The theory of games as a tool for the social epistemologist

Kevin J.S. Zollman*

July 8, 2019

Abstract

Traditionally, epistemologists have distinguished between epistemic and pragmatic goals. In so doing, they presume that much of game theory is irrelevant to epistemic enterprises. I will show that this is a mistake. Even if we restrict attention to purely epistemic motivations, members of epistemic groups will face a multitude of strategic choices. I illustrate several contexts where individuals who are concerned solely with the discovery of truth will nonetheless face difficult game theoretic problems. Examples of purely epistemic coordination problems and social dilemmas will be presented. These show that there is a far deeper connection between economics and epistemology than previous appreciated.

Many philosophers make a distinction between epistemic and pragmatic criteria for decision making (e.g. Joyce, 1998; Kelly, 2003; Oddie, 1997). Epistemic criteria are those that pertain principally or exclusively to attaining knowledge or forming reliable beliefs. Pragmatic criteria deal with considerations beyond that like the desire for happiness, the alleviation of suffering, and nice cars.

Mathematical tools originally developed for pragmatic decision making have been imported into epistemic contexts. First, the theory of probability – created to understand gambling – was adopted by epistemology. Bayes is famous for suggesting that probability theory can be marshaled to justify belief change in the face of Hume's skeptical doubts (see

*The title of this paper is an homage to R. B. Braithwaite's insightful book *The Theory of Games as a Tool for the Moral Philosopher* (1954). The author would like to thank Liam Kofi Bright, Remco Heesen, Gurpreet Rattan, Teddy Seidenfeld, Julia Staffel, Katie Steele, and several workshop audiences for their comments.

Diaconis and Skyrms, 2018, Chapter 6). Building on that early insight, Bayesian epistemology has grown into an enormous field where probabilistic reasoning is employed to answer epistemic questions (e.g. Bovens and Hartmann, 2003).

Modern utility theory was developed by von Neumann and Morgenstern (1953) to provide a foundation for individual decision making in economic contexts. Joyce (1998) adapted an argument originally developed by de Finetti (1975) to construct a "purely epistemic" utility function, thereby moving utility theory into the epistemic context. This program has been further developed in many directions (e.g. Pettigrew, 2016).

This paper continues this tradition by illustrating the importance of game theory in purely epistemic reasoning. I provide two examples where individuals, motivated by exclusively epistemic considerations, find themselves in strategically complicated situations. The central claim of the paper is that game theory is like probability and utility theory; it is a necessary tool for understanding certain types of purely epistemic problems.

In so doing, I also illustrate an important game theoretic advance. Normally, in game theory it is assumed that agents have (somewhat) divergent desires formalized with distinct utility functions. In this paper, I consider agents who are epistemically altruistic – all care only about attaining truth, not just for themselves but for everyone in their community. From a game theory perspective it is surprising that strategically complicated situations arise when individuals have identical preferences.[1]

In the first example (a case of disagreement), we find that individuals are facing the common – and difficult to handle – Prisoner's dilemma. Even if they can solve the problem of cooperation, a thorny bargaining problem remains. In the second example, we have a richer strategic setting. Two individuals must choose which of two hypotheses to investigate. By manipulating two parameters in their learning problem, we generate four of the canonical five types of two-player, two-strategy symmetric games. (We can construct Prisoner's dilemma, Prisoner's delight, the Stag hunt, and a Coordination game. Chicken or Hawk-Dove is the only absent game.)

While the individual cases are interesting, my broader desire is to show that the theory

[1]It is common in game theory to assume that both players share a common prior. I do not make this assumption here, which is what drives these examples.

of games is at home as much in epistemology as it is in pragmatic decision making. In so doing, I illustrate that the connection between epistemology and economics is far deeper than usually assumed.

I am not the first to draw a connection between game theory and epistemology. Following Dretske (1981), Skyrms (2010) connects signaling games to epistemic questions. Heesen and van der Kolk (2016) analyze the problem of disagreement by connecting it to reliability in a repeated game of compromise and stubbornness. List and Pettit (2004) appeal to the phenomena of information cascades as a potential example of an epistemic free-rider problem (see Banerjee, 1992, for early work on information cascades). While arguing against List and Pettit, Dunn (2018) finds a slightly different free-rider problem in group beliefs in deliberation. The debate between Dogramaci (2012) and Tebben and Waterman (2015) is one about the role that incentives play in social epistemology. In particular, Tebben and Waterman (2015) suggest that in the context of testimony, there may be a free-rider problem.[2] Similarly, Kummerfeld and Zollman (2016) and Zollman (2018) find examples of social dilemmas in models designed to understand features of scientific research.

These papers offer support for the position I'm advancing. I am presenting new examples, not because previous ones are flawed, but because those papers make assumptions which might be viewed as inappropriate in the "purely epistemic" context (depending on how that is interpreted). In some of these cases, a reader might classify the agents as considering pragmatic rather than epistemic criteria. In other cases, authors focus on the case of epistemic "selfishness" where individuals are aiming to maximize their private accuracy without consideration to how their behavior affects another. One might imagine that a "purely" epistemic agent would care about the beliefs of others as well as her own. The two examples that follow avoid these concerns: they both feature agents that are purely epistemic and care equally about the beliefs of others as they do their own. Even then, they find themselves facing complex game theoretic situations.

[2]Although they don't use the term, Tebben and Waterman (2015) identify what game theorists call a "second-order" free-rider problem: I can gain the benefits of cooperative agreements and shirk on the costs of maintaining the social norms via punishment.

1 Epistemic utility theory

One's beliefs ought to be accurate. Optimally, all of one's beliefs would be exactly correct, but this is a standard few of us can achieve. So instead, it is useful to adopt a gradated method of evaluating accuracy. In qualitative contexts, where I either believe something or its negation, accuracy is relatively easy to measure. If my belief is true, I'm accurate. Otherwise not. This is slightly more complex if agents can withhold belief, but even here it seems relatively simple.[3]

When I have probabilistic beliefs – like assigning a numeric probability to the proposition that my preferred candidate will win the next election – things are more difficult. We need a way to measure accuracy that makes fine distinctions between different degrees of belief. If my preferred candidate wins, I am most accurate if I assigned the proposition probability 1, less accurate if I assigned it 0.5 and even less so if I assigned it 0.1. Our measure of accuracy should represent that.

The desire to have such a measure has driven research into *scoring rules*, rules that enable us to grade quantitative beliefs. Scoring rules should assign a lower score to beliefs that are further away from the truth. As in the election example, if the proposition is true, a belief of 0.5 should be judged more accurate than a belief of 0.1. This is not the only relevant criteria, there are many more. We need not discuss them here, but detailed discussions can be found elsewhere (Pettigrew, 2016; Schervish, 1989; Schervish et al., 2009; Seidenfeld, 1985).

We can define such a rule this way. Suppose a single proposition p which an agent assigns a credence c. If p turns out to be true, our agent is assigned an accuracy score of $S(c, 1)$ and if it turns out to be false $S(c, 0)$. The most obvious scoring rule, advocated by Goldman (1999), is the simple difference between the truth and one's belief:

$$S(c, 0) = -c$$
$$S(c, 1) = -(1 - c)$$

If the proposition is true, then the "correct" probability is 1 and the distance is $1 - c$. Similarly, when the proposition is false.

[3]For a more complete discussion of this situation see (Easwaran, 2016).

Supposing that an agent has credence c, they might ask how accurate they would be if they changed their belief. An agent who has belief c and is evaluating the expected accuracy of alternative belief, c', expects c' would receive the score: $E(c', c) = cS(c', 1) + (1-c)S(c', 0)$ It seems natural that an agent's beliefs should be self-ratifying, that c maximizes one's expected score (Joyce, 2009). If a scoring rule S self-ratifies all (probabilistically coherent) beliefs in this way, the scoring rule is called a *proper scoring rule*.[4]

This is a kind of self-consistency that one expects with belief: one's belief should be one's best-guess about the world. In the context of scoring rules, this means that an individual might refrain from adopting a belief which does not regard *itself* as most accurate according to an agent's preferred scoring rule. It would be strange to say "I believe P and as a result, I regard my belief as inaccurate."

It turns out that the simple rule advocated by Goldman is not proper. This rule rewards extremity. Whatever my current belief, Goldman's rule would endorse adopting a credence of either 0 or 1 (whichever is more probable given my initial belief). An agent who believes that the flip of a coin has a 0.51 probability of coming up heads would endorse only the belief that assigned probability 1 to that proposition.[5] Thought of as a rule for scoring groups of inquires, this rule would reward communities made up of people with extreme credences over communities that are more reasonable.

Instead, many scholars advocate the Brier score (Brier, 1950), which is proper. The Brier score is given by:

$$S(c, 0) = -c^2$$

$$S(c, 1) = -(1-c)^2$$

[4]While I describe this as an agent who is considering a change in belief, in economics it is more common to talk about incentive compatibility. If I have belief c and I know I will be paid in proportion to scoring rule S, then we can ask: would I maximize my expected monetary return by honestly announcing my opinion c instead of an alternative opinion c'? The formal constraints are the same, but they have a different interpretation.

[5]Consider an agent who believes the probability of heads is 0.51. When they evaluate their own belief they assign it an expected accuracy of $E(0.51, 0.51) = 0.51(-0.49) + 0.49(-0.51) = -0.4998$. When that same agent considers the expected accuracy of adopting a different belief (or the expected accuracy of a different agent), say the belief of 1, the expected accuracy is $E(1, 0.51) = 0.51(0) + 0.49(-1) = -0.49$ As a result, the agent's belief in 0.51 is self-undermining: they regard themselves as less accurate than they would be with a different belief.

This is Goldman's rule, except the distance is squared. This results in larger punishment for larger deviations, which converts Goldman's improper rule into a proper one.[6]

At this point one might object to my description of an agent as "changing" her belief. This way of speaking may make belief look too much like action. One can take claims about changing belief as short hand for more complicated claims. So instead of saying that "our agent would prefer to change her belief," we might say "our agent regrets having the belief she does." Or alternatively, we might say "our agent endorses another agent as more rational than herself, not because the other knows more, but because our agent's own beliefs are self undermining."

Beginning with de Finetti (1975), philosophers have employed scoring rules to establish normative conclusions about probabilistic beliefs. Joyce (1998, 2009) argued that proper scoring rules present a purely epistemic utility function – an agent who cared only for accuracy would act so as to maximize her score on a proper scoring rule. In this respect, one can see a proper scoring rule as a formalization of the concept of a purely epistemic agent. An agent who cares only for epistemic considerations would form their beliefs as if they were maximizing a utility function that was a proper scoring rule.

This way of viewing scoring rules has led to a large literature analyzing different epistemic norms (e.g. Greaves and Wallace, 2006; Pettigrew, 2016). This philosophical innovation allowed for the incorporation of decision theory into epistemology. Of course, there are concerns about this program (Carr, 2017; Levinstein, 2012; Greaves, 2013). Without wading into this debate, I will presume that a proper scoring rule should count as a "purely epistemic" measure of an individual's utility. I believe that the examples that follow could be reconstructed from any plausible way of caching out a measure of epistemic goodness or badness. Do no take this assertion too seriously, however; it is no more than a hunch.

2 Disagreement

Suppose two friends, Ann and Bob, discover that they disagree about some proposition of interest: the probability that their favorite author will ever finish writing the book series they

[6]While Brier score is popular for its simplicity, it is not without critics (Fallis and Lewis, 2016).

both love. Ann believes this to be relatively improbable, perhaps she assigns it probability 0.1. Bob thinks it is likely to happen, his credence is 0.75. They debate long into the night, each presenting what they take to be evidence for their preferred probability judgment. But after all reasons have been exhausted, they find themselves no better off than when they started – they continue to disagree.[7] What shall Ann and Bob do in such a situation? Should they "agree to disagree" or should they find some compromise between their beliefs – taking each other as a reasonable epistemic agent and changing their belief in light of this?[8]

This question is closely related to a different question about how third parties should handle disagreement. Suppose Carole, someone who knows nothing about Ann and Bob's favorite author, comes to them and asks for their opinion. What should Carole do when experts disagree? Is she forced to choose one or the other to "believe" or should she construct some other compromise position that takes both experts' opinions into account?

Finally, there is a third question. Suppose Ann and Bob disagree, but must jointly make a decision. Perhaps they represent a publishing company that must decide whether to pay an advance to the author for his last book. They may continue to disagree privately, but they are obligated to take a joint action together. In some ways this is more like the second question than the first, they don't have to change their attitudes, but they must construct a new third attitude on which to base their actions.[9] (In this example, the cooperative Ann-and-Bob-together takes the place of Carole.)

Many of the discussions on disagreement have tried to answer several of these questions at once. Here, I will set aside the latter two issues. In this paper we only address the question of how Ann and Bob ought to handle their particular disagreement. What is said here will be relevant for the second and third questions as well, but I will leave exploring the

[7]The literature on disagreement often focuses on questions about *peer* disagreement, and much of the debate turns on how one analyzes the concept of a "peer." I do not wish to engage with this debate. If the reader would like to call these agents "peers," I have no objection. But if the reader would prefer to call this a case of non-peer disagreement, I will not argue. Whatever the preferred nomenclature, this case is worthy of discussion.

[8]Due to a well-known theorem from Aumann (1976), if Ann and Bob share a prior, are Bayesian rational, learn different evidence, but commonly know the structure of evidence, this cannot happen. I will assume that Ann and Bob are Bayesian rational but one of the other conditions fail. For example, they may not have a common prior.

[9]For reasons not critical to this paper, the cooperative action problem is very difficult (see Seidenfeld et al., 1989).

implications for another time.

Moss (2011) was the first to suggest employing epistemic utility theory to address this problem in disagreement. Moss focuses on my third question, by supposing that Ann and Bob must find a compromise, and then asks what compromise would be best for them. I will not argue with her conclusions, but what I hope to show here is that when we shift to look at the question of disagreement things become rather more complicated.

Let us formalize the situation this way: suppose a single proposition under discussion (although this could easily be expanded to multiple propositions). Ann assigns the proposition credence c_A and Bob assigns it c_B. Given that Ann has credence c_A, she evaluates the expected accuracy of a different belief, call it c', by calculating its expected accuracy: $E_A(c', c_A) = c_A S_A(c', 1) + (1 - c_A) S_A(c', 0)$. Ann determines what she expects the score of the alternative belief c' would be given her current belief c_A and her preferred scoring rule S_A. She thinks that the probability of the event occurring is c_A and in such a case the belief will be awarded score $S_A(c', 1)$. Similarly, for the event not occurring. We will assume all the same for Bob with the beliefs and scoring rules swapped out for those used by Bob.

As described in the previous section if Ann is utilizing a proper scoring rule, then Ann will always regard her own belief, c_A as maximizing her expected accuracy. This is the definition of propriety. And, this has a sense of consistency to it: I should regard my own belief as the most accurate, otherwise I should have a different belief.

In keeping with the general method of the epistemic utility program, we assume that Ann and Bob care for accuracy. If they cared only for their own accuracy, we would have nothing to discuss. Each would stay steadfast in their beliefs. Instead, let us assume that Ann and Bob are maximally altruistic: they care equally about their own accuracy as they do about the other. This complicates things. Ann desires to maximize their joint accuracy $JA_A(c_A, c_B) = E_A(c_A, c_A) + E_A(c_B, c_A)$ – the sum of her accuracy and what she expects Bob's accuracy to be. Similarly for Bob.

It is at this point that our discussion has diverged from Moss. Put in my notation, Moss asks both Ann and Bob to maximize this function: $E_A(c_A, c_A) + E_B(c_B, c_B)$. There are two odd assumptions built in here. First, Ann is using Bob's scoring rule S_B to evaluate Bob rather than using her own. While this assumption strikes me as strange, it doesn't

really matter in the discussion that follows. Instead, the troubling assumption is that Ann is evaluating Bob's credence relative to *Bob's* belief c_B. Why should Ann do this? Bob's credences are not Ann's. If Ann wants to determine what she thinks Bob's expected accuracy will be, Ann should use Ann's beliefs about the states of the world. We are asking Ann what she thinks of Bob's credences. Perhaps this is a reasonable assumption in the context of compromise (although I remain skeptical). It is certainly not a reasonable assumption in the context of responding to disagreement.

To put the issue more concretely: if you ask me what is my opinion of the beliefs of a politician about the prospect for economic growth, I do not report how I would feel if I were the politician. I would (and should!) say, "The politician believes that the prospects for growth are strong because she believes that manufacturing will continue to grow. But I don't think manufacturing will continue to grow, therefore I think the politician is overly optimistic about prospects for growth." Analogously for Ann, she should not say "Bob thinks that he is accurate, so I do too." Instead she should say, "Bob's belief is very far from the correct one."

Moss and I share the assumption that Ann and Bob should measure their "joint accuracy" by taking the sum of their individual accuracies. This is not the only way to combine their individual scores into a measure of joint accuracy. One might use the geometric average or some other way of combining the score. Or one might judge the group in some other way: according the most accurate or least accurate one of them.[10] I will stick with the simple case of arithmetic averaging, while noting that this is not an innocuous assumption.

Even using the arithmetic average, we see the situation is more complicated than Moss' discussion would suggest. Now Ann and Bob no longer agree about their expected accuracy. However, when Ann and Bob disagree, it remains the case that there are compromise positions that *both* regard as better than their current beliefs.

Consider the average of their beliefs $\bar{c} = \frac{1}{2}(c_A + c_B)$. So long as Ann and Bob each use

[10]Evaluating the accuracy of two (or more) agents is formally very similar to evaluating the accuracy of a single agent whose beliefs are represented by a set of probabilities rather than just one (this is the "imprecise probability" framework). There are thorny issues in evaluating the accuracy of imprecise credences (Seidenfeld et al., 2012). However, various types of averaging (like arithmetic and geometric) are more defensible in the context of groups than in the context of a single individual with imprecise credences. Shifting to other measures like the minimum or maximum score will likely result in undesirable results.

concave, strictly proper scoring rules S_A and S_B (they need not be the same rule), they both regard the compromise position $\bar{c}$ as superior to each staying the same.[11] A scoring rule is concave if for all c_1 and c_2 $S(\frac{1}{2}(c_1 + c_2)) \geq \frac{1}{2}(S(c_1) + S(c_2))$. Put it words, a scoring rule is concave if the score of the average of two beliefs is better than the average of the score of those same two beliefs.[12]

Proposition 1. *Suppose that S_A and S_B are concave, strictly proper scoring rules. $JA_A(\bar{c}, \bar{c}) \geq JA_A(c_A, c_B)$ and $JA_B(\bar{c}, \bar{c}) \geq JA_B(c_A, c_B)$. When $c_A \neq c_B$ and S_A and S_B are strictly concave, the inequalities are strict.*[13]

It is worth one short technical aside. The Brier score is strictly concave; so too are many other proper scoring rules. However, not all proper scoring rules are. The discontinuous scoring rule presented in (Schervish, 1989, 1863) is not concave and the theorem is no longer true for some pairs of beliefs (although it is true for others). Schervish's discontinuous scoring rule looks like this:

$$S(c, 0) = \begin{cases} -c^2 & x \leq \frac{1}{2} \\ -c^2 - 1 & x > \frac{1}{2} \end{cases}$$

$$S(c, 1) = \begin{cases} -(1-c)^2 & x > \frac{1}{2} \\ -(1-c)^2 - 1 & x \leq \frac{1}{2} \end{cases}$$

This is the Brier score with an added penalty for being on the "wrong" side of 1/2. An agent who assigns a true proposition probability 0.51 does substantially better than one who assigns it 0.49.

Proposition 1 is not true under Schervish's discontinuous rule. Suppose Ann has belief $c_A = 0.75$ and Bob has belief $c_B = 0.25$. Ann would regard the compromise belief 0.5 as

[11] While the arithmetic average $\bar{c}$ will be better than sticking to one's beliefs, depending on scoring rule, it may not be the optimal compromise. Moss provides several examples where different compromises are better.

[12] There is an unfortunate duality in how these rules are discussed. One can, as I have, talk about measures of accuracy, where higher scores are better. In that context, Brier score is concave. Alternatively, one can talk about measures of inaccuracy where lower scores are better. In that case, Brier score is convex. Joyce (1998, 2009) defends what he calls "convexity." This is a defense of what I call "concavity."

[13] Both propositions presented in this section are straightforward application of definitions. Proofs are omitted.

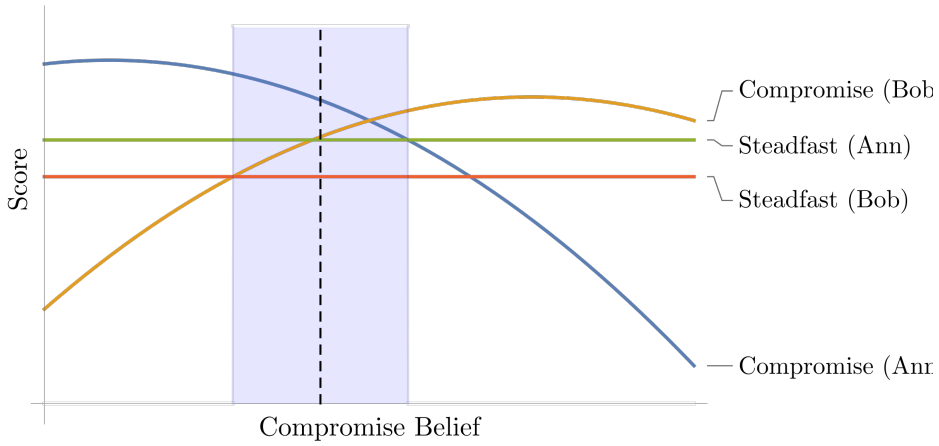

Figure 1: *An illustration of possible compromise positions for Ann and Bob. For this illustration Ann and Bob both employ the Brier scoring rule.* $c_A = 0.1$ *and* $c_B = 0.75$*. The expected joint accuracy for refusing to compromise is given by the two horizontal lines. One line represents Ann's judgment of the consequences of staying steadfast, the other represents Bob's point of view. The other two curves represent the expected score from Ann's and Bob's perspective if they compromise by both adopting the value on the x-axis. When both Ann's and Bob's compromise curves are above their respective horizontal lines, they both prefer the compromise to both sticking with their current belief. That region is shaded. The vertical dotted line represents the average,* $\bar{c}$*, of their beliefs.*

worse for both of them than sticking to their current belief. Bob on the other hand would prefer the compromise. This limits the generality of Moss' point, but in a way that might not be troubling for those who prefer concave scoring rules.

Given that restriction, Proposition 1 appears to provide a strong defense of compromise. Both Ann and Bob would prefer a compromise by averaging their beliefs instead of sticking to their current credences. Not only would Ann and Bob prefer compromising on the linear average of their beliefs, there is a whole range of possible compromise positions that the two might agree on. One illustration is provided in Figure 1 using the Bier score for both S_A and S_B. Figure 2 illustrates how the size and location of the region of compromise changes with different beliefs.

These considerations might seem largely in line with Moss' pro-compromise position. If Ann and Bob both, together, shifted their beliefs to $\bar{c}$ they would both regard their joint accuracy as improved.

While all of that is true, this does not entail that Ann and Bob should compromise. Ann

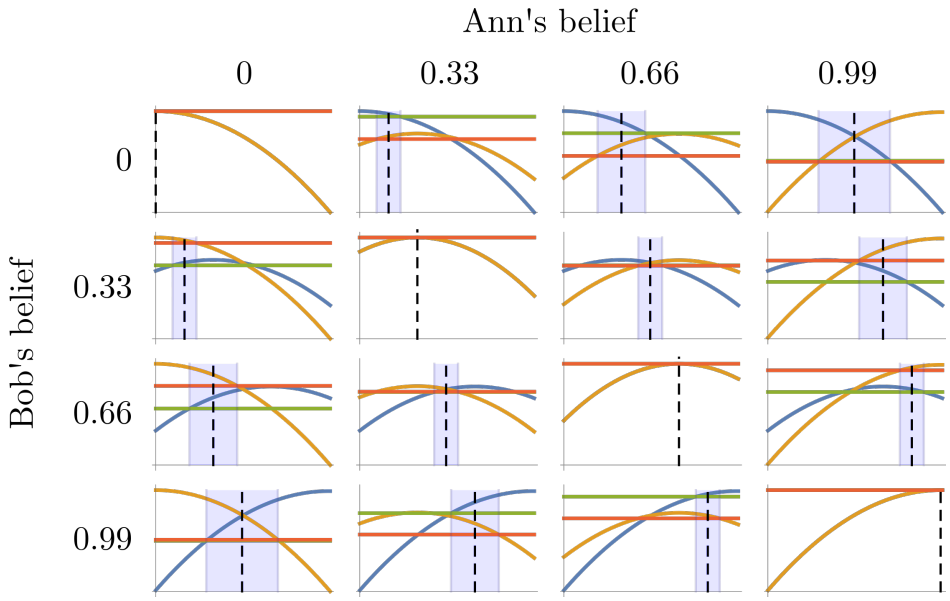

Figure 2: *An illustration of how the graph from Figure 1 changes as Ann's and Bob's beliefs change. Each graph illustrates the compromise position for the corresponding two beliefs listed at the top and left of the figure.*

and Bob are not jointly choosing their credences, each is choosing independently. (What is more private than belief?) So while Ann and Bob would prefer that they both compromised, each has an incentive not to do so. Whatever Bob does, Ann regards it is superior for *her* to stay with her current belief. And likewise for Bob.

Proposition 2. *Suppose that S_A is a strictly proper scoring rule. $JA_A(c_A, \bar{c}) \geq JA_A(\bar{c}, \bar{c})$ and $JA_A(c_A, c_B) \geq JA_A(\bar{c}, c_B)$ (same for JA_B). If $c_A \neq c_B$ then the inequalities are strict*

The two propositions collectively show that Ann and Bob occupy the familiar problem known as the Prisoner's dilemma. In the classic story of the Prisoner's dilemma, both players would prefer that they both cooperate. But regardless of what the other does, each has a private incentive to defect. As a result, cooperation is untenable.

Ann and Bob both agree they would do better by both compromising on their beliefs. Perhaps they might even form an agreement to do so. But, this agreement would be unstable. Ann might say to herself privately, "I know that I promised Bob to change my belief, but that would make us both less accurate. So I should not." And Bob would say the same. Note these propositions are fully general: for any beliefs that Ann and Bob have, so long as they disagree and they care about their joint accuracy, they face a Prisoner's dilemma. This

illustrates that the problem of disagreement is irreducibly a game theoretic problem.

Some scholars argue that one ought to cooperate in the Prisoner's dilemma. While I am highly skeptical of this position, space prevents a complete discussion. Allow me to point out that, unlike the classic story of the Prisoner's dilemma, Ann's and Bob's reasons are not selfish. It is not that Ann is refusing to cooperate with Bob because she is pursuing private ends and ignoring Bob's interests. Instead, she is acting in the way that she believes will benefit them both. She knows Bob disagrees, but she thinks Bob is wrong. While perhaps one might regard this as hubris or a type of epistemic paternalism, it is not the same thing as selfishness.

Even if one believes that there is some epistemic obligation for Ann and Bob to cooperate here, this would require philosophical defense.[14] Given an obligation to compromise, there are other game theoretic problems illustrated in Figures 1 and 2. Notice that there are many possible compromise beliefs in the shaded region, where both Ann and Bob regard compromising as superior to both staying steadfast. Within that region, Ann and Bob face a difficult bargaining problem. Ann would prefer beliefs in that region closer to her current belief (in Figure 1 those are on the left). The same for Bob (in Figure 1 those are on the right). So even if one could convince Ann and Bob that they had an obligation to pursue a strategy that is Pareto superior to both staying steadfast, they would have to figure out which of an infinite number to choose. Some would reward Ann more than Bob, others would do the reverse.[15]

My hope is that this discussion demonstrates my central point: game theoretically complex situations arise even in purely epistemic contexts. First, Ann and Bob face a Prisoner's dilemma. If it is an epistemic or moral obligation for Ann and Bob to cooperate, that is a significant philosophical thesis in need of special defense. That philosophical defense must engage with the game theoretic problem present here, thereby showing that the game theoretic considerations are inseparable from the epistemic situation. Should a philosopher successfully argue that Ann and Bob should cooperate, a distinct game theoretic problem

[14]There remain other thorny problems to deal with if one thinks that Ann and Bob should compromise (e.g. Staffel, 2015).

[15]Notice that this is a problem even for Moss' version of the compromise question. Even if Ann and Bob agree that they must adopt a single belief which represents their "joint" opinion, they will have to bargain over which belief they should adopt.

emerges: how should they divide the epistemic benefits of compromise between them.

This is not the only way that interesting game theoretic phenomena might arise in the context of discussions about disagreement. Heesen and van der Kolk (2016) describe a situation of disagreement where belief is characterized qualitatively. They imagine a scenario where individuals can repeatedly attempt to come to agreement and receive payoffs according to their individual accuracy (they are epistemically selfish). They illustrate – convincingly – that even in this very simple scenario there is substantial game theoretic complexity. In addition to applying to a quantitative, rather than qualitative, framework this paper demonstrates how this complexity remains even when we make individuals epistemically altruistic.

While the Prisoner's dilemma is the most well known game in game theory, it is not the only one of interest. In the next section, I will provide a novel situation where we can generate four of the five canonical two-person symmetric games of game theory.[16]

3 Choosing an experiment

The previous section demonstrated that complicated game theoretic problems always arise in the context of disagreement. One might concede this point but regard disagreement as special. Perhaps this is the only setting where such problems arise. Or perhaps, one might think that the Prisoner's dilemma and bargaining problems are the only types of games that might be of interest. In this section, I will offer a more stylized example which allows me to generate a larger class of games. While there may be few situations that are identical to this idealized example, many of the basic properties will be common to situations where there is joint investigation into a common problem. As such, this simple example should illustrate that game theoretic problems are common in such settings.

[16]The games we will generate are: Prisoner's dilemma, Prisoner's delight, a pure coordination game, and the Stag hunt (a.k.a. assurance game). The only two-person symmetric game we do not create is Chicken (a.k.a. Hawk-Dove). I have yet to be able to create a plausible situation which is uncontroversially epistemic for this game.

3.1 Individual choice

Before we turn to the game theoretic example: let us begin with a simple one-person decision problem to illustrate some important formal properties of scoring rules. Consider a detective, Diego, who is investigating a string of nearly identical robberies. It seems likely that these robberies are being performed by the same crew. Diego has narrowed his list down to two principle suspects: the northern Neerdowells and the southern Sleazeballs.

Diego is a good detective. He knows his intuitions might be wrong, it might be some other group entirely. It is also possible that the groups are working in coordination, performing the robberies together. Diego decides that the probability that each group is involved is probabilistically independent of the other one being involved.[17]

Time is of the essence: Diego must decide which of the two groups to investigate. He does not have the time or resources to investigate both.[18] We will also assume that Diego knows more about the northern Neerdowells. If he investigates them, the evidence he acquires is likely to be more probative; it will more clearly point to the guilt or innocence of that group. But, Diego thinks that they are probably not involved. Instead, he thinks it's more likely to be the unfamiliar group, the Sleazeballs. He knows his limitations, however. He knows that the evidence he will uncover about the Sleazeballs is less informative than the evidence he would uncover about the Neerdowells.

To make this concrete, suppose that the probability Diego assigns to the Neerdowells being involved is 0.1. Diego is maximally uncertain about the Sleazeballs, he assigned their involvement probability 0.5. Suppose further that Diego believes he will turn up some evidence about the group he investigates no matter what, and the evidence will either point to their involvement or their innocence. That evidence will have different probative value. Let N_D^- represent the proposition that the evidence Diego secured points toward the innocence of the Neerdowells. N_D^+ means the evidence points to their guilt. (S_D^+ and S_D^- mean the

[17] Although I have not investigated this completely, I do not believe the independence assumption is critical. However, the case does require that both groups might be involved, that one's involvement is not mutually exclusive with the other's involvement. Making the assumption of independence greatly simplifies the mathematics, and since this is merely for illustration, I take as an acceptable case.

[18] Perhaps one might claim that constraints of this sort bring in "pragmatic" considerations into the purely epistemic. It strikes me that if we define purely epistemic so narrowly as to exclude any notion of constraint, epistemology will become largely irrelevant to most questions relevant to knowledge generation (like what investigations to pursue).

same for the Sleazeballs.) Let N^G and S^G represent the guilt of each group and N^I, and S^I their innocence.

The relevant probabilites are neatly summarized here where $x > y$:

$$P_D(N^I) = 0.9$$
$$P_D(S^I) = 0.5$$
$$P_D(S_D^-|S^G) = x$$
$$P_D(S_D^+|S^I) = x$$
$$P_D(N_D^-|N^G) = y$$
$$P_D(N_D^+|N^I) = y$$

Recall that both groups could be guilty or innocent. S^G and N^G are not exclusive propositions. (But of course no single group could be *both* guilty and innocent.) x is the false positive and false negative rate for investigating the Sleazeballs. It is the probability that given the Sleazeballs are innocent the evidence will point toward their guilt and vice versa. Similarly for y and the Neerdowells.

Diego is a pure epistemic agent: he doesn't care for the politically expedient solution of closing a case. Like his hero Sgt. Friday, Diego cares only for the facts. Specifically, he cares only for the expected accuracy of his own beliefs after he concludes his investigation.

Diego is confronted with a tricky question: should he investigate the Neerdowells or the Sleazeballs? He knows that he will likely uncover very informative evidence about the Neerdowells, but he also thinks he already knows what that evidence will say – that they aren't involved. He might investigate the Sleazeballs, where evidence would be more helpful, but he recognizes that the information is less reliable.

Figure 3 illustrates what choice Diego should make. Depending on the magnitude of x and y, an agent motivated by only epistemic concerns might prefer either investigative choice. When x and y are relatively close in value, then he should investigate the Sleazeballs. When they are not, he should investigate the Neerdowells.

This fact is well understood among Bayesian statisticians: whether an experiment is worth performing depends both on its reliability and one's uncertainty regarding the propo-

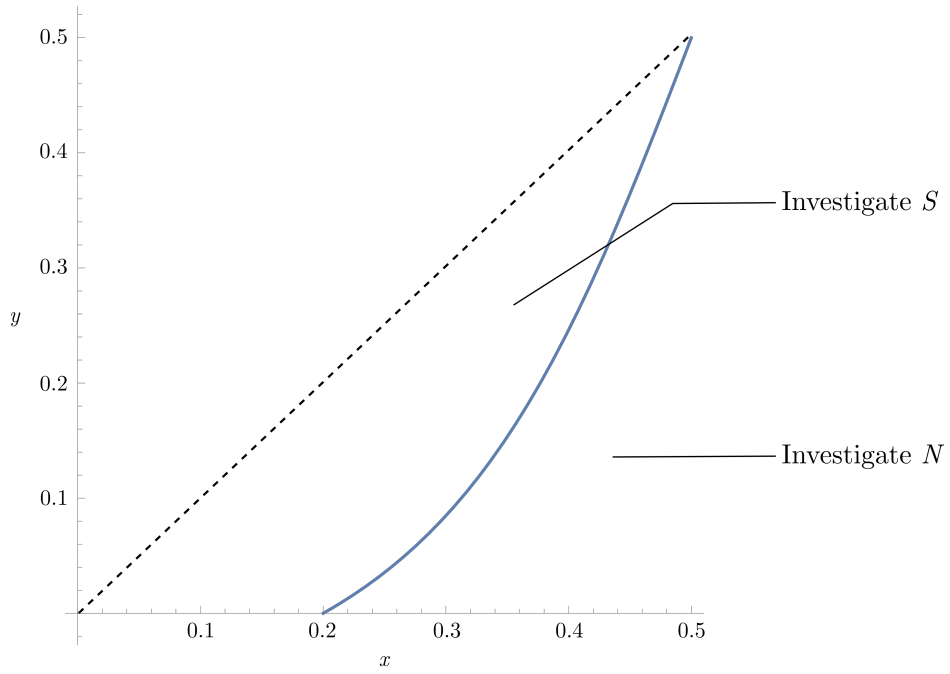

Figure 3: *An illustration of Diego's choices. Each point in the graph represents one possible value for Diego's reliability in investigating each group. The x-axis is Diego's error-rate for investigating the Sleazeballs (x) and the y-axis is his error-rate for the Neerdowells (y). We restrict our attention to the region below the dotted line, when $y < x$. In that region, if Diego's errors are above the solid line, then Diego would prefer to investigate the Sleazeballs; doing so would maximize his expected accuracy. However, if his error rates are below the solid line, then he would prefer to investigate the Neerdowells.*

sitions under scrutiny. If one is already relatively certain about something, even a very reliable test may not be worth performing. This also coincides with everyday life: I am confident that the temperature has not changed radically in the last hour, so I don't bother to check. It doesn't matter that my thermometer is very accurate.

3.2 Multi-person choice

The story of Diego illustrates a trade-off faced by the purely epistemic agent. When their resources are scare, and they are not able to collect all the relevant evidence, they must choose what to investigate. This choice is driven both by the expected reliability of their investigation and also by their current uncertainty.

So far, however, there is no game theory. We have considered only the choice of a single agent. However, the situation becomes more interesting when we have two investigators who have different skills.

Consider, now, two detectives, Diego and Emilia. Diego works the northern area and Emilia works the southern parts of the same city. These two detectives are now jointly assigned the case of the robberies and they have the same two suspects: the southern Sleazeballs and northern Neerdowells.

Like before, each knows the criminals in their regions well. Diego knows the Neerdowells and Emilia knows the Sleazeballs. The detectives are not constrained to their region, each can investigate either group. Because they are less familiar with the group outside their region, they are less reliable when investigating them.

Each will independently choose who to investigate, collect evidence (which we will assume is independent conditioned on the guilt or innocence of the group), and then return to share the results of their work. Each will update their beliefs based on the total evidence collected by both detectives.

Diego and Emilia both have priors over the guilt and innocence of the two groups. They disagree (concrete numbers will be given in a moment). Like Ann and Bob from before, these two have exhausted all methods of convincing each other. For the moment they must agree to disagree. They do, however, agree about each other's abilities: Diego knows that Emilia is a better investigator in the south and Emilia knows the same about Diego in the north.

For the moment, we will make them selfish: they don't care about the other's accuracy. They are each acting to maximize their individual expected accuracy. We will revisit this assumption in a moment, however.

Here are the concrete beliefs of Diego and Emilia:

Diego	**Emilia**
$P_D(N^I) = 0.9$	$P_E(N^I) = 0.5$
$P_D(S^I) = 0.5$	$P_E(S^I) = 0.9$
$P_D(S_D^-\|S^G) = x$	$P_E(S_E^-\|S^G) = y$
$P_D(S_D^+\|S^I) = x$	$P_E(S_E^+\|S^I) = y$
$P_D(N_D^-\|N^G) = y$	$P_E(N_E^-\|N^G) = x$
$P_D(N_D^+\|N^I) = y$	$P_E(N_E^+\|N^I) = x$

Diego and Emilia represent exact opposites of one another. Each believes that the group they are unfamiliar with is more likely to be the culprit. Each is good at investigating their own group, but bad at investigating the other. Now what will happen if each is tasked to choose, independently of the other, which group to investigate? It turns out this picture is far more complicated than the picture with Diego alone. Like that picture, it depends critically on the values of x and y.

The first thing to note about Figure 4 is how strategically rich it is. Depending on the values of x and y, one can construct four of the five canonical two-person, symmetric games. Figure 4 illustrates the locations of various games for different value of x and y. Figure 5 provides examples for each type of game with labeled outcomes.

Let's start with a coordination game. Here, players want to ensure that their investigative talents are distributed across the two different gangs. They both agree *someone* should investigate the Neerdowells and *someone* should investigate the Sleazeballs. The obvious solution would be for both to investigate the group they know well. On the other hand, there is an equilibrium where each investigates the group they don't know well. In our little story, this equilibrium seems implausible. However, real life social groups sometimes coordinate on inferior equilibria (Bicchieri, 2005). These coordination failures are more common in groups larger than two people, and may occur in analogous epistemic situations.

Even with only two, this possibility becomes more realistic when we move to the Stag

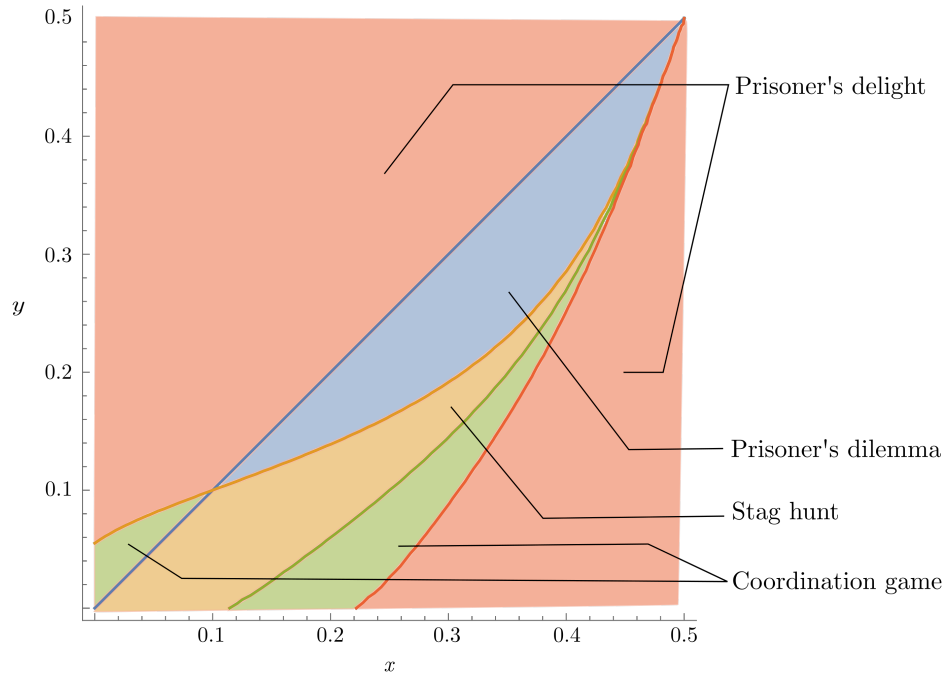

Figure 4: *A figure displaying different game theoretic games played by Diego and Emilia. x and y represent their respective error rates. Each region represents a different potential game theory game. Four of the five canonical two-player, symmetric games occur here (only Chicken/Hawk-Dove is missing.*

	S	N
N	$-0.227382^{\dagger}$	-0.304411
S	-0.204315	$-0.234844^{\star\S}$

$x = 0.2;\ y = 0.19$

(a) *Prisoner's dilemma.*

	S	N
N	$-0.144878^{\star\dagger\S}$	-0.303058
S	-0.17867	-0.32867

$x = 0.4;\ y = 0.1$

(b) *Prisoner's delight.*

	S	N
N	$-0.115831^{\star\dagger}$	-0.282484
S	-0.150008	$-0.234844^{\star\S}$

$x = 0.2;\ y = 0.075$

(c) *Stag hunt.*

	S	N
N	$-0.0191489^{\star\dagger\S}$	-0.258036
S	-0.0996843	$-0.234844^{\star}$

$x = 0.2;\ y = 0.01$

(d) *Coordination game.*

Figure 5: *Illustration of four games for non-altruistic Brier score interactions. All payoffs are for Diego acting as the row player. $\star$ marks the Nash equilibria, $\dagger$ marks a Pareto superior outcome, and § marks the risk dominant equilibrium. The games are symmetric in the following sense: the strategy "investigate what you are good at investigating" represents the same strategy for each player, although it involves investigating a different group. The games are presented in this way in the figures.*

Hunt. Like in the coordination game, there are two equilibria where each group is investigated. However, in the Stag Hunt the inferior equilibrium – where both investigate the group they are bad at investigating – is also "safe."[19] Epistemic safety arises here because each player wants to ensure that a particular group is investigated. And in this context epistemic safety can conflict with epistemic efficiency.

The Prisoner's dilemma arises because Diego would rather have two pieces of evidence about the Sleazeballs rather one piece of evidence about each group. (He would also rather have one piece about the Sleazeball rather than none.) Emilia feels the same way about the Neerdowells. So, each investigates the group they are less efficient at investigating. This is obviously worse than having both switch and investigate the group they know well.

In the scenario I described there might be a possibility for contracts. Diego and Emilia might solve their dilemma by making the following agreement: I will only share my evidence with you on condition that you "cooperate" by investigating the group you know well.[20] Space prevents me from fully exploring this possibility here, but it illustrates how interesting game theoretic considerations are in social epistemology. This would be an interesting case where the threat of withholding evidence might serve an epistemic good.

The last of the possible games, the Prisoner's delight is the happiest of all outcomes. In this setting the efficient outcome is also dominant, meaning that the players will choose to investigate the group they know well no matter what the other one does.

One might criticize the story so far by suggesting that Diego and Emilia are not "purely" epistemic. In all the games presented thus far Diego is only concerned about his own accuracy. Emilia's accuracy does not concern him. (And vice versa for Emilia.) This is in contrast to the story of Ann and Bob from before. One might argue that since Diego and Emilia are only concerned with their own beliefs they are not properly epistemically motivated.

The strategic complexity does not vanish if we make Diego and Emilia altruistic. If we suppose that each wants to maximize their joint accuracy, the picture changes, but the strategic complexity remains. All four games remain in figure 6 that existed before.

[19]Safety has a formal definition in this context. For this region of the parameter space, the inferior equilibrium is *risk dominant* (Harsanyi and Selten, 1988). Skyrms (2004) argues that the Stag Hunt models the fundamental problem of cooperation better than the Prisoner's dilemma.

[20]My thanks to Alexandru Baltag for the suggestion.

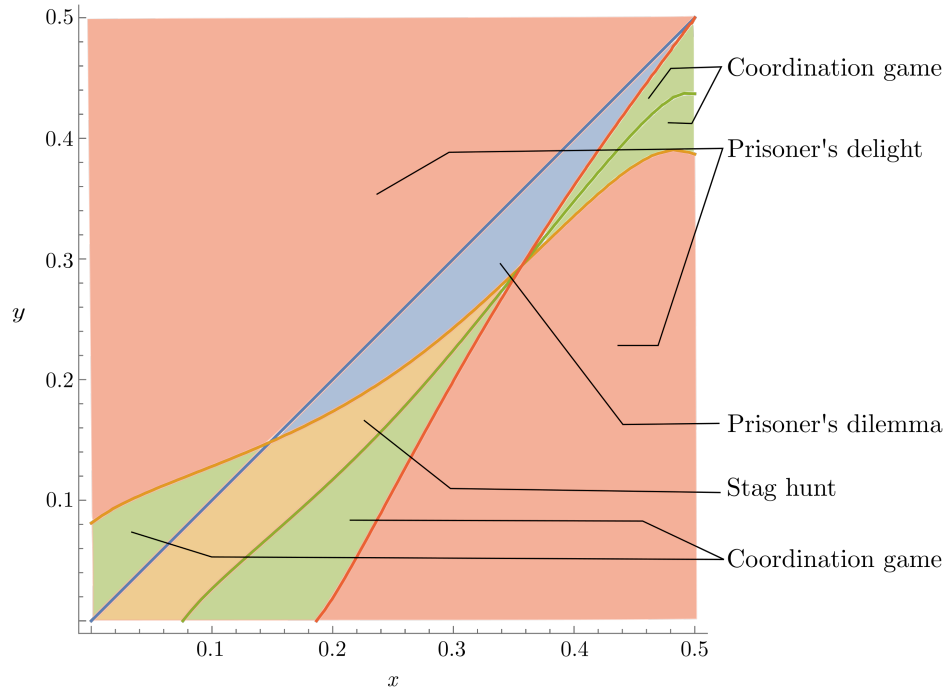

Figure 6: *A figure displaying different game theoretic games played by Diego and Emilia when both are altruistic – they care about the joint accuracy of both their scores. x and y represent their respective error rates. Each region represents a different potential game theory game.*

Even in the altruistic setting, there strategically interesting problems can arise. Game theory remains relevant even in the case of epistemic altruism. I cannot say whether epistemic selfishness or altruism can lay claim to the correct purely epistemic attitude, but it does not matter for my central claim. In both cases, complicated game theoretic problems can arise in purely epistemic settings.

4 Conclusion

In economics, game theory has demonstrated how a substantial disconnect can arise between the aims of individuals and the social outcomes they bring about. Most famously, the Prisoner's dilemma and related games have shown that individuals who pursue one goal can make everyone worse off according to that same goal. Translated to the epistemic case, this shows how individuals pursuing the truth might make the group worse at pursuing the truth (see also Kummerfeld and Zollman, 2016). Dunn (2018) argues that the presence of epistemic Prisoner's dilemmas support the Independence Thesis (see also Mayo-Wilson et al., 2011), that norms of social epistemology are distinct from norms of individual epistemology. In this case, the norms of social epistemology might enjoin individuals to cooperate while norms of individual epistemology might require they defect.

Beyond the Prisoner's dilemma, game theory is replete with complicated social interactions that give rise to difficult and thorny questions about the design of social institutions. What I hope to have demonstrated here is that these very same questions have a natural home in social epistemology. These questions cannot be dismissed as merely a problem for pragmatic or prudential rationality – they are systemic problems present in all forms of social interaction. With the tools of game theory, social epistemologists might hope to provide answers.

References

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

Banerjee, A. V. (1992, aug). A simple Model of Herd Behavior. *The Quarterly Journal of Economics 107*(3), 797–817.

Bicchieri, C. (2005). *Grammar of Society*. Cambridge: Cambridge University Press.

Bovens, L. and S. Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Braithwaite, R. (1954). *Theory of Games as a Tool for the Moral Philosopher*. Cambridge: Cambridge University Press.

Brier, G. W. (1950, jan). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review 78*(1), 1–3.

Carr, J. R. (2017). Epistemic Utility Theory and the Aim of Belief. *Philosophy and Phenomenological Research 95*(3), 511–534.

de Finetti, B. (1975). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons.

Diaconis, P. and B. Skyrms (2018). *Ten Great Ideas About Chance*. Princeton: Princeton University Press.

Dogramaci, S. (2012). Reverse Engineering Epistemic Evaluations. *Philosophy and Phenomenological Research 84*(3), 513–530.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge: MIT Press.

Dunn, J. (2018). Epistemic free riding. In *Epistemic Consequentialism*. Oxford University Press.

Easwaran, K. (2016). Dr. Truthlove or: How I Learned to Stop Worrying and Love Bayesian Probabilities*. *Nous 50*(4), 816–853.

Fallis, D. and P. J. Lewis (2016). The Brier Rule Is not a Good Measure of Epistemic Utility (and Other Useful Facts about Epistemic Betterness). *Australasian Journal of Philosophy 94*(3), 576–590.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford: Clarendon Press.

Greaves, H. (2013). Epistemic decision theory. *Mind 122*(488), 915–952.

Greaves, H. and D. Wallace (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind 115*(459), 607–631.

Harsanyi, J. C. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.

Heesen, R. and P. van der Kolk (2016). A Game-Theoretic Approach to Peer Disagreement. *Erkenntnis 81*(6), 1345–1368.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science 65*(4), 575–603.

Joyce, J. M. (2009, oct). Causal reasoning and backtracking. *Philosophical Studies 147*(1), 139–154.

Kelly, T. (2003). Epistemic Rationality As Instrumental Rationality: A Critique. *Philosophy and Phenomenological Research 66*(3), 612–640.

Kummerfeld, E. and K. J. Zollman (2016). Conservatism and the Scientific State of Nature. *British Journal for the Philosophy of Science 67*(4), 1057–1076.

Levinstein, B. A. (2012). Leitgeb and Pettigrew on Accuracy and Updating. *Philosophy of Science 79*(3), 413–424.

List, C. and P. Pettit (2004). Aggregating sets of judgments: two impossibility results compared 1. *Synthese 140*, 207–235.

Mayo-Wilson, C., K. J. Zollman, and D. Danks (2011). The Independence Thesis : When Individual and Social Epistemology Diverge. *Philosophy of Science 78*(4), 653–677.

Moss, S. (2011). Scoring rules and epistemic compromise. *Mind 120*(480), 1053–1069.

Oddie, G. (1997). Conditionalization, Cogency, and Cognitive Value. *British Journal for the Philosophy of Science 48*, 533–541.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Schervish, M. J. (1989, dec). A General Method for Comparing Probability Assessors. *The Annals of Statistics 17*(4), 1856–1879.

Schervish, M. J., T. Seidenfeld, and J. B. Kadane (2009). Proper Scoring Rules, Dominated Forecasts, and Coherence. *Decision Analysis 6*(4), 202–221.

Seidenfeld, T. (1985). Calibration , Coherence , and Scoring Rules. *Philosophy of Science 52*(2), 274–294.

Seidenfeld, T., J. B. Kadane, and M. J. Schervish (1989). On the Shared Preferences of Two Bayesian Decision Makers. *The Journal of Philosophy 86*(5), 225–244.

Seidenfeld, T., M. J. Schervish, and J. B. Kadane (2012). Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning 53*(8), 1248–1261.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*. New York: Cambridge University Press.

Skyrms, B. (2010). *Signals: Evolotion, Learning and Information*. New York: Oxford University Press.

Staffel, J. (2015). Disagreement and Epistemic Utility-Based Compromise. *Journal of Philosophical Logic 44*(3), 273–286.

Tebben, N. and J. Waterman (2015). Epistemic free riders and reasons to trust testimony. *Social Epistemology 29*(3), 270–279.

von Neumann, J. and O. Morgenstern (1953). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Zollman, K. J. (2018). The credit economy and the economic rationality of science. *Journal of Philosophy 115*(1).