Disagreement as Self-Deception About Meta-Rationality

Tyler Cowen
Robin Hanson

Department of Economics
George Mason University
Fairfax, VA 22030
tcowen@gmu.edu
rhanson@gmu.edu

July 8, 2002

*The authors wish to thank Curt Adams, Nick Bostrom, James Buchanan, Bryan Caplan, Wei Dai, Hal Finney, Mark Grady, Patricia Greenspan, Kevin Grier, Robin Grier, Hans Haller, Claire Hill, Mary Hirschfeld, Dan Houser, Stephen Hsu, Michael Huemer, Maureen Kelley, Tom Morrow, William Nelson, Mark Notturno, David Schmidtz, Susan Snyder, Aris Spanos, Damien Sullivan, Daniel Sutter, Alexander Tabarrok, William Talbott, Nicolaus Tideman, Eleizer Yudkowsky, and participants of the Virginia Tech economics department seminar for useful comments and discussion.

ABSTRACT

Honest truth-seeking agents should not agree to disagree. This result is robust to many perturbations. Such agents are "meta-rational" when they act as if they realize this result. The ubiquity of disagreement, however, suggests that very few people, academics included, are very meta-rational. Instead, we seem self-deceived in thinking ourselves to more meta-rational than others. Since alerting us to this fact does not much change our behavior, we must not really want to know the truth.

I. Introduction

Disagreement is a ubiquitous feature of human life. Most people disagree frequently, especially on politics, morality, religion, and relative abilities, and often with people they consider intelligent and honest. Virtually any two intelligent people can find topics of disagreement quickly. Disagreements usually persist, and often become stronger, when people become mutually aware of them. Nor is disagreement usually embarrassing; it is often worse to be considered a "fence-sitter" without distinctive opinions.

People believe that many of their disagreements are about what is objectively true, rather than how they each feel or use words, and in such cases people usually consider themselves to be *truth-seekers*. That is, they believe their opinions to be the best estimate of the truth they can achieve given their information and effort, and they believe that they are honestly stating their opinions.[1]

Yet in theory these disagreements are irrational for such truth-seekers. Robert Aumann (1976) first developed general results about the irrationality of "agreeing to disagree." He showed that if two or more Bayesians would believe the same thing given the same information (i.e., have "common priors"), and if they are mutually aware of each other's opinions (i.e., have "common knowledge"), then those individuals cannot knowingly disagree. Merely knowing someone else's opinion turns out to provide something like a sufficient statistic of everything that person knows. In particular it provides enough information, in reflective equilibrium, to eliminate any differences of opinion due to differing information. This result turns out to be robust to many variations, and reflects a general principle of "meta-rationality": when seeking to estimate the truth, you should realize you might be wrong and others right. Since the opinions of others may be based

[1] Note that we here consider only individual level truth-seeking and rationality, and do not consider how individual deviations from this ideal might aid the overall advance of knowledge (Everett 2001; Kitcher 1990). Note also that we will offer only necessary, but not sufficient, conditions for truth-seeking. We do not delve into the murky philosophical waters of "justified belief." A complete account of truth-seeking behavior would likely include many other stipulations and properties.

on evidence that you do not possess, you should regard their opinions as highly relevant evidence.

This logic holds whenever people believe there are objectively correct answers to the question at hand, and whenever opinions are considered to be estimates of such objective answers (i.e., "expected values of random variables"). It holds for facts both specific and general, both hard and easy to verify. It covers the age of a car, the correctness of quantum mechanics, whether God created the universe, and which political candidate is more likely to induce prosperity. If there are objectively correct answers to questions like "is this action moral?" then this logic holds for such questions of value and morality. This logic, however, need not cover questions of taste. People should agree on which of them like French fries, but some can like French fries while others do not.[2]

Reality, however, is at odds with this theory. So we ask: Why do people disagree? Our analysis, based on simple observations about the phenomena of disagreement, suggests some specific and striking conclusions. It suggests that people are not truth-seekers, that they are self-deceived about this fact, and that they are self-deceived in over-estimating their relative meta-rationality. Furthermore people seem to want to be this way. We refer not only to people in general, but also to most academics, and to ourselves.

The paper proceeds as follows. We first present the basic theory of disagreement, and how it has been generalized, paying special attention to the rationality of non-common

[2] The Bayesian literature is cited in more detail throughout the paper. In the philosophy literature, Everett (2001) argues that a rational individual should often abandon his or her own estimates of truth for the estimates of others, provided he or she has rational reason to believe those other estimates are no worse than his or her own. Brandt (1944) examines the implications of disagreement for "ethical rationalism" and heavy reliance on intuitions, but does not consider the Bayesian formulation. Coady (1992), drawing upon earlier work by Thomas Reid (1997 [1764]), argues for the importance of evidence based on the testimony of others, but does not consider whether opinions should then converge on agreement. The earliest philosophic precursor arguably is in Sextus Empiricus (2000, first edition predates 235 A.D.). Sextus Empiricus argued that when people disagree, we face an "equipollence" of reasons, and cannot adjudicate between competing claims, or judge our own perspective to be superior to that of others.

priors. We then present some stylized facts about disagreement, and consider how these facts can be reconciled with the basic theory of disagreement. We conclude by asking how one should rationally proceed in disagreements, and by asking whether we, the authors, agree about what the paper means.

II. The Basic Theory of Agreeing to Disagree

Most analysis of agreeing to disagree, like most analysis of inference and decision-making, has used Bayesian decision theory. Bayesian theory may not be fully satisfactory or fully general (see Nozick 1993), but the core results in agreeing to disagree have been generalized beyond Bayesian agents to a surprising extent. Such generalizations have been possible because these core results rest mainly on a few simple intuitions about rationality. To see the intuitive appeal of the basic argument, consider the following simple parable.

Imagine that John hears a noise, looks out his window and sees a car speeding away. Mary also hears the same noise, looks out a nearby window, and sees the same car. If there was a shooting, or a hit-and-run accident, it might be important to identify the car as accurately as possible.

John and Mary's immediate impressions about the car will differ, due both to differences in what they saw and how they interpreted their sense impressions. John's first impression was that the car was an old tan Ford, and he tells Mary this. Mary's first impression was that the car was a newer brown Chevy, but she updates her beliefs upon hearing from John. Upon hearing Mary's opinion, John also updates his beliefs. They then continue back and forth, trading their opinions about the likelihood of various possible car features. (Note that they may also, but need not, trade evidence in support of those opinions.)

If Mary sees John as an honest truth-seeker who would believe the same things as Mary given the same information (below we consider this "common prior" assumption in detail), then Mary should treat John's differing opinion as indicating things that he knows but she does not. Mary should realize that they are both capable of mistaken first impressions. If her goal is to predict the truth, she has no rational reason to give her own observation greater weight, simply because it was hers.

Of course, if Mary has 20/20 eyesight, while John is nearsighted, then Mary might reasonably give more weight to her own observation. But then John should give her observation greater weight as well. If they can agree on the relative weight to give their two observations, they can agree on their estimates regarding the car. Of course John and Mary might be unsure who has the better eyesight. But this is just another topic where they should want to combine their information, such as knowing who wears glasses, to form a common judgment.

If John and Mary repeatedly exchange their opinions with each other, their opinions should eventually stop changing, at which point they will be mutually aware (i.e., have "common knowledge") of their opinions (Geanakoplos and Polemarchakis 1982).[3] They will each know their opinions, know that they know those opinions, and so on.

We can now see how agreeing to disagree is problematic, given such mutual awareness. Consider the "common" set of all possible states of the world where John and Mary are mutually aware that John estimates the car age to be (i.e., has an "expected value" of) X, while Mary estimates it to be Y. John and Mary will typically each know many things, and so will know much more than just the fact that the real world is somewhere in this common set. But they do each know this fact, and so they can each consider, counterfactually, what their estimate would be if their information were reduced to just knowing this one fact. (Given the usual conception of information as a set of possible worlds, they would then each know only that they were somewhere in this common set of states.)

[3] For a basic exposition of Aumann-like results, see Geankoplos (1994).

Among the various possible states contained within the common set, the actual John may have very different reasons for his estimate of X. In some states he may believe that he had an especially clear view, while in others he may be especially confident in his knowledge of cars. But whatever the reason, everywhere in the common set John's estimate has the same value X. Thus if a counterfactual John knew only that he was somewhere in this common set, this John would know that he has some good reason to estimate X, even if he does not know exactly what that reason is. Thus this John's estimate should be X.

Similarly, if a counterfactual Mary knew only that she was somewhere in the common set, her estimate should be Y. But if counterfactual John and Mary each knew only that the real world is somewhere in this common set of possible worlds, they would each have exactly the same information, and thus should each have the same estimate of the age of the car. If John estimates the car to be five years old, then so should Mary. The same logic applies to any estimate of fact, such as the probability that the car is a Ford. This is Aumann's (1976) original result, that mutual awareness of opinions requires identical opinions.[4]

A more detailed Bayesian analysis says not only that people must ultimately agree, but also that the discussion path of their alternating expressed opinions must follow a random walk. If John and Mary are Bayesians, then Mary should not be able to tell John how John's next opinion will differ from what Mary just said. Mary's best public estimate of John's next estimate must instead equal Mary's current best estimate (Hanson 2003).

[4] The irrationality of agreeing to disagree can be seen as implicit in the classic "Dutch book" arguments for Bayesian rationality. These arguments showed that if an agent is willing to take bets on either side of any proposition, then to avoid combinations of bets that guarantee losses, that agent's betting odds must satisfy the standard probability axioms. Furthermore, any predictable rule for changing betting odds in the light of new information must make the new betting odds equal to the old odds conditional on that new information. An analogous argument applies to a group of agents. If the group is to avoid combinations of bets that guarantee losses for the group as a whole, then each

Yet in ordinary practice, as well as in controlled experiments (Hanson and Nelson 2002), we know that disagreement is persistent, i.e., that people can consistently and publicly predict the direction of other people's opinion relative to their own opinion. For instance, if John first says the car is six years old, and Mary then says the car is three years old, a real Mary can usually accurately predict that John's next estimate will probably be more than three years. If Mary is rational, this implies that John is not efficiently using the information contained in Mary's forecast.

Generalizations of the Basic Theory

While Aumann's results depended on many strong assumptions, similar results obtain when these assumptions are considerably relaxed. For example, rather than knowing the exact values of each other's estimates, John and Mary need only be mutually aware of the fact that John thinks the car is at least as old as Mary thinks it is. (That is, a mutual awareness of the fact that X >= Y also implies that X=Y.) Larger groups of people need only identify the "extremist" among them, the person who has highest estimate (Hanson 1998). It is also enough for people to be mutually aware of a single number that increases whenever any person's estimate increases (McKelvey and Page 1986).

We also can relax the requirement that John and Mary be absolutely sure of the things they are mutually aware of, i.e., that they have "common knowledge." We need instead assume only "common belief." That is, we need only assume that there is some common set of possible states of the world where 1) some condition like X>=Y holds, and 2) both John and Mary believe that they are in this common set. John and Mary can sometimes be mistaken in this belief, but the higher their confidence, the smaller can be the difference between their estimates X and Y (Monderer and Samet 1989).

Thus John and Mary need not be absolutely sure that they are both honest, that they heard each other correctly, or that they interpret language the same way. Furthermore, if John

group member must offer the same betting odds on every proposition. For the related literature on "no-trade" theorems, see Milgrom and Stokey (1982).

and Mary each assign only a small chance to such confounding factors being present, then their difference of opinion must also be proportionately small. This is because while payoff asymmetries can induce non-linearities in actions, the linearity of probability ensures linearity in beliefs. A rational Mary's estimate of the car's age must be a linear weighted average of her estimate conditional on confounding factors being present, and her estimate conditional on the absence of such factors.

These results are also robust to John and Mary having many internal biases and irrationalities, as long as they also have a "rational core." Consider a corporation with many irrational employees, but with a truth-seeking Bayesian CEO in charge of its official statements. This CEO should treat inputs from subordinates as mere data, and try to correct for their biases. While such corrections would often be in error, this company's official statements would be rational, and hence would not agree to disagree with statements by other companies with similar CEOs. Similarly, if John and Mary were mutually aware of having "clear head" rational cores capable of suspecting bias in inputs from the rest of their minds, they should not disagree about the car.

We also need not assume that John and Mary know all logical truths. Through the use of "impossible possible states," Bayesians do not need to be logical omniscient (Hintikka 1975; Garber 1983). John and Mary (or their rational cores) do not even need to be perfect Bayesians, as similar results have been proven for various less-than-Bayesian agents (Rubinstein and Wolinsky 1990, Samet 1990, Geanakoplos 1994). For example, agents whose beliefs are represented by sets of probability distributions can be said to agree to disagree when they are mutually aware that their sets do not overlap (Levi 1974).

Real people's beliefs usually depend not only on their information about the problem at hand, but also on their mental context, such as their style of analysis, chosen assumptions, and recent thoughts. This does not justify disagreement, however. If there can only be one best estimate, a truth-seeker should prefer to average over the estimates produced in many different mental contexts, instead of relying on just one random context.[5] So John should pay attention to Mary's opinion not only because it may embody information that John does not have, but also because it is the product of a different mental context, and John should want to average over as many mental contexts as he can.

This intuition can be formalized. When Mary has limited computational powers, we can say that she is a "Bayesian wannabe" if she can imagine counterfactually being a Bayesian, even if she is not actually one. It turns out that Bayesian wannabes who make a few simple calculations, and who would not agree to disagree in a certain strong way about a state-independent random variable, for which private information is irrelevant, cannot agree to disagree about any matter of fact (Hanson 1997).

III. Can rationally differing priors save the notion of honest disagreement?

While Aumann's result is robust to generalizing many of his assumptions, two of his assumptions are harder to generalize: truth-seeking and common priors. There seems little reason to expect agreement from people who usually lie, or who prefer to believe, for example, whatever seems the most self-flattering. And since information and priors jointly determine Bayesian beliefs, there is little reason to expect agreement from Bayesians with arbitrarily differing priors.

Might rationally differing priors therefore explain most disagreement? It all depends on how different rationally differing priors can be, compared with the typical variation in human opinions that disagree. We will now argue that, for truth-seekers, rational priors should be largely common, and that any differences in priors that might be rational are insufficient to explain most human disagreement. Thus to the extent that real human disagreement can be described in terms of differing priors, this seems to mostly be just another way to model non-truth-seeking, dishonest, or irrational behavior.

[5] John may have information suggesting that his mental context is better than random, but Mary may also have information on this topic. Persistent disagreement on this topic should be no less problematic.

Let us first review the nature of a prior. In general, agents can have beliefs not only about the world, but also about the beliefs of other agents, and so on. When modeling agents for some particular purpose, the usual practice is to collect a "universe" of all possible states of the world that any agent in the model considers, or suspects that another agent may consider, and so on. It turns out that one can then translate such agent beliefs into a "prior" probability distribution for each agent (Aumann 1998, Gul 1998). An agent's prior describes the probability she would assign to each state if her information were reduced to knowing only that she was somewhere in that model's universe of possible states. Many dynamic models contain an early point in time before the agents acquire their differing private information. In such models, the prior is intended to describe each agent's actual beliefs at this earlier time. In models without such an early time, however, the prior is interpreted counterfactually, as describing what agents would believe if sufficiently ignorant.

Note that even when priors are interpreted counterfactually, they have as much standing to be considered "real" as any other construct used to explain or justify spoken human opinions, such as information sets, or epistemic principles. All such constructs are intrinsically counterfactual. So it seems hard to reject the irrationality of uncommon priors on the grounds that priors are counterfactual, or on the grounds that priors are a "mere methodological construct," without also rejecting most normative epistemic analysis.

The most common argument given for the rationality of common priors is that differences in beliefs should depend only on differences in information. If John and Mary were witnesses to a crime, or jurors deciding guilt or innocence, it would be disturbing if their honest rational beliefs -- the best we might hope to obtain from them -- were influenced by personal characteristics unrelated to their information about the crime. And as truth-seekers, John and Mary should usually have no good reason to believe that the non-informational inputs into their beliefs have superior predictive value over the non-informational inputs into the beliefs of others. If so, differing Bayesian priors would describe a situation where at least one of the individuals has an innate propensity to rely on non-evidentiary considerations.[6]

A belief in the irrationality of uncommon priors is also implicit in the claim that rational Bayesians should change their beliefs by conditioning when they learn (or forget) information. That is, consider an earlier "self" who is the immediate causal ancestor of a later "self" who has learned a new fact about the world. These different selves are logically two different agents who can in principle have different preferences and beliefs. But it is usually said that the beliefs of the later self should be equal to the beliefs of the earlier self, conditional on that new fact. This is equivalent to saying that the earlier and later selves should base their beliefs on the same prior.[7]

If these two selves should have the same prior, then why should not all agents have the same prior? One might say that these two agents are special in having a relation of immediate causal ancestry. But if immediate causal relatives should have the same prior, then by transitivity so should all causal ancestors and descendants. Arguably the process of conception, connecting parents and children, is just as relevant a causal relation as that

[6] Some critics of the common prior assumption (e.g. Morris 1995, Gul 1998) claim correctly that common priors cannot be derived rigorously from an underlying maximization problem, or demonstrated to be "rational" in this general sense. We argue only that truth-seeking individuals cannot justify disagreement on the basis of differing priors. We do not argue that common priors maximize utility or provide the optimal solution to some game. Other authors (Bernheim 1986, Morris 1995) argue that multiple equilibria provide a rationale for differing priors. Since in game-theoretic terms each different equilibrium describes a different set of priors, if there can rationally be two different equilibria, there must be two distinct rational priors. In our view, this argument mistakenly elevates the notion of multiple equilibria -- a useful theoretical construct -- to a final description of the real world. Short of the quantum level, science finds no indeterminacy in the world. Only one equilibrium is possible and that is determined along with the correct set of accompanying priors. In any case we would still expect truth-seekers to agree on priors within whichever equilibrium comes to pass.

[7] Van Fraassen (1984) poses a version of the “agreeing to disagree” problem for a single individual over time. He considers the question of whether this logic can justify "self-fulfilling" beliefs, namely an individual deciding rationally to believe something, on the grounds solely that he expects to believe that same thing in the future. See also

between the selves of a person on different dates. In that case, the fact that all humans share a common evolutionary ancestor makes all humans relevant relatives. John and Mary should then share a common prior. Yes, different people acquire different genetic imprints, but we can think of these genetic imprints as nature giving us each different information, via a "different set of eyeglasses," leaving the common prior intact.

More general considerations of the causal origins of priors also seem to argue for the typical irrationality of non-common priors. If John and Mary have different priors, then they should realize that some physical process produced that difference. And if that difference was produced randomly or arbitrarily, it is not clear that John and Mary should retain it. After all, if John realized that some sort of memory error had suddenly changed a belief he had held for years, he would probably want to fix that error (Talbott 1990). So why should he be any more accepting of random processes that produced his earliest beliefs?

These intuitions can be formalized. A rational Mary should be able to form coherent, albeit counterfactual, beliefs about the chance that nature would have assigned her a prior different from the one she was actually given. If so, it seems that Mary's actual prior should be consistent with her extended "pre-prior" in the sense that the first prior should be obtained from the second by updating on the fact that nature assigned her a particular prior. Even if John and Mary have different pre-priors, one can show that if Mary thinks that it was just as likely that nature would have switched the assignment of priors, so that John got Mary's prior and vice versa, then John and Mary's priors must be the same. The priors about some event, like the car being a certain age, must also be the same if John and Mary believe that this event was irrelevant to estimating the chances that nature would have assigned them particular priors (Hanson 2001).

Typical beliefs about causal origins (i.e., actual pre-priors) seem to preclude rationally differing priors about most events, even if such differing priors might otherwise be

Christensen (2000) on this idea. In the modern literature, Hurley (1989) stresses the analogies between rationality in the interpersonal and intrapersonal cases.

rational. For example, standard models of genetic inheritance predict that siblings (and parents and children) have almost the same ex ante chance of getting any particular DNA, and that these chances are correlated with little else. Thus if priors are encoded in DNA, and if our pre-priors accept standard inheritance models, then it seems that we must have virtually the same priors as our siblings (and parents and children), and so cannot rationally agree to disagree with them. Thus if John and Mary are at all related, they must have nearly common priors.

There is another reason to question whether most human disagreements can be attributed to rationally differing priors, even if in principle some disagreements can be so attributed. Psychologists observe that human disagreement typically depends heavily on each person believing that he or she is better than others at overcoming undesirable influences on their beliefs (e.g., innuendo), even though people in fact tend to be more influenced than they realize (Wilson, Gilbert, & Wheatley 1998). More generally, though there is a genetic component to some general attitudes (Olson, Vernon, Harris, & Jang 2001), people are not endowed at birth with detailed context-specific prior opinions on most topics. Instead, cognitive psychologists find that much of the variation in human belief is due to various random features of how exactly each person is exposed to a topic, such as how the subject was first framed and what other beliefs were easily accessible then. These random beliefs then become the basis of disagreements. Each person holds the faith that he or she can reason better than other people, or that his or her basic emotional temperament is somehow more conducive to finding the truth.

Even if it were rational to be endowed at birth with differing priors specifying opinions on the vast number of topics that one might one day express opinions on, it seems much harder to justify having a differing prior that simply declares one's innate mental superiority. This seems prima facie self-serving. Thus we again see that even if some differences in priors might be rational for truth-seekers, such differences seem insufficient to justify most actual human disagreement.

IV. The Phenomena of Disagreement

Having reviewed a theory of disagreement, let us now review some stylized facts about human disagreement, so that we can compare theory to the phenomena.

Virtually any two people capable of communication can quickly find a topic on which they substantially disagree. In such disagreements, both sides typically believe themselves to be truth-seekers, who honestly say what they believe and try to believe what is true. Both sides are typically well aware of their disagreement, and can reliably predict the direction of the other's next statement of opinion, relative to their own last statement.

Many people dismiss the arguments of others, often on the grounds that those others are less smart, knowledgeable, or otherwise less able. At the same time, however, such people do not typically accede to the opinions of those who are demonstrably equally or more able, be the relevant dimension IQ, life experience, or whatever. People are commonly more eager to speak than they are to listen, the opposite of what a simple information-transmission model of discussion would predict (Miller 2000).

Disagreements, and dismissals of others, do not typically embarrass us. People are often embarrassed to discover that they have visibly violated a canon of rationality like logical consistency. Upon this discovery, they often (though not always) change their views to eliminate such violations. And in many cases (though again not always) fewer such violations tend to be discovered for individuals with higher IQ or superior training. Disagreements, however, happen even though people are usually well aware of them. Not only are disagreements not embarrassing, but more social shame seems to fall on those who agree too easily, and so lack "the courage of their convictions."

Real world disagreements seem especially frequent about relative abilities, such as who is smarter than whom, and about subjects, like politics, morality, and religion, where most people have strong emotional reactions. Discourse seems least likely to resolve disagreements of these kinds, and in fact people often move further away from each other's views, following a sustained dialogue.[8]

The positions taken in many disagreements seem predictable, as in "where you stand depends on where you sit." In general, people seem inclined to believe what they "want" to believe, such as that they are especially able. For example, most people, especially men, estimate themselves to be more able than others and more able than they really are (Waldman 1998). Gilovich (1991, p.77) cites a survey of university professors, which found that 94% thought they were better at their jobs than their average colleagues. A survey of sociologists found that almost half said they expected to become among the top ten leaders in the field (Westie 1973).[9]

High-IQ individuals seem no less likely to disagree, nor do academics, such as the authors, who are aware of and accept the arguments described in this paper. In response to these arguments, numerous academics have told us that trying to disagree less would be dishonest, destroy their identity, make them inhuman, and risk paralyzing self-doubt.[10]

V. Explaining disagreement

How can we reconcile the phenomena of persistent disagreement with theory that says such disagreement is irrational for truth-seekers? While the theory is robust to relaxing many of the original assumptions, it is not robust to relaxing all of them. Our suspicion should thus fall on the remaining assumptions, to which we give the collective name *meta-rationality*. People are meta-rational when they are rational, truth-seeking, and aware of their own fallibility relative to others. That is, they choose and report their opinions as if they know the basic principles of rationality, including the basic theory of

[8] On the tendency for polarization, see Sunstein (1999).

[9] For a survey of the psychology literature on this point, see Paulhus (1986).

[10] The description of the Houyhnhnms in Jonathan Swift's (1962 [1726]) Gulliver's Travels can be considered such a critique of inhumanity due to excessive agreement. In contrast, the concluding dream in Fyodor Dostoevsky's (1994 [1866]) Crime and

disagreement. Given their information and their effort level, meta-rational people find and report the truth.

The ubiquity of disagreement suggests that people are not usually meta-rational. But which component of meta-rationality is most to blame? One possibility is honesty; do people usually lie and not honestly stating their opinions? Unfortunately for this hypothesis, people usually have the strong impression that they are not lying, and it hard to see how people could be so mistaken about this. While there is certainly some element of sport in debates, and some recognition that people often exaggerate their views for effect, most people feel that they believe most of what they say. People sometimes accuse their opponents of insincerity, but rarely accept this same label as a self-description. Even when they are conscious of steering a conversation away from contrary evidence, people typically perceive that they honestly believe the claims they make.

Another possibility is that most people simply do not understand that disagreement is irrational or a sign of non-truth-seeking behavior. The arguments summarized in this paper are complex in various ways, after all, and recently elaborated. If this is the problem, then just spreading the word about the irrationality of disagreement should eliminate most disagreement. One would then predict a radical change in the character of human discourse in the coming decades. The reactions so far of people who have learned about the nature of disagreement, however, do not lend much support to this scenario. Not only do such people continue to disagree frequently, it seems hard to find any pair of them who, if put in contact, could not frequently identify many persistent disagreements on matters of fact.

If we reject dishonesty and ignorance about the irrationality of disagreement as causing most non-meta-rationality, we must accept the only remaining option - a lack of truth-seeking when forming opinions. While this possibility conflicts with our self-image, we cannot conclude that we are usually truth-seekers simply because we feel like truth-

Punishment seems to describe disagreement as the original sin, from which arises all other sins.

seekers; we may be *self-deceived*. That is, even if we often attempt at some conscious levels to find truth, at other levels our mental programs may systematically bias our beliefs in the service of other goals. Our mental programs may under-emphasize evidence that goes against some favored ideas, and distract the critical mechanisms that make us so adept at finding at finding holes and biases in other people's arguments (Mele 2001).

Our hypothesis of self-deception and non-truth-seeking as explaining most disagreement is consistent with the behavioral evidence discussed above. This hypothesis fits with our tendency to believe what we "want" to believe, and what we want others to believe, such as that we are especially able and virtuous. The hypothesis also fits our discrimination against those who agree too easily, perhaps because they seem less informed and more submissive. This hypothesis is further supported by the continued disagreement among those who accept disagreement is irrational. Finally, evolutionary arguments suggest that we should have evolved to be biased and self-deceived to some extent.[11]

We therefore posit that the main cause of our lack of meta-rationality, and therefore of persistent disagreement, is our lack of truth-seeking, i.e., our tendency to unconsciously choose beliefs for reasons other than closeness to truth, and to be self-deceived about this propensity. We feel confident even when we ought to know better and hold a more reserved judgment. Either few of us have rational cores, or those rational cores are not usually in control.

[11] Many have considered the evolutionary origins of excess confidence one's own abilities and self-deception (Waldman 1994). For example, if truth-seekers could benefit by saying certain things they do not believe, but find it hard to lie, then maybe they should just believe those things. If by thinking highly of himself, John can induce Mary to think more highly of him, then Mary may be more willing to associate with John (Trivers 1985; Trivers 2000). Argument more generally may serve the evolutionary function of showing off one's mental abilities. On topics like politics or religion, which are widely discussed but which impose few direct penalties for mistaken beliefs, our distant ancestors may have mainly demonstrated their cleverness and knowledge by inventing original positions and defending them well (Miller 2000).

We are reluctant to admit this lack of truth-seeking, either publicly or to ourselves. Few people say "What I've been telling you is not my best estimate of the truth, given everything I know." But when forced to confront the issue, people consistently choose to continue to disagree, suggesting that they fundamentally accept not being a truth-seeker.[12]

VI. How few truth-seekers?

So how much non-truth-seeking behavior have we really shown? We have not shown that everyone is self-deceived. If there were one meta-rational person, for instance, who was justified in assuming his own meta-rationality, then he or she would be justified in disagreeing with all the other irrational and/or dishonest people. So if one truth-seeker is possible, how many more are possible?

If meta-rational people were common, and able to distinguish one another, then we should see many pairs of people who agree with each other on just about all matters of fact. In reality, however, it seems very hard to find any pair of people who, if put in contact, could not identify many persistent disagreements. This suggests that there are either extremely few meta-rational people, or that they have virtually no way to distinguish each other.

Yet it seems that meta-rational people should be discernable via a distinct conversation style. We know that the sequence of alternating opinions between a pair of people who are mutually aware of both being meta-rational must follow a random walk. And we know that the opinion sequence between typical non-meta-rational humans is nothing of the sort. If, when responding to the opinions of someone else of uncertain type, a meta-rational person acts differently from an ordinary non-meta-rational person, then two meta-rational people should be able to discern one another via a long enough

[12] It is perhaps unsurprising that most people are not truth-seekers, in the sense that they do not always spend the effort required to overcome known biases. What may be more

conversation. And once they discern one another, two meta-rational people should no longer have persistent disagreements.

Since most people have extensive conversations with hundreds of people in their lives, many of whom they come to know very well, the fraction of us who are meta-rational must be very small. For example, with *N* people, a fraction *f* of whom are meta-rational, if each person participates in *C* conversations with random others that last long enough for two meta-rational people to discern each other, there should be on average $f^2CN/2$ pairs who no longer disagree. If, across the world, *C* was one hundred, *N* was two billion, and *f* was one in ten thousand, we should see one thousand pairs of people who agree. If, within academia, *C* was one thousand, *N* was two million, and *f* was one in ten thousand, we should see ten agreeing pairs of academics. And if meta-rational people had any other clues to discern each another, and preferred to talk with one another, there should be far more such pairs. Yet we know of *no* such pairs, with the possible exception of some cult-like or fan-like relationships.

We therefore conclude that only a tiny non-descript percentage of the population, or of academics, can be meta-rational. Either few people have truth-seeking rational cores, and those that do cannot be readily distinguished, or most people have such cores but they are in control very infrequently and unpredictably. Worse, since it seems unlikely that the only signals of meta-rationality would be purely private signals, we should have little grounds for confidence in our own meta-rationality, however much we would like to believe otherwise.

VII. What to do now?

Let us assume that you, the reader, are trying to be one of those rare meta-rational souls in the world, if indeed there are any. How guilty should you feel when you disagree?

surprising is the ubiquity of this behavior, and that it explains most disagreement, even though a rather low effort (e.g. stop disagreeing) can greatly reduce this bias.

If you and the people you disagree with completely ignored each other's opinions, then you might tend to be right more if you had greater intelligence and information. And if you were sure that you were meta-rational, the fact that most people were not could embolden you to disagree with them. But for a truth-seeker, the key question must be how sure you can be that you, at the moment, are substantially more likely to have a truth-seeking, in-control, rational core than the people you now disagree with. This is because if either of you have some meta-rationality, then your relative intelligence and information are largely irrelevant except as they may indicate which of you is more likely to be self-deceived about being meta-rational.

One approach would be to try to never assume that you are more meta-rational than anyone else. But this cannot mean that you should agree with everyone, because you simply cannot do so when other people disagree among themselves. Alternatively, you could adopt a "middle" opinion. There are, however, many ways to define middle, and people can disagree about which middle is best (Barns 1998). Not only is there disagreement on many topics, but there is also disagreement on how to best correct for one's lack of meta-rationality.

Ideally we might want to construct a model of the process of individual self-deception, consistent with available data on behavior and opinion. We could then use that model to take the observed distribution of opinion, and infer where lies the weight of evidence, and hence the best estimate of the truth. Ideally this model would also satisfy a reflexivity constraint: when applied to disputes about self-deception it should select itself as the best model of self-deception.[13]

A more limited, but perhaps more tractable, approach to relative meta-rationality is to seek observable signs that indicate when people are self-deceived about their meta-rationality on a particular topic. You might then try to disagree only with those who display such signs more strongly than you do.

[13] If most people reject the claim that most people are self-deceived about their meta-rationality, this approach will be more difficult, though perhaps not impossible.

For example, psychologists have found numerous facts about self-deception. Self-deception is harder regarding one's overt behaviors, there is less self-deception in a galvanic skin response (as used in lie detector tests) than in speech, the right brain hemisphere tends to be more honest, evaluations of actions are less honest after those actions are chosen than before (Trivers 2000), self-deceivers have more self-esteem and less psychopathology, especially less depression (Paulhus 1986), and older children are better than younger ones at hiding their self-deception from others (Feldman & Custrini 1988). Each fact implies a corresponding signal of self-deception. Other possible signs of self-deception include idiocy, self-interest, emotional arousal, informality of analysis, an inability to articulate supporting arguments, an unwillingness to consider contrary arguments, and ignorance of standard mental biases. Each of these results suggests potential clues for identifying other people as self-deceivers.

Of course, this is easier said than done. For example, it is easy to see how self-deceiving people, seeking to justify their disagreements, might try to favor themselves over their opponents by emphasizing different signs of self-deception in different situations. So looking for signs of self-deception need not be an easier problem than trying to overcome disagreement in the first place.

We therefore end this section on a cautionary note. We do not yet have a general recipe for how to respond to widespread disagreement. But the difficulty of this problem should at least make us wary of our own judgments when we disagree.

VIII. Conclusion

A literature spanning several decades has explored the finding that, on matters of fact, “honest disagreement" is problematic. Both Bayesian and other approaches point towards this same conclusion, which seems robust to many permutations. Academics, however, have not considered seriously enough what this implies about their own rationality. While academics like to think of themselves as rational and seeking and telling the truth, the ubiquity of disagreement strongly suggests that this is simply not so, and that academics are self-deceived about this fact and about their own meta-rationality.

What Tyler Thought Before Talking to Robin About This Problem

Tyler thought that widespread and persistent disagreement was not especially problematic.

What Tyler Thinks Now

Tyler now sees disagreement as highly problematic. He believes that non-truth-telling behavior, combined with self-deception, drives most observed disagreement. Most individuals think, wrongfully, that their opinions should count for more than the opinions of their peers. Most people (though not all) are sincere in their intentions to discover truth, yet they use algorithms that make non-true arguments look better than they deserve. Tyler is persuaded by the arguments of Freud, Lacan, Nietzsche, and the seventeenth century French moralists that individuals are not generally transparent to themselves. People are not Bayesian "wannabes," nor would a simple investment of effort make them as such, given the complexity of the underlying psychological distortions.

There is no simple answer as to when, and with whom, Tyler is willing to disagree with others. He acts as if he has multiple selves. Tyler reports that one of his parts is somewhat dogmatic, and thinks he is smarter than most other people, and better able to know truth, while ignoring issues of meta-rationality in the typically self-deceived fashion. Tyler reports another of his parts to be deeply agnostic and skeptical of his own abilities to best others in discovering truth. Those who listen to Tyler (or read his work) usually receive some combination of the two parts. Tyler claims he doesn't know any way that the latter part in him could drive out the former. Tyler often thinks he simply has a good idea what the right answer is. Tyler feels that, in many cases, to hide behind a cloak of (supposed) meta-rationality that he does not fully believe would be an even greater form of self-deception than what already goes on. Tyler would like to try harder to be more truth-seeking, but he is not sure what this would consist of or how it might succeed.

What Robin Thought Before Talking to Tyler

Robin thought that disagreement was irrational, but had not thought much about self-deception, and had suspiciously avoided drawing any conclusions regarding his own rationality and self-deception.

What Robin Thinks Now

Robin now agrees with Tyler that non-truth-seeking behavior and self-deception seem to best explain most observed disagreement, and Robin accepts this description as usually applying to himself. People greatly overestimate their ability to consciously control their minds, much like someone riding a boat on the ocean, who claims to cause the boat to rise and fall as the ocean swells underneath. And yet Robin still believes that with perhaps great effort he can choose which direction he rows his boat.

Robin is also very concerned that since we do not consciously decide the topics on which we allow ourselves to be self-deceived, this decision probably relies heavily on ancient genetic and cultural habits. As the environment of modern society diverges from our ancestors' environments, this reliance seems increasingly dangerous.

Robin thus insists on trying hard to seek truth, starting with the known observable signs of self-deception. Robin tries to avoid disagreements, especially with smarter people like Tyler. Robin is most willing to disagree with those who find his opinion silly or crazy, but who will neither articulate critical arguments nor consider his supporting arguments. This scenario is often played out regarding Robin's opinions on the potential of envisionable future technologies.

What Robin Thinks About the Point of Disagreement With Tyler

Meta-rationality seems hard, but hardly impossible. We have hardly begun to seriously study the phenomena of self-deception. Yet Tyler seems to have given up, apparently saying "If God meant man to fly, he would have given him wings." Our civilization is the accumulation of ways to let primates do what once seemed impossible. Our core nature may not change, but it appears to as we embed ourselves in new physical and institutional contexts. In the next century we may well construct computer-based "rational cores" outside ourselves which can correct for our biases, and we may also rely heavily on betting markets to provide a visible Bayesian consensus “middle” to which all can contribute and defer (Hanson 1995). We have only begun to fight irrationality, and we need not change our nature to do so.

What Tyler Thinks About the Point of Disagreement With Robin

I haven't given up, Robin and I simply do not agree on the right way to proceed. Robin says he thinks Tyler is smarter, but he still disagrees with Tyler on what this paper means. So is Robin really trying, or is he just investing in a yet deeper form of self-deception? Psychologists also note that we overestimate our capacity to improve ourselves, so has writing this paper really made Robin less self-deceived? Human nature is hard to change.

How Robin Rationalizes the Disagreement

Most people have mental blocks to thinking seriously about the future. Tyler too.

How Tyler Rationalizes the Disagreement

Most people agree with Tyler that human nature is hard to change. Robin is deceiving himself.

References

Arnauld, Antoine and Nicole, Pierre. *Logic of the Art of Thinking*. Cambridge: Cambridge University Press, 1996 [1683].

Aumann, Robert J. "Agreeing to Disagree." *The Annals of Statistics*, 1976, 4, 6, 1236-1239.

Aumann, Robert J. "Common Priors: A Reply to Gul." *Econometrica*, July 1998, 66, 4, 929-938.

Barns, EC, "Probabilities and epistemic pluralism." *The British Journal for the Philosophy of Science*, March 1998. 49,1, 31-47.

Bernheim, R. Douglas. "Axiomatic Characterizations of Rational Choice in Strategic Environments." *Scandinavian Journal of Economics*, 1986, 88, 3, 473-488.

Bikchandani, Sushil, Hirshleifer, David, and Welch, Ivo. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, October 1992, 992-1026.

Bonnanno, Giacomo and Nehring, Klaus. "How to Make Sense of the Common Prior Assumption Under Incomplete Information." *International Journal of Game Theory*, 1999, 28, 409-434.

Brandt, Richard B. "The Significance of Differences of Ethical Opinion for Ethical Rationalism." *Philosophy and Phenomenological Research*, June 1944, 4, 4, 469-495.

Caplan, Bryan. "The Logic of Collective Belief." Forthcoming in *Rationality and Society* , 2003.

Christensen, David. "Diachronic Coherence versus Epistemic Impartiality." *The Philosophical Review*, July 2000, 109, 3, 349-371.

Coady, C.A.J. *Testimony: A Philosophical Study*. Oxford: Clarendon Press, 1992.

Dostoevsky, Fyodor. *Crime and Punishment*. Barnes and Noble Books, New York, 1994 [1866].

Everett, Theodore, "The Rationality of Science and the Rationality of Faith" *Journal of Philosophy*, 2001, 19-42.

Feinberg, Yossi. "Characterizing Common Priors in the Form of Posteriors." *Journal of Economic Theory*, 2000, 91, 127-179,

Feldman, Robert, and Custrini, Robert, "Learning to Lie and Self-Deceive" in *Self-Deception: An Adaptive Mechanism?,* edited by Joan Lockard & Delroy Paulhaus, Prentice Hall, 1988.

Garber, Daniel, "Old Evidence and Logical Omniscience in Bayesian Decision Theory", in *Testing Scientific Theories*, edited by John Earman, University of Minnesota Press, 99-131.

Geanakoplos, John. "Common Knowledge." *Handbook of Game Theory*, volume 2, edited by R.J. Aumann and S. Hart. Elsevier Science, 1994, 1437-1496.

Geanakoplos, John D. and Polemarchakis, Heraklis M. "We Can't Disagree Forever." *Journal of Economic Theory*, 1982, 28, 192-200.

Geanakoplos, John D. (1994). Common Knowledge.In Aumann,R.,&Hart,S.(Eds.), *Handbook of Game Theory*, Volume 2 pp.1438-1496. Elsevier Science.

Gilovich, Thomas. *How We Know What Isn't So*. New York: Macmillan, 1991.

Gul, Faruk. “A Comment on Aumann’s Bayesian View.” *Econometrica*, July 1998, 66, 4, 923-927.

Goldstein, Michael. “Temporal Coherence” *Bayesian Statistics* 2, edited by Jose Bernardo, Morris DeGroot, Dennis Lindley, and Adrian Smith, 1985.

Goodin, Robert. “The Paradox of Persisting Opposition” *Politics, Philosophy, Economics*, 1, 2002.

Hacking, Ian. “Slightly more realistic personal probability”, panel discussion, *Philosophy of Science*, 1967.

Hanson, Robin. “Could Gambling Save Science? Encouraging an Honest Consensus” *Social Epistemology* 9(1):3-33, 1995.

Hanson, Robin. “Consensus By Identifying Extremists.” *Theory and Decision* 44(3):293-301, 1998.

Hanson, Robin. “For Savvy Bayesian Wannabes, Are Disagreements Not About Information?” Chapter in Ph.D. thesis, California Institute of Technology, October 1997.

Hanson, Robin. “Uncommon Priors Require Origin Disputes.” Unpublished manuscript, George Mason University, 2001.

Hanson, Robin. “Disagreement is Unpredictable.” *Economics Letters*, to appear, 2003.

Hanson, Robin and Nelson, William. "An Experimental Test of Agreeing to Disagree." Unpublished manuscript, George Mason University, 2002.

Harsanyi, John. “Bayesian Decision Theory, Subjective and Objective Probabilities, and Acceptance of Empirical Hypotheses” *Synthese* 57, 341-365, 1983.

Hintikka, J. “Impossible Possible Worlds Vindicated” *Journal of Philosophical Logic* 4, 475-484, 1975.

Hurley, Susan. *Natural Reasons: Personality and Polity*. New York: Oxford University Press, 1989.

Kitcher, Philip. “The Division of Cognitive Labor” *The Journal of Philosophy*, 87, 1, 5-22, 1990.

Levi, Isaac. “On Indeterminate Probabilities” *Journal of Philosophy* 71, 391-418, 1974.

McKelvey, Richard D. and Page, Talbot. "Common Knowledge, Consensus, and Aggregate Information." *Econometrica*, January 1986, 54, 1, 109-127.

Mele, Alfred R. *Self-Deception Unmasked*. Princeton: Princeton University Press, 2001.

Milgrom, Paul and Stokey, Nancy. "Information, Trade, and Common Knowledge." *Journal of Economic Theory*, 1982, 26, 17-27.

Miller, Geoffrey. *The Mating Mind, How Sexual Choice Shaped the Evolution of Human Nature*, Random House, New York, 2000.

Monderer, Dov and Samet, Dov. "Approximating Common Knowledge with Common Beliefs." *Games and Economic Behavior*, 1989, 1, 170-90.

Morris, Stephen. “The Common Prior Assumption in Economic Theory.” *Economics and Philosophy*, 1995, 11, 227-253.

Nau, Robert F. “Coherent Decision Analysis with Inseparable Probabilities and Utilities.” *Journal of Risk and Uncertainty*, January 1995, 10(1), 71-91.

Olson, James, Vernon, Philip, Harris, Julie, and Jang, Kerry. “The Heritability of Attitudes: A Study of Twins” *Journal of Personality and Social Psychology*, June 2001, 80(6) 845-860.

Nozick, Robert. *The Nature of Rationality*. Princeton: Princeton University Press, 1993.

Paulhus, Delroy L. "Self-deception and Impression Management in Test Responses." In Angleitner, A. & Wiggins, J. S., *Personality assessment Via Questionnaires*. New York, NY: Springer, 1986, 143-165.

Reid, Thomas. *An Inquiry into the Human Mind*. University Park, Pennsylvania: Pennsylvania State University Press, 1993 [1764].

Rescher, Nicholas. *Pluralism: Against the Demand for Consensus*. Clarendon Press, Oxford, 1993.

Rubinstein, A., 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.

Scharfstein, David S. and Stein, Jeremy C. "Herd Behavior and Investment." *American Economic Review*, June 1990, 80, 465-479.

Schiller, F.C.S. *Must Philosophers Disagree?* London: Macmillan and Co., Limited, 1934.

Sextus Empiricus. *Outlines of Scepticism*. Cambridge: Cambridge University Press, 2000, first edition predates 235 A.D.

Sunstein, Cass. "The Law of Group Polarization." University of Chicago Law School, John M. Olin Law & Economics Working Paper, no.91, December 1999.

Swift, Jonathan. *Gulliver's Travels and Other Writings*, New York: Bantam Books, 1962 [1726].

Talbott, William J. *The Reliability of the Cognitive Mechanism, A Mechanistic Account of Empirical Justification*, 1990, Garland Publishing, New York.

Taylor, S.E. *Positive Illusions: Creative Self-Deception and the Healthy Mind.* New York: Basic Books, 1989.

Trivers, Robert, *Social Evolution*, Benjamin/Cummings, Menlo Park, Ca., 1985.

Trivers, Robert, "The Elements of a Scientific Theory of Self-Deception," in *Evolutionary Perspectives on Human Reproductive Behavior*, Annals of the New York Academy of Sciences, edited by Dori LeCroy and Peter Moller, volume 907, April 2000.

Van Fraasen, C. "Belief and the Will." *Journal of Philosophy*, May 1984, 81, 5, 235-256.

Waldman, Michael. "Systematic Errors and the Theory of Natural Selection." *American Economic Review*, June 1994, 84, 3, 482-497.

Westie, Frank R. "Academic Expectations of Professional Immortality: A Study of Legitimation." *The American Sociologists*, February 1973, 8, 19-32.