*Metaphilosophy* 41 (3): 255-278

# The Value of Knowledge and the Pursuit of Survival[i]

**Sherrilyn Roush**

*In honor of Luciano Floridi, for his leadership in bringing information theory to epistemology*

***Abstract.*** Knowledge requires more than mere true belief, and we also tend to think it is more valuable. I explain the added value knowledge contributes if its extra ingredient beyond true belief is *tracking*. I show that the tracking conditions are the unique conditions on knowledge that achieve for those who fulfill them a strict Nash Equilibrium and an Evolutionarily Stable Strategy in what I call the True Belief game. The added value of these properties, intuitively, includes preparedness and an expectation of survival advantage. On this view knowledge is not valuable because *knowledge* persists, but because it makes the bearer more likely to maintain an appropriate belief state – possibly non-belief – through time and changing circumstances. When Socrates concluded that knowledge of the road to Larissa was no more valuable than true belief for the purpose of getting to Larissa, he did not take into account that one might want to be prepared for a possible meeting with a misleading sophist along the way, or for the possibility of road work.

***Keywords.*** ESS, evolutionary stability, knowledge, Nash Equilibrium, swamping problem, tracking, value problem.

It is thought that externalist views of what knowledge is – that do not require conscious access to reasons and arguments but only certain relations in which a person's belief must stand to the world – have trouble explaining why knowledge is more valuable than mere true belief. Justificatory arguments are intrinsically valuable, some say, but what additional epistemic worth could another relation of your belief to the external world have if the belief already has the relation of being true? (Swinburne 1999, 2000 , Kvanvig 2003)

This surely depends on what the further relation is. The value problem, or "Swamping Problem" as it is called since the property of truth of the belief seems to swamp in significance other external relations, appears particularly acute for process reliabilism, but process reliabilism is not the only externalist view. Besides truth the process reliabilist puts

constraints only on the history of formation of a belief, but once you actually *have* a true belief the extra property of its having been formed in a certain way seems otiose. According to the standard comparison, a beautiful chair does not have additional aesthetic worth for having been produced by a process that produces beautiful chairs most of the time.[ii] If these points stand, then the swamping problem appears to afflict all historical views of knowledge.

However, not all externalist views of knowledge are historical. Counterfactual views do not impose conditions on the genesis of a belief. They put conditions on how the belief is currently disposed to behave or fare in scenarios different from the actual one. The distinction here is analogous to the difference between how the solar system was formed and what laws govern its motions. The laws govern the motion of a planet at every point in time, even if the planet is not in fact moving. The history of its formation is a different matter; it will conform with these laws but involve a lot more information, about initial conditions for example. This is where the analogy ends, of course, for we, unlike the law-governed physical world, are capable of forming beliefs by processes that are not well-behaved in the relevant way, and of forming beliefs that do not conform in their dispositions to any epistemologically nice counterfactual properties. The point is that these are distinct failures. Even if there are correlations between them in the actual world, as there probably are, those would be contingent relationships. The first failure is a defect of the process of forming a belief, the second a defect in the product. The process reliabilist thinks that the first type of failure – formation by a process that does not tend to produce true beliefs – is what deprives a belief of the status of knowledge even if it happens to be true. A counterfactualist thinks the failure of a true belief to be knowledge is a defect in the dispositions that accompany the fully formed belief.

Since ascription of a counterfactual concerning a belief is ascription of a current property, counterfactual views of knowledge are "current time-slice" views, in Alvin Goldman's terminology (Goldman 1979), a property they share with traditional internalist justified-belief views. Accordingly, the value problem for these views looks entirely different from that for process reliabilism. Here the question becomes whether a person's disposition to believe or not believe a proposition p in non-actual situations could add value when she already actually has a true belief in p. It is clear intuitively that counterfactuals might have something to offer here. After all, your spouse's not actually having an affair with Mr. or Ms. X is a good thing, but it would surely be strictly better if it were also the case that he or she wouldn't have that affair even if offered a million dollars. The latter is evidently not swamped by the former.

Love might be a case where the counterfactual enhances an intrinsic kind of value. I am going to argue that if knowledge requires tracking then it enhances the extrinsic value of a true belief, the value it has for achieving or obtaining other things. That is, it will turn out that Socrates was wrong to think that knowledge of which road went to Larissa would be no more valuable than mere true belief about it. I will show that the additional dispositional properties required by the tracking view of knowledge, formulated using conditional probabilities rather than counterfactuals, add payoff and survival value necessarily and that no other conditions

on knowledge have the property that ensures this necessarily. This follows because tracking is the unique Nash Equilibrium and Evolutionarily Stable Strategy in what I will call the True Belief Game.

Intuitively, what fulfillment of the tracking conditions adds to the truth of a belief is a kind of robustness against contingencies. What can be questioned in taking this analysis as a resolution of the value problem is whether robustness of a person's belief behavior as the subject is faced with a world that evolves over time is of value to a person at the time of holding the belief. I will discuss this after explaining the kind of robustness in question.

To begin, consider the situation of a person playing a game with the world, which I will call Nature, on a single occasion. Nature can play p or –p, p a proposition, and the person can play B(p), that is, believe p, or play -B(p), that is, not believe p. Suppose the person's payoffs are positive if she plays B(p) when Nature plays p, and positive when she plays -B(p) to Nature's -p, and they are negative when she plays -B(p) to Nature's p and when she plays B(p) to Nature's -p. These payoffs express the conditions that when p is true, it is more valuable to the subject to believe p than not to believe p, and when p is false it is more valuable to the subject to not believe p than to believe p. The results I am explaining are limited to these conditions, but that is not a limitation on their application to the value problem. There are plenty of p for which it is more valuable to have a false belief than a true belief or no belief – think of crazy metaphysical beliefs that come bundled with other, true, beliefs holding all of which is required to cement your relation to your social group, and consider a situation – of pioneers, for example – where survival depends on membership in a group. However, these cases of p for which true beliefs are not valuable are not relevant to the value problem under discussion here, which is to say whether or how given that true belief is valuable, knowledge has added value. Anyway, cases of p for which true beliefs are not valuable are not cases where we expect knowledge to be valuable either.

Nature is indifferent to what you play when it plays p or -p. It gains nothing and loses nothing, so in our first pass here we are dealing with a degenerate game:

| | B(p) | -B(p) |
|---|---|---|
| p | (0, 2) | (0, -3) |
| -p | (0, -1) | (0, 3) |

Nature is the player choosing a row, and the subject is the player choosing a column. The winning strategy for the subject is to play B(p) when Nature plays p and to play -B(p) when Nature plays –p, which is reflected in the payoffs in the four possible situations written as

ordered pairs with the subject's payoffs second.[iii] The word "strategy" does not imply, in this or any other game I will discuss here, conscious or consciously accessible planning, or even thought. It is simply an intuitive term for what the player does; later in the discussion strategies will be rules in accord with which players act in a given round of play, and we will discuss dispositions to act in accord with a given rule, but doing or having any of these does not require conscious access to them either. We do not assume that the subject knows (or does not know) what Nature played before playing B(p) or -B(p).[iv] Nor are the players assumed to know the structure of the game or their or their opponent's payoff structure. We think merely of which of their options each of the players play, and what they get when they do. This game can thus be compared with Floridi's more sophisticated Knowledge Game (2005). Here we will see how knowledge emerges from a game in which the object is true belief and no knowledge of any sort is assumed. There common knowledge is assumed in order to show how second-order knowledge that one is conscious can emerge from the fact that there is a game one can win that a zombie could not, and that one can see that one can win it and a zombie could not.

This simple game is a way of formulating what is essential in forming beliefs about matters like p in situations where the truth of p matters positively to us; it formulates the starting point of the value problem. The game expresses the assumption that at a given time a true belief about p is not trivially acquired and is valuable by saying that whether p is true or false is not determined by the subject, but by a different player, Nature, and that a correct belief state about p (whether that means believing it or not believing it) has positive value for the subject; it makes her win the game and achieves her best possible outcome given Nature's play.

Thus, the assumption that a merely true belief has value can be represented as a certain kind of payoff structure in a one-shot game played with Nature. In real life, even with a single p, this game would often be repeated over time. The truth value of p may change; the tiger may be gone today, but back tomorrow. We will imagine this repeated play with p and -p understood as states of the world, and belief and absence of belief in p understood as acts of the subject. I will represent this as a more elaborate type of game, which I will introduce using an example about something other than belief. In this game, the states of the world may be different at each round of play, and the players may opt for different acts at each round of play. If we also imagine messages interposed between states of the world and acts of the subject, then we have what David Lewis called a "Signaling Game." (Lewis 1969) Such a game has the following kind of structure

| States of World | Messages Sent | Acts |
|---|---|---|
| p | $m_1$ | Watching football |
| q | $m_2$ | Self-reflection |

There is a Sender and a Receiver in the game, and each will have payoffs associated with strategies for responding to scenarios. Sender is defined by her repertoire of possible plays: here either $m_1$ or $m_2$ when Nature sets p or q. Receiver is defined by his repertoire of possible acts, here Watching football or Self-reflection, when Sender plays $m_1$ or $m_2$. In order for there to be a game of this sort, Sender and Receiver must each have the capacity for a variety of rules of responding when the other player plays in each of the possible ways he might. Thus, if the possible messages are $m_1$ and $m_2$, Sender must be able to respond by taking p to one of these, and taking q to one of these. Sender may send both states to one message or send p and q to different messages and it still be a signaling game. Similarly Receiver must have the ability to act either of his two ways, Watching Football or Self-reflection, and in any of the four possible permutations of rules for responding to each of $m_1$ and $m_2$.

A full set of such rules, that is, that covers the possibilities a player may be paired with, is called a *strategy*. Thus Sender's strategy is a set of two rules, for example, $T_1$:

$p \rightarrow m_2$

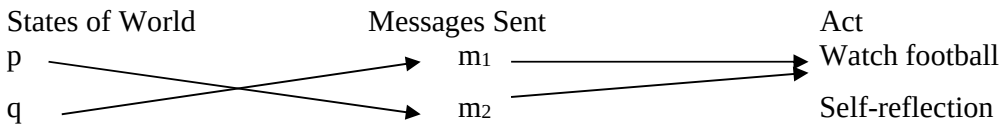
$q \rightarrow m_1$

Likewise for Receiver, for example, $L_1$:

$m_1 \rightarrow$ Watch football

$m_2 \rightarrow$ Watch Football

In a picture:



Sender and Receiver each have a repertoire of other possible strategies that is easy to list:

| | | |
|---|---|---|
| $T_2$: | p → $m_1$ | q → $m_1$ |
| $T_3$: | p → $m_2$ | q → $m_2$ |
| $T_4$: | p → $m_1$ | q → $m_2$ |

| | | |
|---|---|---|
| $L_2$: | $m_1$ → Self-reflection | $m_2$ → Self-reflection |
| $L_3$: | $m_1$ → Watch football | $m_2$ → Self-reflection |
| $L_4$: | $m_1$ → Self-Reflection | $m_2$ → Watch football |

We can assess all possible outcomes for Sender and Receiver in this game by looking at the payoffs for their possible strategies when each is paired with each possible strategy of the other player. This is because a strategy pair, one from each player, determines what each will do whether the world is p or the world is q, the only two world states stipulated to be relevant to our game.

Thus, we consider payoffs for all combinations of $T_1$, $T_2$, $T_3$, $T_4$ with $L_1$, $L_2$, $L_3$, $L_4$. If we were modeling an actual situation we would take these payoffs from the facts. How nice or nasty is a certain consequence for a given player, and how likely is that consequence if he plays a particular way and the other player plays a particular way? In this example I will make the payoffs up, to illustrate some key points.

| | $L_1$ | $L_2$ | $L_3$ | $L_4$ |
|---|---|---|---|---|
| $T_1$ | (-1, 2) | (2, 0) | (-2, -1) | (2, 0) |
| $T_2$ | (-2, -2) | (0, 3) | (-1, 2) | (2, -1) |
| $T_3$ | (0, 1) | (2, -1) | (-1, 2) | (1, -2) |
| $T_4$ | (-2, -1) | (2, -1) | (3, -2) | (-2, 1) |

The first number of each ordered pair is the payoff for Sender when she plays the strategy to the left, and the second is what Receiver gets when he plays the strategy at the top of that row. The numbers could be anything, but the ones I have entered for the current example imply that, for example, if Sender plays $T_3$ and Receiver plays $L_2$ then Sender gains 2 and Receiver loses 1. If Receiver plays $L_4$ to Sender's $T_4$, Receiver gains 1 and Sender loses 2. A feature that I have written in to this particular assignment of payoffs is that there is no one combination of plays (square in the table) that will have both Sender and Receiver better off compared to their other options in that row or column respectively. This implies that if we set Sender and Receiver to play the game in perpetuity, neither of them would settle on one strategy out of their repertoire.[v]

Although the definition of these games does not involve any assumption of intentionality, knowledge, or information transmission, the terms "message," "sender" and "receiver" are meant to be suggestive. This is because phenomena we recognize as information transmission can arise naturally out of the games. We can see how by looking at what the scenario just imagined means intuitively. Suppose Sender plays $T_1$ and Receiver $L_1$, as in the picture above. Sender's dispositions in $T_1$ mean she is cued in to there being a difference between p and q and is revealing that difference by differentiating between them in a uniform way in the messages she sends out. Receiver's dispositions in $L_1$ mean he does not register that difference in his act of Watching football or Self-reflection. Receiver's indifference to the distinction between $m_1$ and $m_2$ in his responding act makes the information about the state of the world, p or q, unavailable to Receiver; we could say that he is not listening. According to our payoff table, and assuming a simple dynamics, it follows that this set of dispositions is not a configuration our two players would stably end up in, since though it is beneficial to Receiver not to hear (+2), it is detrimental for Sender not to be heard (-1), and this is not better than all the players' other options. We can imagine this as a realistic case where Sender has an interest in communicating and Receiver has an interest in not hearing. Since every other square also involves some analogous mismatch of best interest, it follows that, other things equal, this relationship will not become stable no matter how many rounds they play.

The basic condition under which they would become stable is there being a square where both receive the highest payoff they could get given the strategy the other has played. This would happen here, for example, if the top left corner had the payoffs (4, 4). If in the course of play the players happen on such a combination, and stick with it for a while, then they will become stable and resiliently wedded to the corresponding strategies indefinitely.

That there be a possibility of stable convergence depends on there being a payoff possibility reflecting common interest. However, there being a common interest does not imply that the interest is in what we would intuitively call communication, or information transmission. Witness that nothing prevented us imagining the highest mutual payoff in the square $T_1$, $L_1$where Sender makes information available, but in fact is never heard. A strategy combination where Sender talks into the wind can become stable as long as Sender is

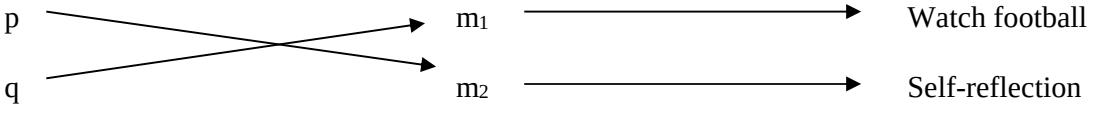
satisfactorily rewarded in it. Whether a stabilizable configuration is also communicative depends on which strategies in fact have the highest payoffs. Thus, the issue of stability and the issue of communication are conceptually independent.

The reason the new payoff structure leads to convergence to a single set of strategies is that if the top left corner is (4,4), then this option dominates every other possible set of strategies. That is, it is better, for both players, than any other option either of them has. This dominating ordered pair of strategies is called a Strict Nash Equilibrium, and such a pair of strategies in a signaling game is called a *Signaling System*. As we have just seen, not every signaling system will lead to what we would intuitively call communication or information transmission.

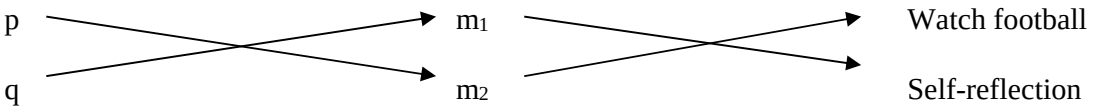
The strategy pairs in our game that would intuitively correspond to information transmission are $<T_1, L_3>$, $<T_1, L_4>$, $<T_4, L_3>$, $<T_4, L_4>$. In all of them Sender can tell, and signals, the difference between p and q, and Receiver can distinguish those messages and respond differentially. Notice that the messages $m_1$, $m_2$ have no preassigned meanings. As long as Sender consistently sends the same message for p each time and q each time, and sends different messages for p and q, it does not matter which of $m_1$ or $m_2$ she uses in which role. As far as whether it is possible to achieve a signaling system or information transfer the qualities of $m_1$ and $m_2$ are conventional. Receiver will be able to cotton on to p versus q if he has available a strategy that can respond differentially to $m_1$ and $m_2$.

Responding *differently* to p and q is one thing, one might say, but what about responding correctly? If the character of the messages is completely conventional, couldn't Receiver get his responses to p and q exactly the wrong way around? The answer is No. Which is the *right* response to p for Receiver is determined by his payoff structure – how good is it for him when he watches football in the situation p? But his payoff structure also determines which combined sets of strategies of him and his playmate can become stable Signaling Systems. Thus, here are the four possible intuitively communicative Signaling Systems in our game:

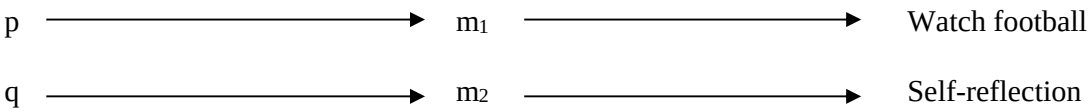
$T_1, L_3$



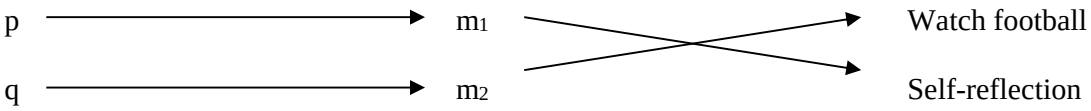

$T_1, L_4$



$T_4$, $L_3$




$T_4$, $L_4$




Suppose it is appropriate for Receiver to respond to state p by Watching football and to q by Self-reflection. These assumptions will be reflected in his payoffs. His rewards will be highest, among all the possibilities, for <$T_1$, $L_4$> and <$T_4$, $L_3$>. The rewards will be the same for both of those strategies because each has him successfully responding to p with Watching football and to q with Self-reflection; it doesn't matter which message route he took to get there. Accordingly, if all of these communicative solutions are winners for Sender, then, according to theorems, Sender and Receiver will be stable in either of <$T_1$, $L_4$> and <$T_4$, $L_3$>. Receiver will have a system in which he is able to make what is for him the appropriate response to p, because the appropriateness will be reflected in his payoffs, and which system is stable, if any, is determined by the payoffs for him and the other player.

To model the task of getting true beliefs about the physical world as a signaling game with two players and repeated play, we take the possible States of the World to be p and -p. We take Sender to be the laws of Nature, those things that determine which indicators flow downstream from the State of the World, p or –p. Receiver is a player about whom we stipulate the values of true belief, false belief, and no belief as in the value problem about knowledge, thus, as in the very first table above. Sender and Receiver each have four possible strategies:

Sender:

| | | |
|---|---|---|
| $N_1$: | p → $M_1$ | -p → $M_2$ |
| $N_2$: | p → $M_1$ | -p → $M_1$ |
| $N_3$: | p → $M_2$ | -p → $M_1$ |
| $N_4$: | p → $M_2$ | -p → $M_2$ |

Receiver:

| | | |
|---|---|---|
| $K_1$: | $M_1$ → B(p) | $M_2$ → - B(p) |
| $K_2$: | $M_1$ → -B(p) | $M_2$ → -B(p) |
| $K_3$: | $M_1$ → B(p) | $M_2$ → B(p) |
| $K_4$: | $M_1$ → -B(p) | $M_2$ → B(p) |

| | $K_1$ | $K_2$ | $K_3$ | $K_4$ |
|---|---|---|---|---|
| $N_1$ | (0, 4) | (0, 1) | (0, -2) | (0, -5) |
| $N_2$ | (0, -1) | (0, -1) | (0, -1) | (0, -1) |
| $N_3$ | (0, -5) | (0, 1) | (0, 1) | (0, 4) |
| $N_4$ | (0, -1) | (0, -1) | (0, -1) | (0, -1) |

The payoffs in this table are not assigned by thinking of values as possessed by the strategies, but as possessed by particular types of outcomes, in our case the reward or punishment for the having or lacking of a true belief. The payoffs could not belong to the strategies per se, because those achieve different payoffs depending on the strategy played by the other player. We assume that a merely true, particular belief p occurring at a particular time is valuable, more valuable than no belief when p is true, etc., and payoffs in our tables always refer to that level of fact. A strategy, which I will eventually associate with that part of knowledge that goes beyond true belief, is obviously instrumentally valuable insofar as it actually gives you one of these valuable things. However, that fact alone would not resolve the swamping problem, since it would still be the case that if you had that true belief you wouldn't need that instrument. What I will eventually show is that a strategy, which in a distinctive way attaches to an actual true belief, adds value beyond that which comes from the truth of the particular belief it is a good tool for getting.

Assuming true belief, false belief, and no belief have the relative values we stipulated at the start, the worst combinations for our subject would be <$N_3$, $K_1$> and <$N_1$, $K_4$>. Both combinations have Receiver believing p when it is false and not believing when p is true, and those belief states are bad for him. The best combinations for Receiver are <$N_1$, $K_1$> and <$N_3$, $K_4$>, since in these cases Receiver believes p when it is true and does not believe p when it is

false, and those outcomes are good for him. In our True Belief game Receiver's payoffs happen to be the worst and the best respectively in these two strategy sets.

Whether a repeated game of this kind converges to a stable signaling system also depends on the payoffs of the other player, who in our case is Nature. It is commonplace in proving epistemological convergence theorems to assume that Nature is cooperative in making separating evidence available, that is, in providing distinct indicators of distinct states of affairs. We cannot expect a subject to gain information about Nature if she obfuscates. The question is then whether the subject's tools will enable him to find or use the messages appropriately. Here this assumption will correspond to Nature's having a preference to play strategy $N_1$ or $N_3$, indifferently between the two, regardless of anything else, since in those two strategies, and only those, distinct messages are given for distinct states of affairs. Writing that in:

| | $K_1$ | $K_2$ | $K_3$ | $K_4$ |
|---|---|---|---|---|
| $N_1$ | (1, 4) | (1, 1) | (1, -2) | (1, -5) |
| $N_2$ | (-1, -1) | (-1, -1) | (-1, -1) | (-1, -1) |
| $N_3$ | (1, -5) | (1, 1) | (1, 1) | (1, 4) |
| $N_4$ | (-1, -1) | (-1, -1) | (-1, -1) | (-1, -1) |

The two best outcomes for Receiver occur in blocks that are also best outcomes for Sender, $\langle N_1,K_1\rangle$ and $\langle N_3,K_4\rangle$, but Sender is indifferent between those. These squares are Nash Equilibria because neither player can do better, given that play of the other, by going to some other square. However, they are not strict because neither of these dominates all options for Sender.

We can find a strict Nash Equilibrium by focusing more closely on the fact that in our game Sender is Nature. It is her laws that determine how she responds to states of the world and produces indicators for them; $N_1$ and $N_3$ are two different possible sets of laws that are equally capable of delivering separating evidence, and that is all we took to matter to her. Yet in a real game Nature, as usually conceived, does not change her laws with each round. She would have chosen $N_1$ or $N_3$ in the beginning, and the repeated game would be a degenerate one that Receiver plays against the background of that one strategy. (In a moment we will see that he plays it in competition with other receivers.) If we assume that the world has only one set of physical laws, then either $N_1$ or $N_3$ will be the unique play Nature always makes. If so, then either $K_1$ or $K_4$ (depending on which laws Nature chose) will be the unique best response of Receiver, and either $\langle N_1, K_1\rangle$ or $\langle N_3, K_4\rangle$ will be a strict Nash Equilibrium.

Since there is symmetry, we can suppose without loss of generality that Nature chose strategy $N_1$. The state of the world remains as it was, potentially changing in each round

between p and -p. Now Sender (Nature) is a degenerate player who is like a set of background conditions, and the game that is left involves a confrontation between the Receiving strategies. We can rewrite this as a non-degenerate game by imagining many Receivers playing with each other. They meet two by two, round by round, each does his thing with Nature and each gets a certain payoff determined by what his payoff was in the previous table when playing against $N_1$. It is not that these two players necessarily interact or oppose each other, simply that each may do better or worse than or the same as the other with which he is paired in a given round; each competes to get higher payoffs than the opponent, as in darts, but not necessarily as in football. What we are now doing is comparing what one's outcomes in the True Belief game would be were one to be this kind of Receiver or that. There are four types of Receiver, each defined by his strategy when faced with $N_1$:

| | $K_1$ | $K_2$ | $K_3$ | $K_4$ |
|---|---|---|---|---|
| $K_1$ | (4, 4) | (4, 1) | (4, -2) | (4, -5) |
| $K_2$ | (1, 4) | (1, 1) | (1, -2) | (1, -5) |
| $K_3$ | (-2, 4) | (-2, 1) | (-2, -2) | (-2, -5) |
| $K_4$ | (-5, 4) | (-5, 1) | (-5, -2) | (-5, -5) |

$K_1$ dominates all other possibilities – it does better than every other possibility in the payoffs – and so is a strict Nash Equilibrium. It is a consequence of this that if you were to always play $K_1$, i.e. play that in every round of the game, then you would always do better than if you had played any other possible strategy. That is, it is not only that a true belief has value but that there is a unique strategy that will deliver a correct belief state no matter what, in particular no matter whether p is true or false (and even if the strategy is offered a million dollars to do otherwise). This added general guarantee over several dimensions of possible variation is what will yield an answer to our value question about knowledge, as I will discuss after relating the present concept of strategies to theories of knowledge.

In this latest game we imagined an individual player being of a certain type, corresponding to a strategy, and him having a true or false, or lack of, belief in each round. I will explain below how what his strategy is can add value, in each round, to his having actually achieved a true belief. This is an advantage that accrues to an individual when he is of a favorable type, but the very same facts also guarantee a value added for a population of individuals of his type. The reason is that a strict Nash Equilibrium in a symmetric game like this one is an Evolutionarily Stable Strategy (ESS). This is a notion used to evaluate the fate of subpopulations of uniform types, here four subpopulations for the four types of Receiver for whom true belief about p is valuable. The proportions of these types in the population change with each round as any individual player's strategy in the next round will be the one determined to be the best by set rules of interaction dynamics applied to his and possibly

others' outcome(s) in this round. The question, as with biological evolution, is how the proportions of the four types evolve with each generation or round of play. The interesting implication of a strategy's being an ESS is that if it comes to be widespread in the population, it will be uninvadable by a mutant strategy; that is, no other single strategy that exists or arose in small numbers could drive this type to extinction.[vi] This is a powerful property because it holds, when it does, no matter what the dynamics of interaction are as the game evolves from round to round (and there are a potentially infinite number of possible interaction dynamics). The basic upshot of this for our case is that if the $K_1$ strategy gets a good start it will be the unique type that rebuffs every competitor in the True Belief Game.

The value, and relative value, of the $K_1$ strategy is the key to answering the value problem for the tracking view of knowledge. We can see what $K_1$ has to do with probabilistic tracking, and knowledge, by looking more closely at what a strategy is and what the winning strategies in the True Belief game look like. Assuming, as above, that Nature chose $N_1$ for her laws, $K_1$ is an ESS and strict Nash Equilibrium (sNE):

| | | |
|---|---|---|
| $N_1$: | $p \rightarrow M_1$ | $-p \rightarrow M_2$ |
| $K_1$: | $M_1 \rightarrow B(p)$ | $M_2 \rightarrow -B(p)$ |

Since $N_1$ is simply assumed to be the case, the player who always uses $K_1$ has a relation to the world, that is, to p and -p, that we can think of as a result of the combined rules $N_1$ and $K_1$. The arrows in these diagrams are normally written in terms of conditional probabilities. So, in the simplest terms, a commitment to following the $K_1$ strategy on assumption of $N_1$ would be written:

$Pr(M_1/p)$ = very high, and $Pr(B(p)/M_1)$ = very high *

$Pr(M_2/-p)$ = very high, and $Pr(-B(p)/M_2)$ = very high **

Notice the similarity of * and ** to the two probabilistic tracking conditions[vii], respectively:

$Pr(B(p)/p)$ is high †

$Pr(-B(p)/-p)$ is high ‡

The two sets of conditions cannot be unconditionally identified because conditional probability is not transitive. However, under the following screening off conditions:

$Pr(B(p)/p.M_1) = Pr(B(p)/M_1)$ and

$Pr(-B(p)/-p.M_2) = Pr(-B(p)/M_2)$

which intuitively say that $M_1$, $M_2$ are the only messages Receiver is getting about p, * and ** imply † and ‡ respectively. (See *Appendix*.) That is, *if you are a faithful follower of the Strict Nash Equilibrium or Evolutionarily Stable Strategy for a given p in the True Belief game, then you fulfill the tracking conditions on knowledge for that p*. This means that no theory of knowledge that does not impose the tracking conditions implies that knowledge gives us an ESS or sNE. If following the rule of an ESS or sNE adds value to having a true belief, then it is a value that only tracking can give. Other conditions on knowledge may have other (non-tracking) properties that make knowledge more valuable than mere true belief and thus address the swamping problem, but those properties must be strictly logically weaker than ESS or sNE, and they will not dominate in the True Belief game, a representation that does seem a fair way of depicting what our task is in forming beliefs.

To see the other direction of relationship between the tracking conditions and the True Belief ESS and sNE conditions, we must consider that the tracking conditions are highly abstract, even more abstract than our imagined signaling game. Magic could be the truthmaker of the counterfactual conditions if magic existed. They involve no requirements that there exist a process of belief formation, or causal connection, the things we familiarly use to get to a knowledge state. How a subject manages to achieve fulfillment of the tracking conditions is not restricted by these conditions for what knowledge is. However, it happens to be a contingent fact about human beings that we can't fulfill the tracking conditions without intermediaries: causal processes, one event indicating another, one trait correlated with another, our eyes, our brains, having dispositions to respond differentially, testimony of witnesses, etc. The minimal description of what these intermediaries give to us that is sufficient to insure tracking is indicators playing the role of messages, $M_1$ and $M_2$ in a signaling system. Thus, what we can say is that if a *human being* fulfills the tracking conditions for a given p then *there are* $M_1$, $M_2$ such that she has an ESS and a sNE.

Having had your belief formed through a reliable process does not imply that your belief also has the tracking properties. Nor does the counterfactual property of safety. (Roush 2005, 118-126) Being justified in your beliefs, or virtuous, also do not yield tracking. (Neither do the advocates of these conditions intend them to.) These alternatives to tracking are all nice properties, but they do not give you an ESS. However, one might be bothered about the fact that none of those conditions implies strategy $K_2$, $K_3$, or $K_4$ either. Having used a reliable process doesn't insure that you would believe p if p were true, but it doesn't

guarantee that if it were true you *wouldn't* believe it either. Are we really comparing actual theories of knowledge at all in the True Belief game?

None of the other theories' conditions on knowledge imply any single one of the four pure strategies, yet they – and every other possible set of conditions – are taken account of in our game. Any set of conditions for knowledge that a subject fulfills will have consequences for whether or how often, or likely it is that, the subject will end up believing p when p is true and avoiding belief when p is false, with possibly different probabilities depending on a variety of conditions. Having formed your belief in p through a reliable process need not imply that you will believe p when it is false in order to confer some probability of doing so given the way the actual world works, or under certain conditions. Being justified in whatever way one prefers may not imply your avoiding belief in p when p is false given the way the actual world works, but there may be some, even significant, probability, .x, of it, and thus a 1-x probability of believing p when p is false.[viii] Similarly for the other rules in the True Belief game. In this way, any conditions on knowledge that are added to the truth and belief conditions is represented as some "mixed strategy" in the True Belief game. It is a fact that because $K_1$ is a sNE and ESS, it not only cannot be invaded by any of the other pure strategies $K_2$ – $K_4$, but none of the mixed strategies can invade it either. So the uniqueness of tracking as an ESS (sNE) is completely general over theories of knowledge that see knowledge as true belief plus a further condition. There may be conditions that are not the tracking conditions but do imply them, and those conditions would also count as an ESS (sNE). But since it would be only in virtue of implying the tracking conditions that they guaranteed that stability, it is still the tracking that confers the value that an ESS and a sNE bring.[ix]

When speaking of strategies in these games I have used locutions in which the players play a strategy or follow a rule, because they are less misleading than talk about choosing options, since the former can be done unconsciously. When a subject needs to know p she rarely just finds herself choosing between her tracking option and other options that could lead to error. There are matters on which a human being does come naturally equipped with a tracking ability – e.g., her eyes can track whether there is a tiger in front of her – but in those cases we would not think of her as, and she would not typically be, choosing to use her eyes. Her doing that is automatic. In nonperceptual cases the subject often would choose the tracking option if she could, but it is not just there for the choosing. A scientist would have to build a hadron collider in order to set up a set of messages that distinguish, for example, the existence and non-existence of the particle of interest. Journalists, and many other types of knowledge-seekers, have to do work to set up a tracking relation with the truth of interest. Even the locutions of "playing a strategy" and "following a rule" can be misleading for the non-automatic cases to the extent that they suggest a mere decision.

We can think of these strategies in the True Belief game as rules that with repeated achievement and use can become dispositions of subjects, but the notion of following a rule has an ambiguity worth clarifying here too. One may follow a rule that corresponds to one of the strategies by using its response types in a given round of play. However, that does not

correspond to tracking; tracking corresponds to playing or following that strategy or rule as a habit or disposition with respect to p – it requires that a player is of a type.

How a subject can or does get herself into the position of having strategies available is not a concern of this game-theoretic analysis. One may stumble into doing a tracking type play, one may be automatically disposed to it, one may make a herculean effort to get to be able to choose and commit to use it. Also, it is the same if you had a strategy available and didn't use it or didn't use it because you didn't have it. What matters to the outcomes and stability properties is only whether one acts in accord with a particular rule, and whether one is or becomes disposed to do so. This restriction of attention does not limit the relevance of these results to epistemology, however, since how one gets oneself into the state of knowledge, despite being a question of central interest in epistemology, is not per se a topic relevant to the questions whether one is in that state or not and what is required for counting as being in that state. A particular view of the criteria for being in that state may stipulate that how one got there matters to whether one is in the state – genetic views of what knowledge is, like process reliabilism, do that – but that is a choice of a particular theory, not a requirement for having an answer at all to the question what knowledge is. Counterfactual theories care only about what your properties now say that you would believe in an alternate situation or whether what you would believe in an alternate situation would be true. What many internalist justification views of knowledge care about is whether you currently have reasons available. The concern in the value problem too, just like the concern of the True Belief game, is not how you got to your knowledge, or your capacity to have strategies at all, but what it is you now have by being there.

What does it follow that you have on the tracking view when you have achieved knowledge? What follows is whatever follows from having true belief in p and the strong disposition to follow the strategy that is the unique strict Nash Equilibrium and the unique ESS in the True Belief game for p. The dominance of your strategy – its being a sNE – brings a number of properties with it. It brings generality over rounds of play: you will win in every round of the True Belief game (except those few stages where you play out of character – your disposition to follow $K_1$ is not assumed perfect). You will always get a higher payoff than any other strategy could get you. Winning a round doesn't necessarily give you knowledge because it does not necessarily bring you belief. The state of the world may be p or -p, and if it is -p then your winning will come from the clause of your strategy that makes you *not* believe in such circumstances. However, this absence of belief is valuable too, we assumed, the most valuable thing you could do given that state of the world.

Having knowledge, on the tracking view, implies having a true belief and along with it a disposition that would make you have the epistemic state – belief or non-belief – that is most valuable given the state of the world, in almost every round, were you to play the True Belief game an infinite number of times. This security may sound so simple as to be trivial, but its power lies in the fact that in an infinite number of rounds of play the game could have an infinite number of different manifestations. As long as the payoff structure holds constant, having your sNE means you will win in all the remotely probable manifestations. What kinds

of variation could there be? It could be that in the actual case of your true belief that a particular road is the road to Larissa you are not having a discussion with a sophist, but in another round though it still is the road to Larissa you are also stopped along the way by a wily, argumentative guy. If in addition to having a true belief you are a $K_1$-type subject on the matter of this road, then you would believe it both were you to be talking to a sophist and were you not. If the sophist were to give you a bad argument that it is not the road to Larissa then you, the $K_1$ type, wouldn't give up your belief. Being $K_1$ implies that somehow or other you would know better. The subject with a mere true belief would be a sitting duck for the sophistical trick.

In this case the truth value of p did not vary from the actual, but the circumstances did. There could also be a variation in truth value of p over different rounds. Though this is in fact now the road to Larissa, there could be a round of the game where it isn't. If a person merely has a true belief that this is the road then nothing follows about whether she would pick up on the circumstance where it wasn't. There may be signs and trustworthy authorities to tell people where the road goes instead, but being a mere true believer gives no insurance that you would pick up on them. The $K_1$ type of subject, by contrast, is prepared to have an appropriate belief state even if there turns out to be road work.

The counterfactual properties that flow from living in a strict Nash Equilibrium give the subject who has knowledge preparedness for all probable circumstances and changes in the truth value of p. However, we have said that properties of the history of a belief cannot save a view of knowledge from the swamping problem, so one might wonder how properties of the subject's potential future could; why isn't the problem symmetric in time? The basic reason is that time flows in one direction. Everything from the past that is relevant to epistemic success now has had its chance to be taken into account in the actual present belief; what the future may hold cannot have been captured already in that belief. And this is not only because the future has not actually happened yet, but also because what does happen in it will not be determined exclusively by the subject's currently believing or even by that belief's truth. There are a million other present and future conditions not determined by this belief or its truth. What the subject will do in response to those circumstances is also not determined by her actually having a true belief now. Her having a disposition to a strategy in the True Belief game does (probabilistically) determine this.

Being $K_1$ now does add something now that is identifiable and not redundant with merely having a true belief, but we can still ask whether that thing is valuable. This comes down to the question whether *preparedness* is valuable, since what being type $K_1$ now gives is robustness of epistemic success against future contingencies. It seems to me undeniable that preparedness has added value, since denying it would require denying that true belief has extrinsic value at any time before it is actually being used. If the value of a true belief is that it aids you in achieving something else, then it is valuable at the times when it is actually aiding you, but it would not be valuable at any earlier time unless we supposed that the *potential* for aiding was also valuable. If we denied that preparedness is valuable then the true belief about which is the road to Larissa wouldn't be valuable at all except at those times

when we were actually walking on the road with the intention of going to Larissa. We don't think that, so preparedness is valuable.[1] The preparedness that $K_1$, or tracking, brings is not redundant with the potential a mere true belief brings, since whatever success the current mere true belief might give the subject in the future will be dominated in payoffs by what the tracking true belief brings. Thus, tracking is both additional and valuable.

The security that the dynamical stability property of tracking brings is a form of persistence over time, and the value of knowledge over true belief has been associated by some, including Socrates, with persistence. The difference between the current view and the others is in what is expected to persist when you have knowledge of p. Socrates compared true belief with the statues of Daedalus, magnificent creations but they run away if not tied down. Mere true belief will not stay around long either, Socrates said, unless it is tethered, in his view by working out the reason. (*Meno*, 97d – 98a) Timothy Williamson has fleshed this idea out by arguing that knowledge is literally more persistent than true belief in part because it is less susceptible to rational undermining. (Williamson 2000, 79) Kvanvig (2003, 13-20) argues against this, and despairs of the prospects for any persistence view of the added value of knowledge. What everyone is missing is that it is not true belief, or knowledge, that will persist in virtue of one's having knowledge. It is not even appropriate for a theory of knowledge to imply that knowledge of any contingent truth persists, or is likely to persist, because the *truth* of a contingent truth cannot be expected to persist.[x] Roads change, tigers come back, and I'm afraid that chocolate shops sometimes go out of business.

If you have a list of truths about where the chocolate shops are in town, an example due to Kvanvig, then you have something valuable (if you like chocolate), but you do not, he points out, have something more valuable if you have the intersection of this list with a list of where the chocolate shops are *likely* to be. However I do have something more valuable if I have in addition to a list of truths about where the chocolate shops are a responsiveness to their probabilities of going out of business over the next month, quarter, and year; I will be disposed now to try them differentially in the future in a way that insures more success and efficiency in getting my chocolate. If a mere true belief that I have now, before deciding which direction to walk to get my chocolate, is valuable to me because it raises my chances of getting chocolate, then my having now a responsiveness of my beliefs to future closings is of value too, and it is evidently not redundant.

What persists over time for a $K_1$-type believer in p, is not knowledge or belief that p, there is a chocolate shop at a certain place, but *appropriateness* of epistemic state – belief or non-belief in p – over time and changing circumstances. The appropriateness is cashed out in my getting the highest payoff a player could get no matter the state of the world, p or - p. And payoffs, of course, are not restricted to non-essential pleasures. They may be food vs. no food, shelter or health care vs. none; they may be any of the goods, services, and cooperative relationships that are relevant to survival. Provided the payoff structure of the game remains

[1] There are other epistemological cases of the added value of preparedness. Many internalists think that having an argument consciously accessible is valuable, but even if an argument that is being used is valuable an argument that is merely accessible and is not actually being used wouldn't have value if preparedness had no value.

the same, the knower type gets the highest payoff that any type could get, and is highly likely to do so in the future, in every round of play; baldly put, the $K_1$-type believer is more likely to survive and flourish. Since persistence and advantage are intuitive features of the value of knowledge, remarked upon by philosophers of various persuasions, the fact that $K_1$ insures a sensible version of them is a point in favor of tracking as a theory of what knowledge is.

The value of knowledge I have been discussing so far flows from $K_1$ being a strict Nash Equilibrium and what follows from this about the fate of an individual who is a tracker in comparison to individuals of other types. That $K_1$ is an ESS brings in a further dimension that seems to have explanatory force when applied to human populations over time, even historical and evolutionary time. An ESS type has a resistance to extinction. In our game this means that, if payoffs remain the same over generations of play, then a widespread subpopulation of knowers of p cannot be eliminated – as a type – by any small population of any style of ignoramus-type. This holds, recall, no matter what the dynamics of interaction or variations in circumstance.

One might think of this in connection with the ideas that education brings greater success, that the truth will prevail, that an educated population will be less likely to enact policies that will lead to its own destruction, that though the meek may inherit the earth the ignorant will not. We might associate it with hope that Karl Rove was wrong to express disdain for the “reality-based community,” people who "believe that solutions emerge from your judicious study of discernible reality," that he was wrong to think that "That's not the way the world really works anymore. We're an empire now, and when we act, we create our own reality.” (Suskind 2004) We might think of the hope that flat-out false beliefs will not determine the outcomes of elections, or of Congressional policy votes.

These hopes are often thought of as naïve, but the ESS property of knowledge gives them some basis. A subpopulation of organisms of a type that has tiger detection in an environment that has tigers that like to eat them, and which is a large fraction of the total population of such organisms, will never be outcompeted as a type – that is, eliminated as aproportion of the entire population – by a small subpopulation of organisms that lack tiger-detection. This case is clean because the connection between accurate representation of reality and positive relative payoff is direct, and the payoffs are not imagined as changing in the course of repeated play. In such cases knowledge, or its counterparts involving tracking via more primitive representations than belief, will have a ratchet effect on the evolution of species. If a knowledge-bearing type of organism arises and goes to fixation – “everybody” knows – then no ignorant variant within the species can stop the future survival of the knower type. It is worth emphasizing that an ESS insures nothing about the fate of individuals – that is a concern we dealt with earlier in the context of an sNE. In interpreting what it means to say that the knower-type will not be driven to extinction, it seems most natural to say that *knowledge* is what will survive; this is not quite true, but close: what will survive is belief states appropriate to the potentially changing states of the world with respect to p, which of course requires lack of belief in p when p is false. This will survive despite potentially being borne by different individuals in each round of play.

In human history and culture, the knower type, or even a large population of them, does not always prevail. We can explain the consistency of this with a tracking view of knowledge by the fact that human beings have capacities that can lead to failure of the conditions for the ESS outcome. One such condition is that the payoff structure of the game remain constant over repeated play; repeated play, in political contexts for example, often changes the payoffs. If one despises Karl Rove then it will be at least partly because of his evaluative claim that we are permitted to change reality however we please if we have enough power. However, he was right in the factual claim that human actions can change reality, not only in the obvious sense that occurs when we build bridges, but also because human actions can change payoff structures.

For example, in a pitched public battle over a government policy, people with enough resources and cleverness can blanket the media with messages psychologically well-crafted to convince the populace of outright falsehoods, with the goal of making them proponents or opponents of the policy. The best payoff for a member of Congress deliberating over whether to vote for the policy might have been determined by whether it truly would improve the lives of his constituents, because after all even if his only concern was re-election his constituents' well-being would be the thing that determined whether they voted for him – right? However, if the election is soon and the consequences of the policy will emerge more slowly, and the member of Congress doesn't have the resources to counter sufficiently the falsehoods that his constituents have come through aggressive advertising to believe, then what is truly in their best interest will not be determining their vote in the next election. The member of Congress now has the highest payoff from voting in the direction of policy that will not help his constituents.

For any p a true belief in which is valuable, etc., the knower type, if in sufficient numbers, will survive as long as conditions relevant to payoffs remain the same. We can see through empirical examples that one condition for their remaining the same is that there be no relevant deception. However, this amounts to saying that the mere *having* of knowledge by many people will not prevent the damaging consequences of deception, and it is not surprising that overcoming the effects of outright obfuscation on a system of interaction will require not just knowledge but also countermeasures.

It is an understatement to say that among human beings the applicability of the ESS result will not be universal; it is a result in an idealization that has the kind of usefulness such tools bring. We have seen that the idealization implies that even perfect knowledge does not alone give us everything we need epistemically, which conforms to the well-known fact that knowledge does not give us everything. However, we also know that knowledge gives a lot, and the tracking view of what knowledge is provides a simple and powerful picture for explaining what and how that is.

## References

Floridi, Luciano. 2005. “Consciousness, Agents, and the Knowledge Game.” *Minds and Machines* 15: 415–44.

Goldman, Alvin. 1979. “What is Justified Belief?” In *Justification and Knowledge*, edited by G. Pappas, 1–23. Dordrecht: D. Reidel.

Kvanvig, Jonathan. 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Lewis, David K. 1969. *Convention*. Cambridge, Mass.: Harvard University Press.

Plato. *Meno* 97d – 98a.

Roush, Sherrilyn. 2005. *Tracking Truth: Knowledge, Evidence, and Science*. Oxford: Oxford University Press.

Swinburne, Richard. 1999. *Providence and the Problem of Evil*. Oxford: Oxford University Press.

Swinburne, Richard. 2000. *Epistemic Justification*. Oxford: Oxford University Press.

Suskind, Ron. 2004. “Faith, Certainty, and the presidency of George W. Bush.” *New York Times Magazine* October 17, 2004.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press.

***Appendix***

From

$$Pr(M_1/p) = \text{very high}$$

$$Pr(B(p)/M_1) = \text{very high } (*)$$

and

$$Pr(B(p)/p.M_1) = Pr(B(p)/M_1)$$

we wish to derive

$$Pr(B(p)/p) \text{ is high } (\dagger)$$

By (*),

$$Pr(M_1 \,.\, p) \approx Pr(p)$$

$$Pr(B(p) \,.\, M_1) \approx Pr(M_1) \; (!)$$

$$Pr(B(p)/p.M_1) = Pr(B(p)/M_1)$$

implies

$$Pr(B(p) \,.\, p \,.\, M_1)/Pr(p \,.\, M_1) = Pr(B(p) \,.\, M_1)/Pr(M_1)$$

By !,

$$Pr(B(p) \,.\, p \,.\, M_1)/Pr(p \,.\, M_1) \approx 1$$

which implies

$$Pr(B(p)/p \,.\, M_1) \approx 1.$$

but by !

$$Pr(M_1 \,.\, p) \approx Pr(p)$$

so

$$Pr(B(p)/p) \approx 1.$$

Two approximate equalities were used in the derivation of the final inequality, so there are two sources damping down the final correlation and it will be strictly lower. This is why I claim only that high tracking correlations come from very high signaling system correlations rather than that very high comes from very high. The derivation for the second half of strategy $K_1$ is analogous. The nature of probabilistic tracking is discussed in Roush (2005); lower bounds on the loss of correlation over tracking links are shown in Chapter 5, where the

links are chains of evidence. Both parts of a strategy like $K_1$ are actually tracking conditions, though not the specific ones about p and B(p) that we use to analyze knowledge.

---

[i] This work was supported by NSF Award No. SES - 0823418. Special thanks to Jason Alexander for helpful discussions, and to Brian Skyrms for teaching me most of what I know about game theory.

[ii] We may wonder, though, why in the case of *really* beautiful chairs having been designed or made by an artist or establishment who usually makes very beautiful chairs tends to make the chair sold under that label have higher market value.

[iii] The particular numbers are important here only for some of the ordinal relationships: the payoff for a true belief must be greater than that for no belief when p is true, and no belief must be of greater value than belief when p is false. How much greater may be different with the two types of mistake one might make, depending on how costly a false positive or false negative is for the subject.

[iv] Such an assumption would anyway trivialize the representation. The question would then be If the subject knows that Nature made p true, should she believe p? Since knowledge implies true belief the answer would be automatic – she already would believe it – and the representation would swing independent of whether true belief has value or not.

[v] That is, since there isn't a strict Nash Equilibrium there will be no evolutionarily stable strategy (ESS). The notion of ESS makes sense for this asymmetric game if we think of ourselves as referring to its symmetric counterpart in which every player has both a sender strategy and a receiver strategy which he or she plays depending on whether he or she is assigned the role of sender or receiver in a given round of play. All of the claims about stability in the Watching football game should be taken to be referring to that symmetric game.

[vi] *Two* player types could successfully gang up on an ESS, so an ESS is uninvadable, but not unbeatable. The mixed strategies discussed below are not equivalent to two strategies ganging up in the relevant way.

[vii] See Roush (2005) for a fuller discussion of these conditions.

[viii] If there are no such probabilities for how fulfillment of a proposed requirement for knowledge would make you do in the task of believing p when p and not believing when -p, then the requirement has no truth connection at all, and so could not invade our tracking-based strategy anyway.

[ix] Tracking with closure (Roush 2005) is weaker than tracking because it allows one also to know p in virtue merely of tracking some q which one knows implies p. On that view knowing p would not imply one had an ESS for p. However, knowing p would imply that either one had an ESS for p or one had an ESS for some q that implies p.

[x] The truth of necessary truths does, of course, persist, so tracking is not the appropriate kind of responsiveness to have to necessary truths. The appropriate kind is proposed and defended in Roush (2005, 134-147).