argument and bounded (see Cesa-Bianchi & Lugosi, 2006, pp. 17–20 for more details).

[15] We have changed N to M in the quote to avoid confusion between the number of experts and the novice N. We also changed n to T to make it consistent with the notation here.

her own judgements, she need not *choose* between the experts. Rather, she can take a kind of average of their predictions as her best judgement. There is a sense in which this answers the question by changing it: instead of giving a procedure to choose an expert, we give a procedure to form a judgment based on expert advice.

Furthermore, the results from PWEA show that there are strategies that yield predictions which are optimal in a precise sense. Goldman noted that simply suggesting that the novice look at past track records did not give the novice a precise algorithm to follow (2001, p. 108). PWEA gives us specific algorithms, and we are rewarded for our formal precision with strong performance guarantees.

Notice that in order to apply the algorithms from PWEA the novice needs to be in a position to know all of the outcomes after they have occurred.[16] This is analogous to observing the card after all predictions have been registered in the card prediction game. In Goldman's language, the outcome must be exoteric for the novice. This puts limits on the kinds of cases for which PWEA can help our novice. They have to be cases for which the novice can verify the outcomes of the predictions.[17]

Furthermore, these methods are set up to deal with *sequences* of events. Even though there are many types of events we care about that have a sequential nature (weather, stock prices, drug effectiveness), this is an additional limitation.

Finally, the novice must have access to all of the expert predictions in advance of making her prediction. But this condition makes sense for the novice/2-expert problem; if the novice did not have access to the experts' advice, she wouldn't be facing this problem. She'd be facing the much worse novice/0-expert problem.

Thus, we see that PWEA gives us good methods to address the novice/2-expert problem, but that these methods will only work under certain conditions. This makes sense: most methods we reason about and deploy work only under certain conditions.

Following Goldman's insight that this problem is not merely academic but also “practically quite pressing" (2001, p. 85), in the remainder of the paper we investigate the extent to which PWEA methods can help resolve a particular practical instance of the novice/*n*-expert problem: policy making. We will see that this context satisfies the conditions for PWEA to applied, to an extent. Furthermore, we will see that PWEA does fairly well at satisfying a number of strategic considerations of this context not present in the basic novice/2-expert problem.

5 Prediction in policy making

PWEA is a well-studied and successful framework and set of methods, and, given its applicability to the novice/2-expert problem, it is plausible that it will be helpful for policy makers who also need to form judgments based on expert advice. Various

[16] There is actually a rich theory of how to make good predictions with expert advice when this condition is violated—that is, when the novice does not get to learn all of the outcomes. Of course, this does make things harder, and the guarantees are not quite as good. A detailed analysis of how this works is beyond the scope of the paper, but we encourage the interested reader to see chapter 6 of *Prediction, Learning, and Games* (2006).

[17] In the remainder of the paper we will see the implications this has for applying PWEA methods to real-world problems.

variants of it have already been applied successfully to numerous real-world contexts including forecasting wind speed (Zamo et al., 2020), subseasonal meteorology forecasting (Brayshaw et al., 2020), and disease progression (Morino et al., 2015). Let us work through the details of how a real policy maker might use it.

5.1 Predictions and decisions

One important aspect of policy making is that the policy maker doesn't only make predictions about the world, but also decisions about what to do. When we think about how a policy maker should evaluate expert opinions, it is important to track this fact.

The PWEA framework is fairly general, in the sense that we can think of it as giving us methods to aggregate predictions, but we can also think of it as giving us methods to aggregate *decisions*. For example, consider the game of rock, paper, scissors. Suppose we are playing this game many times with our friend, and we want to do as well as we can. There are two ways we might use a PWEA method to help us.

The first way is to try to form good judgements about what act (rock, paper, or scissors) our friend is likely to make in the next round. If we have access to some predictors, then we can use the method to try to predict what our friend will play. Once we have made such a prediction using the method (for example, we predict that our friend will play rock with probability 0.5, paper with probability 0.3, and scissors with probability 0.2), then we can use this probability to help make decisions about what to play. For example, we might use the decision procedure that recommends we take the act that maximizes our probability of winning. In this situation, our decision procedure is separate from our prediction procedure: we are not using PWEA to learn what act to take directly.

The second way we might use the methods from PWEA to play against our friend is to learn directly which acts to take. For example, instead of our experts predicting what our friend will play as above, the experts might themselves directly suggest certain acts, such as “play rock" or “play rock with probability 0.5 and paper with probability 0.5". The framework is applicable in this case as well, since we interpret the mixtures of the different acts not as probabilities that our opponent will play a certain move, but as a mixed strategy that we ourselves take.[18]

The core difference is that in the first way we use PWEA to generate probabilities (make predictions) and then use these probabilities as input to some decision procedure, while in the second way we use PWEA to generate mixed acts directly.

For this paper, we focus on the case in which the policy maker uses the first strategy: she uses PWEA to make predictions about events relevant for a decision, and then uses these probabilities as input to a separate decision procedure. It has a number of advantages over the second approach.

First, there are certain decision problems in which a policy maker might find it awkward to consider mixed acts. For example, consider a policy maker who is deciding whether or not to declare war.[19] The importance of such a decision may prohibit a

[18] In fact, if one applies the PWEA framework to learning acts directly, then one can construct a simple proof of the minimax theorem for zero-sum games (see Blum & Yishay, 2007 for details).

[19] We thank an anonymous referee for this example.

policy maker from basing what she does on the outcome of a coin toss;[20] the public might not be sympathetic. If she uses PWEA to learn which act to take directly, then this may be what the method recommends. However, if she uses PWEA to make predictions about various war-related events, and then makes a decision based on that procedure, she need not consider mixed acts.

Second, using PWEA to make decisions directly would collapse value judgements and epistemic judgements into a single judgment of what act to take. This has a number of serious costs in the policy context. One is that it gives the experts more direct influence over which act the policy maker takes, and that may be undesirable in the context of malicious experts.[21] Furthermore, this would push the policy maker into a situation in which she runs into a challenging version of the fact-norm-problem. In general, it is not true that having good predictions of events settles what one ought to do: one needs to combine one's epistemic judgements with one's value judgments. By collapsing these two into a single judgement, the policy maker is less able to disentangle values from the epistemology. In contrast, if she uses PWEA to generate probabilities, and then uses those probabilities to *inform* her judgement in conjunction with her values, this allows her to separate value questions from epistemic questions. This matters greatly, for example, for securing the buy-in of different political groups, since the groups can agree on the outcome, but not necessarily on how to evaluate the outcome. We discuss this more in Sect. 7.

Third, if one pursues the second strategy, then one must update the weights of the experts based on the value of the outcomes. However, it may take a long time for the quality of a decision to become clear. During this time, the policy maker will have to make more decisions, and yet she will be unable to use PWEA to refine her judgements about the quality of each expert (since she has not yet observed the value of the outcome). In contrast, if the policy maker is focused on predicting frequent events, then she can evaluate the quality of the experts more quickly, thus improving the probability judgements she uses to inform her decisions.

Finally, separating predictions about events from decisions allows the policy maker to consider the extent to which her policy decisions are *independent* of the events she is predicting. This will matter for determining contexts in which PWEA will prove helpful, from those in which her interventions distort the system on which she is intervening in a way that might undermine her goals.

5.2 Act-state (in)dependence

This last point is subtle. Suppose the policy maker is trying to predict some sequence of events. She might be in a position in which she expects that the policy she enacts will have an effect on the events. For example, if she is trying to make decisions about housing prices, she will want to know the future trajectory of housing prices. But this trajectory itself might depend on the policy she enacts. Situations like this exhibit a type of *act-state dependence*. Savage's decision-theoretic framework (1972) assumes act-state independence, and is still the dominant framework for decision making in

[20] More realistically, a random number generator.

[21] We expand on this more later in this section and in Sect. 8.1.

economics. In this framework, the agent maximizes her unconditional expected utility. Thus, she does not need to calculate the probability of a state on the supposition that she takes some act; this is a direct consequence of act-state independence.[22] In situations in which act-state independence is satisfied, the agent does not expect her policy choices to affect the events, and the PWEA framework can be applied as usual. Some policy contexts will exhibit act-state independence and some will not. Let us consider the extent to which PWEA methods can be applied in each case.

Policy contexts in which (approximate) act-state independence holds are natural candidates for the application of PWEA. This is because the decisions that the policy maker makes will not affect the decision-relevant states. For example, consider a coastal city that is deliberating about different strategies for mitigating the effects of climate change. The city might be deciding how to allocate its resources between hurricane preparedness and heatstroke support. In this context, the policy maker will want to assess the probability of both hurricanes and heatwaves of various degrees of severity. Notice that, at the level of a city making this sort of decision, the decision problem exhibits (approximate) act-state independence. Whether the city allocates 80% or 60% of its disaster budget to heatstroke support will not affect the probability of a hurricane. Many decisions contexts are like this.

In situations in which there is act-state dependence, decision making must be more sophisticated. When act-state independence is violated, in order to make intelligent decisions the agent will want to know the probability of the events, *supposing* some act were taken.[23] In these kinds of situations, it is less clear how or if PWEA will help her. This is because PWEA gives the policy maker unconditional probabilities, not suppositional probabilities. We will give a brief sketch of one approach we might pursue to address this, but leave the bulk of the analysis for future work.

Suppose the policy maker is in a situation in which she has two possible acts,[24] and she thinks there might be act-state dependence. She might ask each expert to make *two* predictions: one on the supposition that she takes the first act, and the other on the supposition that she takes the second act. The policy maker will then use these suppositional predictions to inform her decision. Of course, once the event is revealed, the policy maker will not get to observe the counterfactual outcome—what would have happened had she taken the act she did not in fact take. The question is how to shift the weights on experts for the counterfactual case. As briefly mentioned in Sect. 4, there is a version of PWEA in which the novice does not get to observe every outcome.[25] It would be interesting to see whether not such methods could be used to deal with the counterfactual event; treat it as an unobserved outcome, and proceed according to the recommendations by PWEA.

[22] There are actually two kinds of act-state independence: causal and probabilistic. It is not entirely clear which Savage was assuming.

[23] The kind of supposition the agent should make is a central debate in decision theory, and is one way of understanding the difference between causal decision theory and evidential decision theory. For a discussion of different kinds of supposition see chapter 6 of Bradley (2017). Luckily, for a broad range of cases, the two decision theories will recommend the same acts.

[24] In general, *n* acts.

[25] For an overview see chapter 6 of *Prediction, Learning, and Games* (2006).

Before we conclude this section, notice one final thing. In the context of act-state independence, the policy maker can use PWEA to aggregate expert advice about the probability of various states.[26] She can then use these judgements to help her make a decision. This a a *separate* task from judging how the policy maker's various acts might interact with the state to produce an outcome.[27] For example, in the example of the coastal city, the policy maker uses PWEA to aggregate expert opinions about the probability of various events. However, she does not use PWEA to form judgements about how her acts (such as shifting spending from one area to another) interact with states to produce outcomes. In many contexts, this will be obvious: spending more money on hurricane preparedness is likely to benefit the city if a hurricane occurs. Other contexts may be less clear. It may be possible to use some version of PWEA to do this in the policy context, but in this paper we focus on the application of PWEA to policy making in which the policy maker uses it to aggregate the judgments of experts about the probability of various events.

Thus, we see that, in the context of policy making, act-state independence[28] ensures that the policy maker can use PWEA to help her make predictions. We see that things are less clear when there is act-state dependence. An important open question is whether or not methods from PWEA can be modified to help the policy maker in such contexts.

5.3 Sequential prediction and policy making

Given that we are concerned with a policy maker who uses probabilities and values to make decisions, one immediate question we might have is: given the success of the Bayesian approach to epistemology and decision making, why shouldn't the policy maker just be Bayesian? A Bayesian agent has a well-defined probability function and utility function capturing her degrees of belief and desires respectively. When deciding on a policy from a set of possible policies she chooses the act that maximizes her expected utility.[29]

For such a Bayesian, learning from expert advice is simple. When she has access to expert advice captured by a proposition EA she uses her conditional probability function $p(\cdot|EA)$ in order to calculate her expected utility of different policies. This is the straight-forward Bayesian approach to learning from expert opinion, and following Bradley we will call this the "the gold standard for coherent revision of belief" (p. 6, 2018). The issue with this approach is, of course, that it is difficult to apply (see Bradley, 2018 for a discussion), and thus fails our first criterion from Sect. 1.[30]

[26] Of course, she can do this in other contexts as well; it is just tricky to evaluate how helpful it will be.

[27] In the framework of Savage, the proposal is to use PWEA to form the probability distribution over states; not to form judgements about how a particular act maps states to outcomes.

[28] Where the states are the events in the sequence.

[29] Breaking ties in some way.

[30] Another aspect that might make Bayesianism unattractive is the subjective choice of a prior. Though many happily embrace the subject choice of priors (for example, Finetti, 1974; Jeffrey, 2004), others find this to be a defect (for example, Jaynes, 2003; Williamson, 2010). In contrast, one can view the methods from PWEA as learning an optimal prior distribution by taking a weighted average of various probability

The setup from Sect. 3 is almost directly applicable to the context of a policy maker making predictions based on her epistemic community. We simply identify the policy maker with the forecaster, and the members of her epistemic community with the experts in $\mathcal{E}$. Then the results of the framework all hold, and thus the policy maker can reap these benefits.

The main components of the framework that warrant discussion in this context are the connection of sequential prediction to policy, the loss function, and the accessibility of expert prediction and outcomes. We discuss these to show that this framework would be practical for policy makers to deploy.

In order to understand the connection of sequential prediction to policy it will be helpful to contrast it with the optimal but infeasible Bayesian approach. One obvious difference between the fully Bayesian approach discussed above and the PWEA framework is the sequential nature of the PWEA framework, which we can think of as imposing a restriction on the domain of the probability function in the more standard Bayesian approach. In the fully Bayesian approach an agent has probabilities over an algebra consisting of all of the propositions an agent can entertain, whereas in the prediction with expert advice framework the probabilities (or, more generally, predictions) are confined to a much smaller domain—the set of possible observations that the agent might make in the next time step (which will be exoteric for the policy maker).

The explicitly sequential nature of the prediction with expert advice framework seems to differ from the more general Bayesian approach. This sequential aspect is in the spirit of Dawid's *prequential probability* (1984, 1992, 1992b), which emphasizes the role of predictions of observables in a sequence. Though on the surface this seems different to the Bayesian approach, in which agents have degrees of belief over a more general algebra of events without a necessarily sequential nature, there are rich historical and technical connections between the two approaches (Dawid, 1992).

There is a real way in which this approach is limited: it only works in cases where the policy maker needs to predict events that she will be able to verify in the future (exoteric events). However, we do buy some benefits at the cost of this limitation. The first is that this helps makes the method feasible, and the second is that it makes us focus on *observable* events. We defer a discussion of this second point to my discussion of the feedback the policy maker must have.

To understand the first point, note that the fully Bayesian picture is infeasible, even putting aside the computational issues of carrying out Bayesian updating. An actual policy maker could not write down all of the propositions she has the conceptual resources to consider, let alone a probability function over all such propositions that accurately represents her degrees of belief. It is much more realistic to imagine that policy makers use very local judgements about particular events (or types of events) to inform their policy making. For example, policy makers might want to know how likely it is that inflation will move in a particular direction in order to inform economic policy, or might want to know how likely it is that average global temperature will

distributions where the weights are sensitive to the empirical success of the different distributions. For the details of this interpretation of PWEA see Schurz (2019, p. 167) and Sterkenburg (2020, sect. 3).

increase by a certain amount in a particular year in order to inform environmental policy. Note that these (and many other) events have an obvious sequential nature.

This restriction to a very simple type of judgment is common in social epistemology. For example, two-armed bandit problems are commonly used to model epistemic problems (see Zollman, 2010, also for example O'Connor & Weatherall, 2018; Weatherall & O'Connor, 2020). The main thing the agents learn about in these models is which of two (or occasionally multiple) bandit arms has the highest chance of success. Other models focus on agents learning things such as which of two possible states of the world obtain (see for example Mohseni & Williams, 2019). Following the simplicity of these models in this literature and the observation that much public discourse and polarization happen around single topics (is climate change real? are vaccines effective?) it is clear that modelling the policy maker's dilemma without a large idealized algebra is useful and well-motivated.[31]

So much for the connection of sequence prediction to policy. With this on the table we can better understand what is at stake when we choose our loss function.

Notice that, in general, the predictions the forecaster makes depend on the choice of loss function; for example, the prediction of the exponentially weighted average forecaster is a function of $L_{E,t-1}$, which in turn depends on the choice of loss function. How should the policy maker select the loss function she uses? There is no general answer to this question, but there are a few technical and pragmatic considerations.

On the technical side, different loss functions can have better or worse bounds in the prediction with expert advice framework.[32] What this means is that, depending on the loss function, getting close to the best expert can be easier or harder. Even though there are a few constraints that all loss functions have,[33] there are still many possible loss functions to choose from.

It is important to stay very close to our pragmatic goal here: helping policy makers aggregate expert advice about the probabilities of various events, in order to take successful action. Thus, the choice of loss function will be context dependent. It will depend on the policy maker's goals, and in particular on the decision making method she plans to use. For an example of this kind of analysis carried out in a slightly different context, see Babic (2019) and King and Babic (2020). The authors consider how an agent's attitudes towards errors of different kinds can inform the choice of risk function the agent uses to evaluate her credences. This is very similar to the context we are considering here, except we consider the choice of loss function instead of risk function, and a more decision theoretic as opposed to purely epistemic flavour.

It is also clear that there many contexts in which there is *no* clear decision method that policy makers would plan to use—broad strokes of policy would be forged in the crucible of public debate and negotiation, followed by the details finalized by the aides writing the actual policy. Given this slightly chaotic nature of actual policy implementation it is dubious that the ultimate policy choice is sensitive to minor changes in the prediction in a principled way. Thus the specific choice of loss function

[31] See Dietrich and Christian (2017) for a more general discussion about dropping the condition that the set of relevant events in the context of probabilistic pooling must be a σ-algebra.

[32] See chapters 3, 8, and 9 in Cesa-Bianchi and Lugosi (2006).

[33] For example, most results in the literature assume bounded and convex loss functions.

is not terribly crucial; more important are the qualitative guarantees this framework yields. On this approach, it would be best to choose one of the easy to use and well studied loss functions in the literature, like absolute loss or logarithmic loss.

Finally, in order to apply the framework, it is necessary that the policy maker have access to the various experts' predictions and that she can clearly determine which state of the world obtains so that she can properly score the different experts.

Although we do not currently have the infrastructure in place for experts to register their predictions about specific events, this would be very feasible to create. Prediction markets like Metaculus and forecasting projects like The Good Judgment Project are proof that systems for posting and aggregating predictions can work in practice. If such infrastructure were put in place, accessing predictions would be straightforward for policy makers.

Less straightforward is the question of whether we can use the existing history of prediction to evaluate experts *right now*. The answer is that, in cases in which we have records of certain experts making clear predictions, we can use those records to help choose good experts. Where such records do not exist at the level of clarity needed to apply the framework we can make some qualitative judgments resembling that of the framework—for example, give more weight to experts with a better track record. However, the proposal is that, moving forward, experts would need to register their predictions on a platform of some kind.

More subtle is ensuring that the policy maker can determine which outcome of the world obtains. In general it can be quite challenging to formulate unambiguous propositions. Consider, for example, having experts assign probabilities to the proposition that a particular sports team will win a particular game. We might think of this a binary event, where 1 corresponds to one team winning and 0 corresponds to the other team winning. It is easy to see how this could go awry. What if the game gets called off? The current description doesn't make it obvious how we should evaluate this outcome. We could make 1 the outcome that the first team wins, and 0 the catch-all. Even this is not entirely satisfactory. What if there are 5 minutes left in the game, and the first team is ahead by a large amount of points. However, the game has to be called off due to an earthquake. Even though 0 technically did obtain, it seems that those who predicted that 1 was more likely were *more* right than those who predicted 0 was likely.[34]

Furthermore, the fact that these predictions are used for informing policy makes them high-stakes. Whenever even the slightest ambiguity is present, experts and political pundits on the losing side of the prediction will likely not concede. Consider again our example. Suppose that during the first half of the game that three of the players on the first team are injured and cannot play. Suppose that the first team then loses. The motivated expert can claim that what they *really* predicted was that the first team would win, but that since the players were injured part-way through and removed that it was not the same team, and that this event should not be used to update the weights the policy maker uses to make her prediction. In light of current political dynamics this kind of appeal does not seem at all unlikely—again, especially since whoever

[34] Thorn and Schurz in fact apply PWEA methods to predict the outcomes of the Monash University Footy Tipping Competition (2019). They chose to represent the first team winning with a 1 and the first team *not* winning with a 0.

performs well has more effect on the predictions of the policy maker, and thus of future policy decisions.

Fortunately for our purposes there is a body of literature on this very problem in an applied context. In particular, Tetlock has carried out an ambitious research program wherein he engaged with actual political experts making predictions about events, which he describes in *Expert Political Judgment* (2017). In the Methodological Appendix of *Expert Political Judgment* he details specific strategies he used to greatly reduce the ambiguity of such propositions. Though a full discussion would take us too far afield, to give the reader a flavour of his approach one test he used to asses the ambiguity of propositions was the *clairvoyance test*:

> our measures had to define possible futures so clearly that, if we handed experts' predictions to a true clairvoyant, she could tell us, with no need for clarifications ("What did you mean by a Polish Person or…?"), who got what right. (p. 14)

This approach and others provide a set of tools that policy makers could use to ensure that the world provides them with clear, unambiguous feedback about the degree to which different experts were successful. Such tools will not lead to perfect, infallible ways of constructing propositions. However, the evidence is clear that in practice they are fairly successful.

The necessity of specifying such clear, *observable* propositions might have certain benefits for charged political discourse. Consider the case of climate change. Climate change is a notoriously polarized topic.[35] The public debate happens at the level of whether or not climate change is *real*, or some variant of this. This is not an observable proposition; of course we can observe *evidence* for and against it, but the proposition itself is a theoretical construct we use in order to predict actual phenomena.

The insistence of the prediction with expert advice framework on making predictions only about observable events has the potential to help sidestep this whole issue of polarization.[36] The term "climate change" need never appear in an application of this framework devoted to the climate predictions we care about. We only need to focus on the actual consequences—frequency of storms in a given period, changes in average global temperature in a given period, etc. This can even help experts who would otherwise have incentive to lie or hedge (for example, if their political base favours a certain view on climate change) make honest predictions about these observable events that aren't as politically charged. An expert could even be a climate change denier and still make predictions in line with climate change models if that is her best guess, making up whatever story for why she made this prediction that will satisfy her base—which is completely fine in this framework. All that matters for the policy maker is success-

[35] For an empirical survey see McCright and Dunlap (2011); for modeling polarization in the case of climate change see Cook and Lewandowsky (2016).

[36] Of course, if there is no way for a certain thing we care about to connect up to observable events, then we would not be able to apply the framework. However, such a case would not lend itself easily to any kind of empirical investigation—if the multiple sides of the debate do not make any different empirical predictions, then it is likely the kind of case in which scientific expertise is not the kind of expertise needed. For example, certain fundamental moral disagreements might have this character.

ful prediction of phenomena.[37] Observable propositions can help remove some of the black powder from politics.[38]

I have shown that there are some contexts in which a policy maker wishing to aggregate the opinion of experts could use the prediction with expert advice framework to do so. In the remaining sections we discuss the virtues and limitations of the framework.

6 Feasibility and good predictions

In light of our discussion of how to actually apply PWEA, we now assess whether it meets the desiderata we laid out in Sect. 1. In this section we show there are methods from the prediction with expert advice framework that satisfy the first two conditions. *Feasibility.* In the previous section we showed that policy makers could apply the framework under the right conditions. My concern here is its feasibility—in particular, whether an actual policy maker could execute the methods.

Fortunately the answer is an enthusiastic yes. The most common methods (for example, the weighted average forecaster) are all computable. Additionally, their computational complexity is proportional to the number of experts,[39] which makes them feasible when the set of experts is not incredibly large.[40] For our purposes this is satisfied; the set of experts the policy maker will be aggregating in any real world context is likely to be quite small.

Two more features of our context make things even easier. The first is that the policy maker herself will not be computing the output of each expert system,[41] but rather aggregating them, so she herself does not suffer the cost of computing those predictions. The second is that, in most contexts a policy maker would use this framework, she has lots of computational time at her disposal. For example, aggregating economic predictions each quarter leaves plenty of time for a policy maker to compute the update to weights from the previous prediction.

Good predictions. Obviously for an aggregation method to be useful it must make good predictions. Or, rather, it must make good predictions *given the available expert*

[37] There are some cases in which we might want predictions of certain things which we will not get to observe, such as counterfactuals. See the discussion in Sect. 5.2.

[38] One might worry that without explicit causal models of underlying processes, we might lack theoretical reasons to think that the methods that predict well now will predict well in the future. We can see that this framework handles the issue nicely. If any of the experts' prediction methods use causal models, then the benefits of causal modelling for prediction will transfer directly to the policy maker via the aggregation process. For example, one of the experts might use a causal model to predict climate observations, and by hypothesis this has good theoretical guarantees for success in future predictions. Since the policy maker uses the kind of method discussed here, if the causal modeler is making good predictions, the expert will assign that expert high weight. But then since we have the theoretical guarantee of future success of that expert's predictions, the policy maker will also reap the benefits of such guarantees. Thus the framework does exactly what we intend it to do: it offloads all of the domain relevant reasoning to the experts—theoretical guarantees included—and lets the policy maker enjoy the fruits of their labour.

[39] See chapter 5 of Cesa-Bianchi and Lugosi (2006).

[40] Even when the set of experts is incredibly large, many contexts have structures that can be exploited to recover good bounds. For a detailed discussion see chapter 5 of Cesa-Bianchi and Lugosi (2006)

[41] Unlike the case of *simulateable experts*—see Cesa-Bianchi and Lugosi (2006, section 2.9) for details.

advice. Recall from Sect. 3.2 that in the prediction with expert advice framework the goal is to minimize the regret of the forecaster's predictions compared to that of the best expert in the group, and that we in fact have methods with such guarantees that hold no matter which sequence obtains.

This is exactly what we would want in this context. The policy maker herself is not an expert, so we want to equip her with tools to effectively leverage the expert advice she has available to make good predictions. Furthermore, the fact that these guarantees hold no matter the actual sequence ensures that this is a fully general method. This means that in cases in which one of the experts does well so will the policy maker. In cases where no expert does well, then it is unreasonable to expect a non-expert to do better.

As emphasized in Sect. 3.3, the specific method the policy maker uses will depend on the context, and on the loss function she chooses. Taking these features of her situation into account she can then choose the appropriate method.

7 Buy-in from different groups

Different interests groups have different policy goals. Many groups use expert opinion in order to support their policy recommendations. This tends to happen in the public sphere, with each group presenting their own evidence and experts. We have seen that prediction with expert advice supplies principled and effective methods to aggregate expert opinion and inform policy makers. In order to transition to a political world in which we would deploy these methods we would need support from multiple interest groups. It is important to see if these methods would incentivize such a switch.

This consideration is especially important in the current political context, rife with polarization. Polarization has attracted recent interest in social epistemology (see for example Bramson et al., 2017; O'Connor & Weatherall, 2018, and Singer et al., 2019). Polarization makes it that much more challenging for opposing groups to agree on a common action.

It is very plausible that different interest groups would find it desirable for policy to be informed by the prediction with expert advice framework in the way we have sketched. This is for two main reasons: the methods are fair and transparent, and successful prediction leads to policy control.

Fairness and Transparency. All of the standard aggregation methods assign equal initial weight to each expert.[42] It is in this sense that the methods are fair. No particular opinion or political group is favoured *a priori*; the expert's track record speaks for itself. These methods are also transparent. Once the particular method is selected, the method, expert opinion at each step, and outcome at each step can all be public knowledge. Individual groups could compute for themselves the policy maker's prediction to verify that the procedure is being followed. Thus there is little room for manipulation.

[42] Rather, they do not need to do so, but they can do so, and in this context it is an advantage.

Success Leads to Policy Control. Each interest group believes that the experts to which their group appeals are correct.[43] Given their interest in controlling policy and the fact that policy depends in some part on the predictions of the policy maker, each interest group should welcome the chance to influence the policy maker's predictions through their own experts' predictions. The rationale is straightforward. An interest group believes that, on average, their expert will make more successful predictions than the experts of their political opponents. Thus they expect to be able to exert more control over the policy makers predictions at each round, allowing them to control policy through this channel. Since different interest groups will reason in this way, we can achieve buy-in from the different groups.

As an example, this feature allows minority opinions to do well. Consider, the case of climate change. There has long been a near consensus among climate scientists that anthropogenic climate change is both likely occurring, and likely to pose grave risks. Despite this near consensus, there are of course dissenters (for example, Linden, 1993). Given the lack of consensus among the public (McCright & Dunlap, 2011), it is clear that many people do not take the scientific near-consensus to be trustworthy. In this case it would be in the interests of both sides of this divide to shift the debate to a context like that of prediction with expert advice. Since all that matters to the aggregation methods is success, even groups with minority views would be able to affect policy—if their views are correct![44]

8 Prediction in adversarial contexts

We have shown that the methods from the prediction with expert advice framework satisfy important desirable features of aggregation methods. Indeed, not only are their *formal* properties exactly the kind we want, but they also have properties that make them *politically* feasible to use.

In this section we consider how the PWEA framework fares in policy contexts in which there are bad actors. It is important to remember, however, that we are trying to provide a useful way for policy makers to aggregate expert opinion. The world is a messy and strategic place; no method will ever be perfect. We believe that the discussion in this paper so far has demonstrated that there is much to be gained in considering this framework.

8.1 Prediction with malicious experts

The guarantees described in Sect. 6 hold regardless of what the experts predict. Thus they of course hold even if there are malicious experts. Despite this, it is still important to consider how susceptible this framework is to attacks by malicious experts. The concern is that even though the worst-case scenario guarantee still holds, a malicious

[43] Or, more cynically, at the very least they must *publicly* declare that the experts to which their group appeals are correct.

[44] Specifically, if their views about the observable events being predicted are correct.

expert might be make strategically dishonest predictions in order to degrade the quality of the prediction.

Consider two distinct cases. The first we call *loss maximization* and the second we call *policy manipulation*.

Loss Maximization. We imagine that there is at least one expert in the pool that is trying to maximize the loss of the policy maker. This kind of goal might be appropriate, for example, in a situation in which a hostile foreign power has somehow gained control of one of the expert systems that the policy maker queries. In this situation we imagine that the malicious expert is trying to maximize the loss of the policy maker—that is, she wants her to make bad predictions. The foreign power isn't trying to manipulate the policy maker so that she chooses some particular policy, but instead is simply trying to degrade their predictions.

Truong et al. (2017) have analyzed this very case (of course, without the political interpretation). They identify the optimal strategy for the malicious expert when playing against the weighted average prediction algorithm. They consider maximizing the loss with two common loss functions: logarithmic loss and absolute loss. Surprisingly, for the logarithmic loss case, the optimal strategy is the greedy policy that lies at every step. The absolute loss case is much more challenging to analyze, but they are able to show in a restricted setting that the optimal policy is a threshold policy, in which the malicious expert is honest until she has a large enough relative weight and then lies after.

Thus it seems that the optimal malicious strategy is easy to carry out. However, it is important to note that there are a few assumptions that the result requires that are quite implausible in our context. In particular, the malicious expert needs to know both the prediction distribution of each other expert and they need to know the *true future outcome* in the sequence. This last requirement is clearly not going to be satisfied in our context. Thus, it seems very likely that this kind of strategy will be less effective when the malicious expert is not a perfect clairvoyant. Future work understanding exactly how successful a more limited malicious expect can be at maximizing the policy maker's loss would be helpful.[45]

Policy Manipulation. Another obvious concern is that experts will attempt to manipulate the policy maker so that she chooses some particular policy. For example, an expert might intentionally predict a lower probability of temperature increase in order to nudge carbon tax policy in a certain direction.

Careful and thorough investigation of this strategic scenario would be quite useful. We make a few preliminary observations, but leave the heavy lifting for the future. The first is that in order for such deception to be effective, the manipulative expert must already have a high weight—otherwise her deception will not affect the prediction much. This means that she must have been fairly successful in the past. But then she would have to be the peculiar kind of expert who can predict well, and yet has policy goals that move against the grain of her honest predictions. This is of course entirely possible—a different interest group with different goals might be in power, and so this might be a prudent thing for her to do. But it still puts a wrinkle of difficulty on things.

[45] Etesami et al. (2020) consider a similar case. The same conditions implausible in our context are required for their analysis.

Furthermore, once she intentionally makes a bad prediction, then by her expectation she will lose weight, and thus have less leverage over future predictions. Thus, this kind of manipulation is quite *costly*. This implies that, even if there are manipulative experts of this character, they would not be able to manipulate policy like this too often. Again, a more precise analysis of this would be useful.

8.2 Prediction with malicious policy makers

It is interesting to consider how the policy maker herself might influence things. It is not the case where she lies about the recommendation of the method—as discussed in §6, other actors could compute the recommendation for themselves, and thus the lie would be exposed. However, the policy maker has some degree of freedom when choosing events in the sequence.

Consider, for example, a policy maker who is concerned about three different policy areas—the environment, the military, and cyber security. Suppose she is in a society where there are two political factions, the greens and the purples. She herself is a green. Suppose furthermore that she expects the green experts to do well predicting events concerning the military and cyber security, but poorly on events concerning the environment. Given the connection between prediction and policy, she knows it would be challenging to justify her preferred environmental policy if the environmental predictions she makes are skewed to the purple experts. Since the method she plans to use will track successful prediction, and she expects the purples to be successful at predicting environmental events, she is in a dilemma.

She comes up with a plan to manipulate the system. Instead of three separate sequences of events, she combines all three sequences into one category—she calls it *security predictions*. Since she expects the greens to do well on military and cyber security predictions, she expects the method to assign more weights to the greens, even though they do poorly on the environmental predictions. But because they are all part of the same sequence, the environmental predictions of the policy maker will skew towards the recommendations of the purple experts. In this way, by diluting the domain in which she expects green experts to fail with those in which she expects them to succeed, she can carry out her preferred environmental policy with support from the aggregation method.

This is a version of the reference class problem with a strategic element.[46] In the context of PWEA, Schurz introduces a more sophisticated way of dealing with contexts in which experts are superior to each other in different domains (sects. 5.9 and 8.2, 2019). This is *conditionalized meta-induction*, in which the prediction method conditionalizes the success rate of the experts on the different domains, and uses conditionalized weights to make predictions about different domains. In order to do this, the method needs access to information about the member of a *reference partition* to which the event to be predicted belongs. In the above example, the reference partition would be whether or not the event to be predicted is of type environment, military, or cyber security. Even with this more sophisticated formulation, if the policy maker can manipulate the reference partition, in the above example by not making these categories

[46] See Hájek (2007) for a discussion of the reference class problem in different contexts.

distinct cells in the reference partition, then she might be able to manipulate the system so that her experts get more weight on domains in which they are unsuccessful.

There are interesting technical questions here that warrant investigation. Perhaps an even more sophisticated method can learn the reference partition itself. However, we do think that for actual political applications this kind of manipulation would likely be obvious and thus difficult to justify. Furthermore, even without this strategic picture in the background, it seems best to apply this framework in cases where there is an obvious repeatable event—inflation per quarter, for example, or average global temperature. With such narrowly defined event sequences this kind of manipulation is less of a concern.

9 Concluding remarks

We have argued that PWEA offers a compelling solution to a version of the novice/2-expert problem. The framework gives a precise way to track the success of the experts and aggregate their judgments that yields good guarantees about the relative success of the novice. In order for her to apply the framework, the events that the novice cares about have to be exoteric for her. We have seen that there are theorems that offer strong theoretical guarantees that predictors who use methods from PWEA will not do much worse than the best expert available to them. We have argued that the underlying facts that make these theorems true in the ideal case are approximately satisfied in more realistic cases, and thus that these methods are applicable in real world contexts. Furthermore, we argued that these methods allow policy makers to separate challenging issues surrounding ideology and theory from possibly more directly relevant questions about the outcomes of future events.

We also outlined a few limitations and areas of future investigation. A careful treatment of the various strategic dimensions of the application of this framework is particularly interesting. Also important is the question about the extent to which a policy maker can apply the framework to situations in which act-state independence is violated. We hope that this paper motivates this kind of work, and motivates social epistemologists to consider prediction with expert advice as a helpful framework.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No.1922424. Many thanks for their support. My deepest thanks to Cailin O'Connor and James Weatherall for invaluable discussion and feedback throughout the entire development of the paper. Additionally, I thank the members of the NSF research group they organize at Irvine for reading an earlier draft. I thank Brian Skyrms, Louis Narens, and the members of the Social Dynamics seminar they run for giving feedback on the project. Two anonymous referees provided many useful comments and suggestions. Finally, I thank the following people for engaging with the project in various ways: Tom Sterkenburg, Josiah Lopez-Wild, Colleen Spencer, Aydin Mohseni, Gerard Rothfus, Bruce Rushing, Benjamin Genta, Clara Bradley, Jacob VanDrunen, Calum McNamara, Max Notarangelo, and John Wilcox.

Funding This work was supported by National Science Foundation Grant 1922424. Many thanks for their support.

Data availibility Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest Not applicable.

Ethical approval Not applicable.

Informed consent Not applicable.

Consent for publication Not applicable.

References

Arnold, E. (2010). Can the best-alternative justification solve Hume's problem? On the limits of a promising approach. *Philosophy of Science, 77*(4), 584–593.

Babic, B. (2019). A theory of epistemic risk. *Philosophy of Science, 86*(3), 522–550.

Blum, A., & Yishay, M. (2007). *Learning, regret minimization, and equilibria.*

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Bradley, R. (2018). Learning from others: Conditioning versus averaging. *Theory and Decision, 85*(1), 5–20.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., et al. (2017). Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of Science, 84*(1), 115–159.

Brayshaw, D., Paula, G., & Florian, Z. (2020). A new approach to subseasonal multi-model forecasting: Online prediction with expert advice. In *EGU General Assembly Conference Abstracts*, 17663.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science, 8*(1), 160–179.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General), 147*(2), 278–290.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian statistics, 4,* 109–125.

Dawid, A. P. (1992b). *Prequential data analysis.* Lecture Notes-Monograph Series (pp. 113–126.)

Dietrich, F., & Christian, L. (2017). Probabilistic opinion pooling generalized. Part one: General agendas. *Social Choice and Welfare, 48*(4), 747–786.

Etesami, S. R., Kiyavash, N., & Poor, H. V. (2020). Adversarial policies in learning systems with malicious experts. arXiv:2001.00543.

Finetti, B. (1974). Theory of probability: A critical introductory treatment. Technical report.

Goldman, A. I. (2001). Experts: Which ones should you trust? *Philosophy and phenomenological research, 63*(1), 85–110.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese, 156*(3), 563–585.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jeffrey, R. (2004). *Subjective probability: The real thing*. Cambridge University Press.

King, Z. J., & Babic, B. (2020). *Moral obligation and epistemic risk.* Oxford studies in normative ethics 10.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*. Vol. 3. Springer.

Linden, H. R. (1993). A dissenting view on global climate change. *The Electricity Journal, 6*(6), 62–69.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly,52*(2).

Mohseni, A., & Williams, C. R. (2019). *Truth and conformity on networks*. Erkenntnis, pp. 1–22.

Morino, K., Hirata, Y., Tomioka, R., Kashima, H., Yamanishi, K., Hayashi, N., et al. (2015). Predicting disease progression from short biomarker series using expert advice algorithm. *Scientific Reports, 5*(1), 1–12.

O'Connor, C., & Weatherall, J. O. (2018). Scientific polarization. *European Journal for Philosophy of Science, 8*(3), 1–21.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science, 75*(3), 278–305.

Schurz, G. (2009). Meta-induction and social epistemology: Computer simulations of prediction games. *Episteme, 6*(2), 200–220.

Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. MIT Press.

Schurz, G., & Thorn, P. D. (2016). The revenge of ecological rationality: Strategy-selection by meta-induction within changing environments. *Minds and Machines, 26*(1), 31–59.

Singer, D. J., Bramson, A., Grim, P., Holman, B., Jung, J., Kovaka, K., et al. (2019). Rational social and political polarization. *Philosophical Studies, 176*(9), 2243–2267.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control, 7*(1), 1–22.

Sterkenburg, T. F. (2018). *Universal prediction*.

Sterkenburg, T. F. (2019). The metainductive justification of induction: The pool of strategies. *Philosophy of Science, 86*(5), 981–992.

Sterkenburg, T. F. (2020). The meta-inductive justification of induction. *Episteme, 17*(4), 519–541.

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?-New edition*. Princeton University Press.

Thorn, P. D., & Schurz, G. (2019). Meta-inductive prediction based on Attractivity Weighting: Mathematical and empirical performance evaluation. *Journal of Mathematical Psychology, 89,* 13–30.

Truong, A., Rasoul Etesami, S., Etesami, J., & Kiyavash, N. (2017). Optimal attack strategies against predictors-learning from expert advice. *IEEE Transactions on Information Forensics and Security, 13*(1), 6–19.

Weatherall, J. O., & O'Connor, C. (2020). Conformity in scientific networks. *Synthese*, 1–22.

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford University Press.

Zamo, M., Bel, L., & Mestre, O. (2020). Sequential aggregation of probabilistic forecasts|Application to wind speed ensemble forecasts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. https://doi.org/10.1111/rssc.12455

Zollman, K. J. S. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17.