ROBIN HANSON

FOR BAYESIAN WANNABES, ARE DISAGREEMENTS NOT ABOUT INFORMATION?

ABSTRACT. Consider two agents who want to be Bayesians with a common prior, but who cannot due to computational limitations. If these agents agree that their estimates are consistent with certain easy-to-compute consistency constraints, then they can agree to disagree about any random variable only if they also agree to disagree, to a similar degree and in a stronger sense, about an average error. Yet average error is a state-independent random variable, and one agent's estimate of it is also agreed to be state-independent. Thus suggests that disagreements are not fundamentally due to differing information about the state of the world.

KEY WORDS: Agree, Bounded rationality, Common belief, Disagree

1. INTRODUCTION

Theory and observation seem to be in conflict. On the one hand, persistent disagreements on matters of fact seem ubiquitous, such as in academia, politics, and speculative trade. Such disagreements persist, even though two or more sides seem well aware of the disagreement. On the other hand, we have theory suggesting that rational agents cannot agree to disagree in this manner. Bayesians with common priors cannot so disagree (Aumann, 1976; Sebenius and Geanakoplos, 1983; McKelvey and Page, 1986; Neilsen et al., 1990), even approximately (Monderer and Samet, 1989; Sonsino, 1995; Neeman, 1996a).

To resolve this conflict, we might posit that people do not actually disagree as much as they seem, or that people are so irrational that it is feasible and profitable for them to disagree less than they do. A third resolution, however, is to posit that existing theoretical results are fragile, and do not hold up under more reasonable and feasible concepts of rationality. Yet many authors have explored

Theory and Decision **54:** 105–123, 2003.

weaker rationality assumptions, and have often still found agreeing to disagree to be irrational (Geanakoplos, 1989; Rubinstein and Wolinsky, 1990; Geanakoplos, 1994; Samet, 1990; Morris, 1994).

Even such weaker concepts of rationality, however, often require unrealistic computational abilities. For example, even agents characterized only by possibility correspondences are assumed to exactly calculate expected values over what are typically truly immense sets of possible states. Some authors have directly considered computational constraints, such as by simulating the consequences of specific computational strategies, or by allowing agents to know all theorems which can be generated in any finite time by a Turing machine (Megiddo, 1989; Shin and Williamson, 1994; Lipman, 1995). The literature has lacked, however, general results regarding agents with more substantial computational limitations.

To fill this gap, this paper considers a general class of agents with perhaps severe computational constraints. For an agent in this class, her estimate regarding any random variable of interest is itself a random variable. Such agents may be unrealistic in the sense that we do not model any confusion on their part regarding the meaning of a random variable, and we restrict attention to finite spaces of possible states. But otherwise, our agents may use any error-prone state-dependent computational strategy to estimate the value of any random variable.

We call such an agent a "Bayesian wannabe" if she can make sense of counterfactual statements about what her estimates would have been if she were capable of computing exact Bayesian expected values. That is, she in principle has a prior and an information partition, but may in practice find it very hard to use them to compute expected values. Even so, she can make sense of the concept of her "error", the difference between her actual estimate and the estimate she would have if she had sufficient computational abilities to be a perfect Bayesian.

We can say that two Bayesian wannabes disagree when the difference between their estimates of some random variable is large enough. And, in the usual way, we can say that Bayesian wannabes agree on some claim when they are both sufficiently confident that the claim is true, *and* that they agree. We can thus ask: when do Bayesian wannabes agree to disagree?

There are three obvious sources of disagreement: priors, information, and computation (i.e., errors[1]). We already know, however, that not all of these sources can by themselves induce agreements to disagree. Previously mentioned results show that for Bayesians with a common prior, differing information can not by itself induce agreements to disagree. In contrast, perfect Bayesians with identical information but differing priors *can* clearly agree to disagree.

Similarly, Bayesian wannabes can agree to disagree purely due to computational errors, since disagreements about computing the value of a shared prior can have the same result as differing priors. Another example would be two agents who were fully aware that in every state one of them estimated $\pi \approx 3.14$ while the other estimated $\pi \approx 22/7$. (Of course, as with differing priors, it is not obvious how rational this behavior is.)

Thus we know that agreeing to disagree can be purely due to differing priors or differing computation, but *not* purely due to differing information (at least for Bayesians). Differing information, however, seems a more rational basis for disagreement than priors or computation. Can we retain a central role for information in persistent disagreements by attributing them to intrinsic *combinations* of differing information, priors, and computation, combinations which do not reduce to pure disagreements of any sort?

For two perfect Bayesians with any combination of differing information and priors, one can easily find state-independent random variables about which the Bayesians will agree to disagree, even though state-information is irrelevant to Bayesians' estimates of such variables. (Consider, for example, the average of some random variable across the universe of states.) Thus persistent disagreements due to combinations of differing information and priors seem to "reduce" to persistent disagreements based purely on differing priors, in the sense that you can't have the first kind without the second kind also being present.

This paper provides a similar result about combinations of differing information and computation, suggesting that you can't have such a combination without also having a pure computational disagreement. Specifically, this paper considers two Bayesian wannabes with identical priors but differing information and computational errors, and gives conditions under which any example of

agreeing to disagree implies that the same agents agree to disagree in a certain strong way about a state-independent random variable.

To get this result, we assume that agent estimates satisfy certain consistency constraints. These assumptions are non-trivial, but they are also not computationally difficult; even quite severely computationally constrained agents can satisfy these constraints. Specifically, we first assume that when two agents, call them Pam and Sue, agree to disagree about some random variable, each agent estimates herself to be unbiased in those estimates. That is, while each agent expects that in some states her estimates will be too high and in other states her estimates will be too low, she expects that her bias, the average of these errors across states where she thinks they agree to disagree, will be zero.

We also assume that Pam keeps her estimates consistent with a simple inequality constraint. This constraint takes as input Pam's estimate of the amount by which Pam and Sue agree to disagree here, and Pam's estimate of their accuracy this agreement, i.e., of how often they actually agree, when they think that they agree. From these, a simple formula gives a lower bound on Pam's estimate of Sue's bias.

Finally, we do not actually require that the above assumptions hold in every state. Instead we require that Pam and Sue *agree* that a direct implication of these assumptions holds. That is, if the agents agree that Sue's estimate of Sue's bias is zero, and that Pam's estimate of Sue's bias is above a positive lower bound, then Pam and Sue by definition agree to disagree about Sue's bias in a strong way.

This disagreement on Sue's bias persists, even though bias is state-independent, and even though Pam and Sue agree on the exact value of Sue's estimate, so that there is nothing more Pam could learn about Sue's estimate (other than perhaps learning that they never actually agreed). If we can consider this to be a situation of "pure" computational disagreement, then we can conclude that, for Bayesian wannabes with a common prior, persistent disagreements due to combinations of differing information and computation "reduce" to persistent disagreements based purely on differing computation, in the sense that you can't have the first kind without the second kind also being present.

We now explain our notation, give some examples, and then derive the above results.

2. THE MODEL

2.1. *Bayesian Wannabes*

Consider a finite set of possible states $\omega \in \Omega$, and real-valued random variables over these states, such as $X(\omega) \in [\underline{X}, \bar{X}]$. A *decision agent* i is characterized by an estimation operator $\tilde{\mathrm{E}}_{i\omega}$, which for every random variable $X(\omega)$ produces another random variable $\tilde{X}_i(\omega) = \tilde{\mathrm{E}}_{i\omega}[X(\omega)]$, which is agent i's estimate in state ω of variable X.

A *Bayesian* decision agent i has information described by a partition I_i of Ω, assigns a prior weight $\mu_i(\omega)$ to the states, and uses an exact expected value decision operator $\tilde{\mathrm{E}}_{i\omega}[X] = \mathrm{E}_{\mu_i}[X \mid I_i(\omega)]$, where

$$\mathrm{E}_{\mu_i}[X \mid I_i(\omega')] = \frac{\sum_{\omega \in I_i(\omega')} X(\omega)\mu_i(\omega)}{\sum_{\omega \in I_i(\omega')} \mu_i(\omega)} \tag{1}$$

When Ω is large, such expected values can be very difficult to compute.

A *Bayesian wannabe* i is a decision agent who can make sense of counterfactual statements regarding what her estimates would be if she had sufficient computational abilities to be a Bayesian. That is, she imagines that with sufficient computational power to consider the matter, she could eventually reach a reflective equilibrium about which prior $\mu_i(\omega)$ is most rational, could eventually combine all the implications of all her state clues into a consistent partition element $I_i(\omega)$, and could then compute exact expected value estimates $X_i(\omega) = \mathrm{E}_{\mu_i}[X \mid I_i(\omega)]$.

A Bayesian wannabe can thus make sense of her *error*, $e_{i\omega}[X] = \tilde{X}_i(\omega) - X_i(\omega)$, the difference between her actual estimate $\tilde{X}_i(\omega)$ and the estimate $X_i(\omega)$ she would use if she had the computational abilities to be a Bayesian. We assume that agent i's estimation oper-

ator $\tilde{\mathrm{E}}_{i\omega}$ is measurable with respect to information I_i, so that for all variables X and $\omega' \in I_i(\omega)$, $\tilde{X}_i(\omega') = \tilde{X}_i(\omega)$.

2.2. *Bias*

An agent's *bias* regarding X on a set S is her average error on that set, $\bar{e}_i[X|S] = \mathrm{E}_{\mu_i}[e_{i\omega}[X] \mid S]$. To reduce the magnitude of her bias,[2] a Bayesian wannabe might adjust a *calibration* $c_{i\omega}[X]$, so that $\tilde{X}_i(\omega) = \tilde{X}_i^0(\omega) - c_{i\omega}[X]$, where $\tilde{X}_i^0(\omega)$ is her estimate with zero calibration.

We can show the following (non-trivial proofs in Proofs Appendix).

LEMMA 1. *When $c_{i\omega}[X] = c$ for all $\omega \in S$ and S is a union of elements of I_i, the c which minimizes $\mathrm{E}_{\mu_i}[(\tilde{X}_i(\omega) - X(\omega))^2 \mid S]$ (or $\mathrm{E}_{\mu_i}[e_{i\omega}^2[X] \mid S]$) sets $\bar{e}_i[X|S] = 0$.*

Let us say that agent i at state ω *expects that she is unbiased* regarding X on S when $\tilde{\mathrm{E}}_{i\omega}[\bar{e}_i[X|S]] = 0$. When a Bayesian wannabe at state ω is aware of Lemma 1 and makes calibration adjustments to minimize her average squared error, it seems reasonable for her to expect that she is unbiased regarding X on her calibration set $D_i^X(\omega)$, the set of states where she makes the same calibration adjustment, so that $c_{i\omega'}[X] = c_{i\omega}[X]$ for all $\omega' \in D_i^X(\omega)$. It also seems reasonable for her to expect that she is unbiased regarding X on any set S that is a union of elements of the partition D_i^X. (We assume D_i^X coarsens I_i.)

2.3. *Agreeing to Disagree*

Standard notions of belief, agreement, and disagreement, can be generalized to apply to a Bayesian wannabe. Such an agent *q-estimates* an event E within the estimation set

$$\tilde{B}_i^q(E) = \{\omega \mid \tilde{\mathrm{E}}_{i\omega}[\mu_i(E \mid I_i(\omega))] \geq q\},$$

and the *accuracy* of agent i on this estimation set is $\mu_i(E \mid \tilde{B}_i^q(E))$. (Bayesians always have an accuracy of at least q in their q-estimation, which for Bayesians is called q-belief.)

A set N of agents *q-agree* that E within any C where

$$C \subset \cap_{i \in N} \tilde{B}_i^q(C \cap E). \tag{2}$$

We will call such an event C a *q-agreement* event of E, and call Equation (2) its agreement equation. Note that this concept has been introduced for Bayesians under other names (Monderer and Samet, 1996; Borgers, 1994).[3]

We will use several concepts of disagreement. A weak concept is that agents i, j are said to *ϵ-disagree* about X when $\tilde{X}_i \geq \tilde{X}_j + \epsilon$. (They are said to disagree when $\epsilon > 0$.) Let $\{\omega \mid \tilde{X}_i(\omega) \geq \tilde{X}_j(\omega)+\epsilon\}$ be the i, j *ϵ-disagreement event* about X. A strong concept is that agents i, j *α, β-disagree* about X when $\tilde{X}_i \geq \alpha$ and $\beta \geq \tilde{X}_j$. Note that a strong α, β-disagreement implies a weak (α–β)-disagreement, and that these definitions are not symmetric between i, j.

We can thus say that i, j *q-agree to ϵ-disagree* about X if they q-agree regarding i, j's ϵ-disagreement event about X. We can similarly define when i, j *q-agree to α, β-disagree* about X. (When $\epsilon > 0$, the agents i, j q-agree to weakly disagree, and when $\alpha > \beta$ they q-agree to strongly disagree.) Note that in any agreement to disagree, each agent has the option to adjust her estimate in the direction of (her estimate of) the other agent's estimate.

2.4. *Agreeing to Disagree about Computation*

Let us say that agents i, j *disagree about the computation* of X when they disagree about X and X is state-independent, and that they q-agree to disagree about the computation of X if they q-agree that they disagree about the computation of X.

How relevant is private information to an agreement to disagree about computation? Let us say that agents i, j *q-agree to clearly α, β-disagree about the computation* of X if they q-agree that $\tilde{X}_i = \alpha$ and $\tilde{X}_j = \beta$, for a state-independent X. Here it is agreed that X, $\tilde{X}_i$, and $\tilde{X}_j$ are *all* state independent, and so knowing the state of the world does not tell any agent directly about X, about her own value of the computation of X, or about the other agent's value for the computation of X.

Let us say that agents i, j *q-agree to strongly α, β-disagree about the computation* of X if they q-agree that $\tilde{X}_i \geq \alpha$ and $\tilde{X}_j = \beta$, for a state-independent X. Here $\tilde{X}_i$, but not $\tilde{X}_j$, is sometimes state-dependent, but only by sometimes moving even farther away from $\tilde{X}_j$. In this case, knowing the state of the world does not tell agent i directly about X, nor about the agent j's value for the computation

TABLE I

One State Example

ω	X	$\tilde{X}_1$	$\tilde{X}_2$	$e_1[X]$	$\bar{e}_1$	$\tilde{E}_1[\bar{e}_1]$	$\tilde{E}_2[\bar{e}_1]$
1	π	22/7	3.14	0.00126	0.00126	0	0.003

of X. While not as clear as the previous case, this still seems a strong candidate for a reasonably "pure" computational agreement to disagree, and this is the case that will be demonstrated in this paper to follow from any agreement to disagree.

It is admittedly not obvious that these cases are "clear" and "strong" as I have described them. This paper will not further consider this issue, however.

2.5. *Examples*

Table I describes an extremely simple example, where agents believe there is only one possible self-consistent description of a possible reality. So $\Omega = \{1\}$, $\mu_i(1) = 1$, and $I_i = \{\{1\}\}$. Two agents, named 1 and 2, each use a different heuristic to estimate the mathematical constant π. Agent 1 estimates that she has zero error in her estimation that $\pi \approx 22/7$, while agent 2, who uses a different estimate $\pi \approx 22/7$, estimates that agent 1 has an error of 0.003. Agents 1, 2 1-agree to clearly 22/7, 3.14-disagree about the computation of π. Since there is known to be only one possible state here, estimates of error are the same as estimates of bias, or average error. All of these features remain if this example is modified to have two states, each with the same row of values shown in Table I, and with $I_1 = \{\{1, 2\}\}$, $I_2 = \{\{1\}, \{2\}\}$, and any μ_i.

Table II describes a five state example where agents 1, 2 are said to 0.94-agree to 15-disagree about a variable $X \in [0, 100]$. That is, two Bayesian wannabes have a common prior $\mu(\omega)$ and differing information partitions $I_1 = \{\{1\}, \{2, 3\}, \{4, 5\}\}$ and $I_2 = \{\{1, 2\}, \{3, 4\}, \{5\}\}$. Consider the event $C = \{2, 3, 4\}$ and agent estimates $\tilde{q}_i(\omega) = \tilde{E}_{i\omega}[q_i]$ about whether they are in event C, where $q_i = \mu(C|I_i(\omega))$. Everywhere in $B_1 = \{2, 3, 4, 5\}$ agent 1 satisfies $\tilde{q}_1 \geq 0.94$, and everywhere in $B_2 = \{1, 2, 3, 4\}$, agent 2 satisfies $\tilde{q}_2 \geq 0.95$. Since $C \subset B_1 \bigcap B_2$, then C is an agreement event. And

TABLE II

Five State Example

ω	μ	$\tilde{q}_1$	$\tilde{q}_2$	X	X_1	X_2	$\tilde{X}_1$	$\tilde{X}_2$	$e_1[X]$	$\bar{e}_1$	$\tilde{E}_1[\bar{e}_1]$	$\tilde{E}_2[\bar{e}_1]$	$\hat{\epsilon}$	$\tilde{p}_2$
1	0.02	0.01	0.95	0	0.0	70.6	5	55	5.0	8.57	0	10	8.6	0.97
2	0.32	0.99	0.95	75	62.5	70.6	70	55	7.5	8.57	0	10	8.6	0.97
3	0.32	0.99	0.99	50	62.5	37.5	70	25	7.5	8.57	0	12	10.7	0.98
4	0.32	0.94	0.99	25	29.4	37.5	40	25	10.6	8.57	0	12	10.7	0.98
5	0.02	0.94	0.02	100	29.4	100	40	95	10.6	8.57	0	5	8.6	0.97

TABLE III

Five State, State-Independent Variable Example

ω	μ	$\tilde{q}_1$	$\tilde{q}_2$	X	$\tilde{X}_1$	$\tilde{X}_2$	$e_1[X]$	$\bar{e}_1$	$\tilde{E}_1[\bar{e}_1]$	$\tilde{E}_2[\bar{e}_1]$
1	0.02	0.01	0.95	π	3.1	3	-0.042	-0.0009	0	0.04
2	0.32	0.99	0.95	π	3.14	3	-0.002	-0.0009	0	0.04
3	0.32	0.99	0.99	π	3.14	3.1	-0.002	-0.0009	0	0.03
4	0.32	0.94	0.99	π	3.142	3.1	0.0004	-0.0009	0	0.03
5	0.02	0.94	0.02	π	3.142	3.14	0.0004	-0.0009	0	0.01

since $\tilde{X}_1 - \tilde{X}_2 \geq 15$ everywhere in C, they are agreeing to disagree on X in the weak sense. (They do not 0.94-agree to disagree in the strong sense, since there are no α, β satisfying $\tilde{X}_1 \geq \alpha > \beta \geq \tilde{X}_2$ for all $\omega \in C$.)

These agents make many errors, and would not even weakly disagree if they were Bayesians. Agent 1's errors in estimating X, averaged across B_1, constitute a bias of 8.57 on that set. (Agent 2 has a bias of -13.53 on B_2.) Agent 1 always estimates herself to be unbiased, but agent 2 estimates that agent 1 is biased. While agent 2's estimate of agent 1's bias varies from state to state, it is always at least 10. So this is an example where agents 2, 1 are said to 0.94-agree to strongly 10, 0-disagree about the computation of $\bar{e}_1$.

As will be explained in the next section, there is a simple formula which agents can use to calculate $\hat{\epsilon}$, which is a lower bound on agent 2's estimate of agent 1's average bias. This formula takes as input the amount of the basic weak disagreement, in this case $\epsilon = 15$, and $\tilde{p}_2$, which is agent 2's estimate of the minimum among the accuracy of agent 1 on B_1, and the accuracy of 2 on B_2. (These accuracies are both equal to 0.98 in this example.) If the agents agree that $\epsilon \geq 15$ and that $\tilde{p}_2 \geq 0.97$, they can agree that $\hat{\epsilon} = 8.6$, and so can 0.94-agree to strongly 8.6, 0-disagree on the computation of agent 1's bias.

Table III describes a five state example with the same prior, information partitions, and estimates q_i as in Table II. The difference here is that the random variable is the mathematical constant π, which is the same in all states. What differs across states is the number of digits in each agent's approximation of π. In this example,

agents 1, 2 0.94-agree to 0.04-disagree about the computation of π, and agents 2, 1 0.94-agree to strongly 0.03, 0-disagree about the computation of $\bar{e}_1[\pi]$. Agents 1, 2 do not 0.94-agree to strongly disagree about the computation of π, however.

This example may help illustrate some issues in attributing disagreements about state-independent variables to private information. While this disagreement is in part due to the fact that agents have access to differing numbers of digits of π in different states, it seems more fundamentally due to the fact that the agents seem to not consider the possibility that the other agent might have access to more digits. It seems that to agree to disagree about π, the agents must agree to disagree about who has more digits. And as we will see, this requires that they agree to disagree about their average bias in estimating who has more digits.

It should be noted that cases of uncertainty about the value of state-independent variables can be re-described as cases of uncertainty about state-dependent variables, by expanding the state space to include "impossible possible worlds" (Hintikka, 1975). So uncertainty about π could be represented using states where the value of π varies across the states. It is not clear, however, how best to do this regarding variables defined in terms of the existing state space, such as an average of random variable over the state space.

3. ANALYSIS

Let us now focus attention on a particular random variable X, and assume that in some states two Bayesian wannabes, named 1 and 2, q-agree to ϵ-disagree about X, with $\epsilon > 0$. Thus there is a non-empty 1, 2 ϵ-disagreement event E, within which $\tilde{X}_1 \geq \tilde{X}_2 + \epsilon$. We assume the agents have a common prior, $\mu_1 = \mu_2 = \mu$, and focus on a particular uniquely defined non-empty q-agreement event C, such as the one with the largest prior weight. Thus C and E satisfy Equation (2).

To further simplify our notation, let $A = C \cap E$ be the analysis set we will focus on, let $B_i = \tilde{B}_i^q(A)$ be its estimation sets, let $\bar{e}_i = \bar{e}_i[X|B_i]$ be each agent's bias regarding X on those sets, and let $p_i = \mu(A \mid B_i)$ be each agent's accuracy on those estimation sets. Note that if agent i considers B_i to be a union of her calib-

ration sets $D_i^X(\omega)$, that is, if she considers her calibration of X as she considers her disagreement, then she should expect that she is unbiased regarding X on B_i, so that $\tilde{\mathrm{E}}_{i\omega}[\bar{e}_i] = 0$.

We will now consider under what conditions agreeing to disagree about X implies that Bayesian wannabes disagree, or agree to disagree, about their biases $\bar{e}_i$. Note that since the sets C and E are defined in a state independent manner, and since the sets B_i were defined uniquely in terms of these sets, biases $\bar{e}_i$ have also been defined to be state-independent random variables. Thus agreeing to α, β-disagree about such biases is agreeing to disagree about computation.

If the agents are accurate ($p_i \approx 1$) in estimating that they agree to disagree, the B_i sets cannot differ much from the analysis set A. This suggests that the difference between the agent's biases $\bar{e}_i$ on B_i is nearly the average difference between the agent's estimates on A.

LEMMA 3. $\bar{e}_1/p_1 - \bar{e}_2/p_2 = \mathrm{E}[\tilde{X}_1 - \tilde{X}_2 \mid A] + \mathrm{E}[\tilde{X}_1 - X \mid B_1 \setminus A](1 - p_1)/p_1 - \mathrm{E}[\tilde{X}_2 - X \mid B_2 \setminus A](1 - p_2)/p_2$.

Since $\tilde{X}_1 - \tilde{X}_2 > \epsilon$ everywhere in A, $\bar{e}_1$ must be positive if $\bar{e}_2 = 0$ and ϵ is not too small. Let us define $p_0 = \min(p_1, p_2)$, $\Delta X = \bar{X} - \underline{X}$, and $\hat{\epsilon}(\epsilon, p) = p\epsilon - 2(1 - p)\Delta X$. Using these definitions, we can show that agent 2 being unbiased implies a lower bound on agent 1's bias.

LEMMA 4. $\epsilon \geq 0$ *and* $\bar{e}_2 = 0$ *imply* $\bar{e}_1 \geq \hat{\epsilon}(\epsilon, p_0)$.

This lower bound is positive for $p_0 \approx 1$ near one, since then $\hat{\epsilon}(\epsilon, p_0) \approx \epsilon$.

If at state ω, Bayesian wannabe 2 is aware of Lemma 4, and expects that she is unbiased regarding X on B_2, she should want to keep her estimates consistent with the inequality constraint given by Lemma 4. Since $\hat{\epsilon}(\epsilon, p)$ is linear in p, this should result in[4]

$$\tilde{\mathrm{E}}_{2\omega}[\bar{e}_1] \geq \hat{\epsilon}(\epsilon, \tilde{p}_2), \tag{3}$$

for $\tilde{p}_2 = \tilde{\mathrm{E}}_{2\omega}[p_0]$.

But if at state ω Bayesian wannabe 1 expects that she is unbiased on B_1, we have

$$\tilde{E}_{1\omega}[\bar{e}_1] = 0. \tag{4}$$

Equations (3) and (4) directly imply our first main result.

THEOREM 1. *Regarding agents* $1, 2$ *q-agreeing to* ϵ*-disagree about X,* if *at some* ω *agent* 1 *satisfies Equation (4) (expecting she is unbiased regarding X on* B_1*), and agent* 2 *satisfies Equation (3) for some* $\tilde{p}_2$*,* then *at* ω*, agents* $2, 1$ $\hat{\epsilon}(\epsilon, \tilde{p}_2), 0$*-disagree about the computation of* $\bar{e}_1$.

Note that the magnitude $\bar{\epsilon}$ of the resulting disagreement depends on $\tilde{p}_2$, agent 2's perception of their accuracy of agreement, but not on q, their common confidence in agreement. Theorem 1 shows that if two Bayesian wannabes each think they are unbiased and agree to weakly disagree by a large enough amount about any random variable X, they must also disagree about one of their biases. And since bias has been defined in a state-independent manner, this implies that they disagree about a computation.

To conclude that two Bayesian wannabes *agree* to disagree about a computation, however, we must assume more. We can simply assume that they agree that they accept the assumptions of Theorem 1.

THEOREM 2. If *agents* $1, 2$ *q-agree (at agreement C)*

1. *that they* ϵ*-disagree about X (for* $\epsilon > 0$*),* and
2. *that they satisfy Equations 3 and 4 for agreed on values of* ϵ *and* $\tilde{p}_2$*,*

then *within C, agents* $2, 1$ *q-agree to strongly* $\hat{\epsilon}(\epsilon, \tilde{p}_2), 0$*-disagree about the computation of* $\bar{e}_1$.

An example to which Theorem 2 applies is described in Table II. Theorem 2 is our main conclusion.[5] It considers two agents who agree both that they weakly disagree by a large enough amount regarding any real-valued random variable, *and* that the assumptions of Theorem 1 holds, which are that agent 1 expects herself to be unbiased, and that agent 2 keeps her estimates consistent with Equation (3) for some agreed-on value of $\tilde{p}_2$, presumably because agent 2 is aware of Lemma 4 and estimates that she is unbiased.

Given this agreement, these agents *must* agree to disagree by a certain amount about agent 1's average error, which is a state-independent random variable. Moreover, they must agree that agent 1's estimate of this variable is also state independent. This is a sort of

agreement to disagree that we earlier argued seems a strong candidate for a case of "pure" computational disagreement, where differing private information is largely irrelevant.

4. CONCLUSION

Since Bayesians with a common prior cannot agree to disagree, to what can we attribute persistent human disagreement? We can generalize the concept of a Bayesian to that of a Bayesian wannabe, who makes computational errors while attempting to be Bayesian. Agreements to disagree can then arise from pure differences in priors, or from pure differences in computation, but it is not clear how rational these disagreements are. Disagreements due to differing information seem more rational, but for Bayesians disagreements cannot arise due to differing information alone.

Can we explain persistent disagreement as due to an intrinsic combination of differing information and differing something else? Given any agreement to disagree due to a combination of differing priors and information, and no computational errors, we can easily find a disagreement purely due to differing priors, so this case seems to reduce to pure prior-based disagreements. Similarly, this paper shows that given any agreement to disagree with differing information and computational errors, but common priors, we can find an agreement to disagree about a state-independent random variable. And this agreement to disagree seems of a strong sort that is plausibly described as a case of pure computational disagreement, though this is admittedly not an obvious interpretation.

If the interpretation proposed here is correct, then it seems that to the extent that pure computational and pure prior-based persistent disagreements are irrational, any persistent disagreement is irrational. This would make it much harder to reconcile the ubiquitous persistent disagreements around us with a presumption of human rationality.

5. PROOFS APPENDIX

LEMMA 1. *When $c_{i\omega}[X] = c$ for all $\omega \in S$ and S is a union of elements of I_i, the c which minimizes* $\mathrm{E}_{\mu_i}[(\tilde{X}_i(\omega) - X(\omega))^2 \mid S]$ *(or* $\mathrm{E}_{\mu_i}[e^2_{i\omega}[X] \mid S]$*) sets* $\bar{e}_i[X|S] = 0$.

Proof. Since $e_i = \tilde{X}_i - X_i$, then $(\tilde{X}_i - X)^2 = (e_i + (X_i - X))^2 = e_i^2 + (X_i - X)^2 + 2e_i(X_i - X)$. The second term on the right is independent of c, and the expectation of the third term on the right over S vanishes because e_i is constant over each $I_i(\omega) \subset S$ and $E[X_i - X \mid I_i(\omega)] = 0$, which follows from Equation 1. Thus to minimize $E[(\tilde{X}_i - X)^2 \mid S]$ is to minimize $E[e_i^2 \mid S]$. We can write $e_i = \tilde{X}_i - X_i = (\tilde{X}_i^0 - c_i) - X_i = m_i - c_i$, where $m_i = \tilde{X}_i^0 - X_i$. We can further write $e_i = m_i - c_i = (m_i - \bar{m}_i) + (\bar{m}_i - c_i)$, where $\bar{m}_i = E[m_i \mid S]$. Then $e_i^2 = (m_i - \bar{m}_i)^2 + (\bar{m}_i - c_i)^2 + 2(m_i - \bar{m}_i)(\bar{m}_i - c_i)$. But the expectation of the third term here over S vanishes by the definition of $\bar{m}_i$, the first term is independent of c_i, and the second term is minimized by $c_i = \bar{m}_i$. So $\bar{e}_i = \bar{m}_i - c_i = 0$. □

LEMMA 2. *Common p-belief implies $(2p-1)$-agreeing, and p-agreeing implies common p-belief.*

Proof. Regarding the second claim, $C \subset B_i^p(C \cap E)$ implies both $C \subset B_i^p(C)$ and $C \subset B_i^p(E)$ due to the general relation that $B_i^p(S) \subset B_i^p(S')$ whenever $S \subset S'$. Regarding the first claim, for all ω in a common set C, $\mu(C \mid I_i(\omega)) \geq p$ and $\mu(E \mid I_i(\omega)) \geq p$. Defining $a_1 = \mu(C \cap E \mid I_i(\omega))$, $a_2 = \mu(C \setminus E \mid I_i(\omega))$, $a_3 = (E \setminus C \mid I_i(\omega))$, and $a_4 = 1 - a_1 - a_2 - a_3$, we thus have $a_1 + a_2 \geq p$ and $a_1 + a_3 \geq p$. This implies $a_3 + a_4 \leq 1 - p$ and $a_2 + a_4 \leq 1 - p$ which implies $1 - a_1 = a_2 + a_3 + a_4 \leq a_2 + a_3 + 2a_4 \leq 2(1 - p)$ so that $a_1 = \mu(C \cap E \mid I_i(\omega) \geq 2p - 1$ for all $\omega \in C$. □

LEMMA 3. $\bar{e}_1/p_1 - \bar{e}_2/p_2 = \mathrm{E}[\tilde{X}_1 - \tilde{X}_2 \mid A] + \mathrm{E}[\tilde{X}_1 - X \mid B_1 \setminus A](1 - p_1)/p_1 - \mathrm{E}[\tilde{X}_2 - X \mid B_2 \setminus A](1 - p_2)/p_2$.

Proof. Since $e_i = \tilde{X}_i - X_i$, we have $\mathrm{E}[e_1 \mid A] - \mathrm{E}[e_2 \mid A] = \mathrm{E}[\tilde{X}_1 - \tilde{X}_2 \mid A] - \mathrm{E}[X_1 - X_2 \mid A]$. The strategy of proof is to find expressions for each of these terms, substitute them, and then solve for $\bar{e}_1/p_1 - \bar{e}_2/p_2$.

First, rearranging equation 1 implies that $\mathrm{E}[X_i \mid S] = \mathrm{E}[X \mid S]$ for any S which is a union of I_i members. And B_i must be a union

of I_i members since $\tilde{\mathrm{E}}_{i\omega'} = \tilde{\mathrm{E}}_{i\omega}$ is the same for all $\omega' \in I_i(\omega)$. Thus $\mathrm{E}[X_i \mid B_i] = \mathrm{E}[X \mid B_i]$. Since $p_i = \mu(A \mid B_i)$, we also have

$$\mathrm{E}[f \mid B_i] = \mathrm{E}[f \mid A]p_i + \mathrm{E}[f \mid B_i \setminus A](1 - p_i) \tag{5}$$

for any f. Using $f = X$ and $f = X_i$, we can then solve for $\mathrm{E}[X_i \mid A] = \mathrm{E}[X \mid A] + E[X - X_i \mid B_i \setminus A](1 - p_i)/p_i$, which implies

$$\mathrm{E}[X_1 - X_2 \mid A] = E[X - X_1 \mid B_1 \setminus A](1 - p_1) \\ /p_1 - E[X - X_2 \mid B_2 \setminus A](1 - p_2)/p_2.$$

Second, using $f = e_i$ in equation 5 yields $\mathrm{E}[e_i \mid A] = \bar{e}_i/p_i - \mathrm{E}[e_i \mid B_i \setminus A](1 - p_i)/p_i$. Finally, we can leave $\mathrm{E}[\tilde{X}_1 - \tilde{X}_2 \mid A] = \mathrm{E}[\tilde{X}_1 - \tilde{X}_2 \mid C \cap E]$ alone, as this must be at least ϵ by the definition of E. Substituting into the original equation, noting that $(X_i - X) + e_i = \tilde{X}_i - X$, and solving for $\bar{e}_1/p_1 - \bar{e}_2/p_2$ gives the result. □

LEMMA 4. *$\epsilon \geq 0$ and $\bar{e}_2 \geq -\delta \leq 0$ imply $\bar{e}_1 \geq \hat{\epsilon}(\epsilon, p_0) - \delta$.*

Proof. Since $\tilde{X}_1 - \tilde{X}_2 \geq \epsilon$ everywhere in E, and $A \subset E$, the first right side term in Lemma 3's equation is at least ϵ. For $\epsilon \geq 0$ the most negative imaginable case for this right side is where $(B_1 \setminus A) \cap (B_2 \setminus A) = \emptyset$, with $\tilde{X}_1 = \underline{X}$ and $X = \bar{X}$ on $B_1 \setminus A$, and $\tilde{X}_2 = \overline{X}$ and $X = \underline{X}$ on $B_2 \setminus A$. This gives

$$\bar{e}_1/p_1 \geq \bar{e}_2/p_2 + \epsilon - \Delta X((1 - p_1)/p_1 + (1 - p_2)/p_2).$$

Multiplying this equation by p_1, the most negative case for the last two right side terms is $p_1 = p_2 = p_0$, and when $\bar{e}_2 = -\delta \leq 0$, the most negative case for the first right side term is $p_1 = 1$, $p_2 = p_0$. This implies the result. □

6. ACKNOWLEDGEMENTS

I thank Kim Border, Pierpaolo Battigalli, Colin Camerer, Erik Eyster, Paolo Ghirardato, John Ledyard, Stephen Morris, Richard McKelvey, Scott Page, Ariel Rubenstein, Illya Segal, Matt Spitzer, Simon Wilkie, several anonymous referees, and participants in a U.C. Berkeley Economic Theory Seminar for helpful comments.

A version of this paper appeared as a chapter in my 1998 Caltech social science Ph.D. thesis. I thank the Center for Study of Public Choice and the Mercatus Center for financial support.

NOTES

1. In this paper, "computation" is whatever agents do to reduce calculation "errors" as we have defined them.
2. When humans suspect their judgment is biased, they often attempt to correct it (Wegener et al., 1998).
3. An related concept is *common p-estimation* of event E, holding at states within any event C where $C \subset \tilde{B}_i^p(C)$ and $C \subset \tilde{B}_i^p(E)$ for all $i \in N$. (For Bayesians, this has been called common p-belief (Monderer and Samet, 1989).) For Bayesians, agreeing and common estimation are very similar concepts.

   LEMMA 2. *For Bayesians, common p-estimation implies* $(2p-1)$*-agreeing, and p-agreeing implies common p-estimation.*
4. As stated, Equation (3) allows no rounding error in the agent 2's maintenance of Lemma 4's constraint. Such rounding error can be allowed by generalizing $\hat{\epsilon}(\epsilon, p)$ to $\hat{\epsilon}(\epsilon, p, \delta) = \hat{\epsilon}(\epsilon, p) - \delta$. Theorems 1 and 2 then trivially generalize to this new expression for $\hat{\epsilon}$, for some rounding error δ. Similar corrections can allow deviations from agents expecting themselves to be exactly unbiased; our conclusions do not fragily depend on such assumptions.
5. Note that Theorem 2 can easily be generalized to distinguish the q'-agreement event C' regarding disagreeing on $\bar{e}_1$ from the q-agreement event C regarding disagreeing about X. There are no obvious constraints relating C and C' and q and q'.

REFERENCES

Aumann, R. (1976), Agreeing to disagree, *The Annals of Statistics* 4(6), 1236–1239.

Bonanno, G. (1997), Agreeing to disagree: a survey, Tech. rep. 97/18, University of California, Davis, Department of Economics.

Bonanno, G. and Nehring, K. (1999), How to make sense of the common prior assumption under incomplete information, *International Journal of Game Theory* 28(3), 409–34.

Borgers, T. (1994), Weak dominance and approximate common knowledge, *Journal of Economic Theory* 64, 265–276.

Feinberg, Y. (2000), Characterizing common priors in the form of posteriors, *Journal of Economic Theory* 91(2), 127–79.

Geanakoplos, J. (1989), Game theory without partitions, and applications to speculation and consensus, Tech. rep. 914, Yale Cowles Foundation.

Geanakoplos, J. (1994), Common knowledge, in R. Aumann and S. Hart, S. (eds.), *Handbook of Game Theory, Vol. 2* (pp. 1438–1496). Elsevier Science, Amsterdam.

Hanson, R. (1998), Consensus by identifying extremists, *Theory and Decision* 44(3), 293–301.

Hanson, R. (2002), Disagreement is unpredictable, *Economics Letters 77*, 365–369.

Hintikka, J. (1975), Impossible possible worlds vindicated, *Journal of Philosophical Logic* 4, 475–484.

Lipman, B. (1995), Information processing and bounded rationality: a survey, *Canadian Journal of Economics* 28(1), 42–67.

McKelvey, R. and Page, T. (1986), Common knowledge, consensus, and aggregate information, *Econometrica* 54(1), 109–127.

Megiddo, N. (1989), On computable beliefs of rational machines, *Games and Economic Behavior* 1, 144–169.

Milgrom, P. and Stokey, N. (1982), Information, trade and common knowledge, *Journal of Economic Theory* 26, 17–27.

Monderer, D. and Samet, D. (1989), Approximating common knowledge with common beliefs, *Games and Economic Behavior* 1, 170–190.

Monderer, D. and Samet, D. (1996), Proximity of information in games with incomplete information, *Mathematics of Operations Research* 21, 707–725.

Morris, S. (1994), Trade with hetereogeneous prior beliefs and asymmetric information, *Econometrica* 62(6), 1327–1347.

Morris, S. (1996), The logic of belief and belief change: a decision theoretic analysis, *Journal of Economic Theory* 69, 1–23.

Morris, S. (1999), Approximate common knowledge revisited, *International Journal of Game Theory* 28(3), 385–408.

Neeman, Z. (1996a), Approximating agreeing to disagree results with common p-beliefs, *Games and Economic Behavior* 12, 162–164.

Neeman, Z. (1996b), Common beliefs and the existence of speculative trade, *Games and Economic Behavior* 16, 77–96.

Neilsen, L. T., Brandenburger, A., Geanakoplos, J., McKelvey, R., and Page, T. (1990), Common Knowledge of an aggregate of expectations, *Econometrica* 58(5), 1235–1239.

Rubinstein, A. and Wolinsky, A. (1990), On the logic of 'agreeing to disagree' type results, *Journal of Economic Theory* 51, 184–193.

Samet, D. (1990), Ignoring ingorance and agreeing to disagree, *Journal of Economic Theory* 52, 190–207.

Sebenius, J. and Geanakoplos, J. (1983), Don't bet on it: contingent agreements with asymmetric information, *Journal of the American Statistical Association* 78(382), 424–426.

Shin, H. S. and Williamson, T. (1994), Representing the knowledge of Turing machines, *Theory and Decision* 37, 125–146.

Sonsino, D. (1995), Impossibility of speculation theorems with noisy information, *Games and Economic Behavior* 8, 406–423.

Wegener, D. T., Petty, R. E., and Dunn, M. (1998), The metacognition of bias correction: naive theories of bias and the flexible correction model, in V. Y. Yzerbyt, G. Lories and B. Dardenne, (eds.), *Metacognition, Cognitive and Social Dimensions* (pp. 202–227). Sage Publications, London.

Address for correspondence: Department of Economics, George Mason University, Carow Hall, MSN 1D3, Fairfax VA 22030, USA
(e-mail: rhanson@gmu.edu; http://hanson.gmu.edu 703-993-2326; Fax: +1-703-993-2323)